Listening. Learning. Leading.®

Practical Considerations in Computer-Based Testing

January 2011

Practical Considerations in Computer-Based Testing

Choosing whether to test via computer is the most difficult and consequential decision the designers of a testing program can make. The decision is difficult because of the wide range of choices available. Designers can choose where and how often the test is made available, how the test items[1] look and function, how those items are combined into test forms, and how those forms produce scores. The decision is consequential because it can impact every aspect of the testing process, from item development and test assembly, through test delivery and response collection, to the scoring and reporting of results.

It is widely believed that all tests will one day be delivered on a computer of some sort (Bennett, 1998, 2002). However, it is difficult to accurately predict when this day will come. It has seemingly been just around the corner since the early 1990s, when a handful of early adopters, including the ASVAB (Sands, Waters, & McBride, 1997) and the GRE® (Mills, 1999), signed on to computer-based testing (CBT). Today, dozens of admissions, placement, certification, and licensure testing programs are administered on computer, with the number growing each year. On the K–12 front, several states already conduct their annual accountability testing on computer, and many others are poised to join in. Hundreds of schools or districts also employ CBTs in a formative or diagnostic role in service of instruction.

As with all trips, we need to start by deciding whether the attractions of the destination outweigh the rigors of travel.

This paper is intended to assist in some small way those practitioners who are struggling with the decision of whether to test on computer and how they might best go about doing so. It is not intended to provide an introduction to the important methods and to the considerations that dictate use of those methods. It is far from comprehensive in this regard. To make truly informed decisions, one would need to be at least conversant — if not quite familiar — with most of the references listed here. In short, this is not a roadmap intended to precisely direct practitioners on their journey to computerized delivery, but rather more a guidebook highlighting some of the landmarks they will likely visit along the way. As with all trips, we need to start by deciding whether the attractions of the destination outweigh the rigors of travel.

Why Test on Computer?

There are three basic reasons for testing on computer. The first is to enable measurement of constructs or skills that cannot be fully or appropriately captured by paper-based tests (Bennett 2002; Parshall, Harmes, Davey, & Pashley, 2010). The second is to improve measurement by increasing the precision or efficiency of the measurement process (Parshall, Spray, Kalohn, & Davey, 2001; van der Linden & Glas, 2000; Wainer, 1990). The third is to make test administration more convenient for examinees, test sponsors, or both. Each of these potential advantages of CBTs is discussed in turn below.

Changed Measurement

Standardized tests often are criticized as artificial and abstract, measuring performance in ways divorced from real-world behaviors. At least some of this criticism is due to the constraints that paper-based administration imposes upon test developers. Paper is restricted to displaying static text and graphics, offers no real means of interacting with the examinee, and sharply limits the ways in which examinees can respond. Computers can free test developers from these restrictions. Computers can present sound and motion, interact dynamically with examinees, accept responses through a variety of modes, and even score those responses automatically. For example:

- A test assessing language proficiency can measure not only how well students can read and write, but also their ability to comprehend spoken language, speak, and even converse.
- A test measuring proficiency with a software package can allow students to interact with that software to generate or express their responses.

[1] The term *item* is synonymous with *question* or *task* and refers to the content of a test form.

- A science test can allow students to design and conduct simulated experiments as a means of responding.
- A medical certification exam can allow examinees to interactively evaluate, diagnose, treat, and manage simulated patients.
- A writing test can allow students to write and edit their essays in a familiar word-processor environment (as opposed to the increasingly less familiar pen-and-paper). Furthermore, the computer is able to score that essay automatically and instantly provide the student with specific, diagnostic feedback, coupled with instruction for improvement.

As these examples illustrate, a CBT can be a richer, more realistic experience that allows more direct measurement of the traits in question.

Items and tasks like those highlighted above have become the subject of considerable research as testing on computer has become increasingly practical and popular (Clauser, 1995; Haladyna, 1996; Huff & Sireci, 2001; Parshall et al., 2010). A wide assortment of options are therefore now available for using a computer to present information, facilitate interaction, and collect responses in ways not possible with traditional text-based items. The question is then how (or whether!) these capabilities can be used to materially and substantively improve measurement.

A CBT can be a richer, more realistic experience that allows more direct measurement of the traits in question.

It cannot be stated vigorously enough that innovative items, tasks, and scoring mechanisms should be a consequence of the constructs a test is required to measure rather than merely a consequence of the test being delivered on computer. The form of a test and its items should directly follow from their function. If a construct can be effectively measured with traditional, multiple-choice items, there is nothing gained through use of more innovative item types. Indeed, such use may run a considerable risk of introducing what amounts to noise or error into the measurement process. As an example, consider a math geometry item that provides various on-screen *tools* (e.g., ruler, protractor, compass) that students can manipulate to determine their answer. Unless the implementation of these tools and the instructions regarding their use are properly handled, this item may end up saying as much about a student's ability (and experience) with computers as it does his or her knowledge of geometry.

The test developer must therefore tread a fine line, adopting innovation when it is necessary to best measure a construct and resisting its temptations when conventional item types will suffice. Like all matters related to test development, the decision process rightly starts with a comprehensive analysis of the construct being measured and how evidence of a student's standing on it is best collected. This analysis may well uncover gaps between what a test *should* be measuring and what it *could* measure under the constraints imposed by paper-and-pencil tests. Item types that appear to best address these gaps can then be identified or designed. Design of the new item types may well be an iterative process, with successive rounds of pilot testing informing design revisions (Harmes & Parshall, 2010).

Improved Measurement Precision and Efficiency

Certain types of CBTs can change not just the nature of what is measured, but the measurement process itself. The key to doing so is, again, the ability of the computer to interact with and tailor itself to the student being tested. A CBT with these capabilities is termed *adaptive*. As an adaptive test proceeds, answers to earlier questions determine which questions are asked later. The test therefore progressively changes as the student's performance level is gradually revealed.

The basic principle behind adaptive testing is simple: avoid asking questions that are much too difficult or much too easy for the student being tested. Because we are fairly sure (but not certain!) that an able student will answer an easy item correctly or that a struggling student will stumble on a hard question, relatively little is learned by asking these items. Much more is learned by administering items that challenge but don't overwhelm the student or, simply put, items that the student has roughly equal odds of answering correctly or incorrectly. Properly identifying these questions and asking them is the goal of any adaptive test.

Three distinct varieties of adaptive CBTs will be described here. But all CBTs consist of two basic steps:

item selection and score estimation. Both steps are repeated each time an item (or collection of items) is presented and answered. The first step determines the most appropriate item or collection of items to administer given what is currently known about the student's performance level. Items are selected from a *pool* containing more items than any single student sees.

The second step uses the response or responses to the item or items previously presented to refine the score or performance estimate so that the next item or collection presented can be more appropriate still. This cycle continues until either a specified number of items have been administered or some measure of score precision is reached. The process is represented schematically by Figure 1.

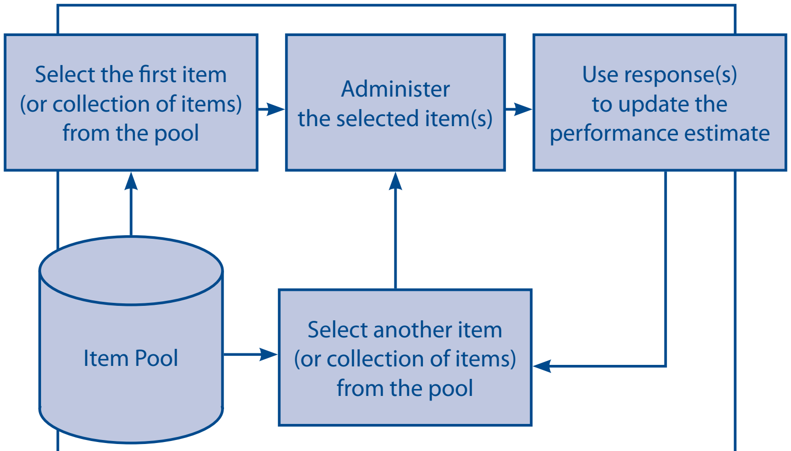
Select the first item (or collection of items) from the pool
Administer the selected item(s)
Use response(s) to update the performance estimate
Item Pool
Select another item (or collection of items) from the pool

Figure 1. The adaptive testing process

Adaptive CBTs can be more *efficient* than conventional tests that present the same items to every student. It is not uncommon for an adaptive test to match the precision of a conventional test containing 25% more items. Conversely, an adaptive CBT can match the length of a conventional test but return more precise measurement, particularly of the students at either extreme of the performance continuum.

Adaptive CBTs can be more efficient than conventional tests that present the same items to every student.

Researchers have developed and proposed a host of procedures and options for implementing each of the basic tasks needed to assemble and score an adaptive test. Methods have proliferated largely because none can be recommended as ideal for all testing situations and circumstances. Instead, the procedures that are best depend on the unique characteristics of a given testing program. Test content, item formats, the examinee population, and even the subjective values of the test's owners and score users are all relevant considerations. The process of deciding among the various design possibilities and choosing those best suited for a particular testing program will be taken up below.

Increased Convenience

The third major benefit of computerized testing is operational convenience for students, test administrators, and those who use test scores. These conveniences include:

Self-proctoring. Conventional paper-and-pencil tests usually require someone to distribute test booklets and answer sheets, keep track of time limits, and collect materials after the test ends. Administering a CBT can be as simple as parking a student in front of a computer. The computer can collect identification data, orient the student to the testing process, administer and time the test, and produce a score report at the conclusion. Different students can sit side-by-side taking different tests with different time limits for different purposes. With conventional administration, these two students might need to be tested at different times or in different places.

Immediate scoring. The value of any information degrades over time. A score report based on a test taken six weeks ago is a description of what that student *was* rather than what she or he currently *is*. CBTs can address this distinction by providing students with score reports immediately upon conclusion of their test. The test can therefore have instant impact. At the student level, this might involve quickly changing the instructional approach taken with a particular concept. At the school or district level, immediate information might allow similar but more global tactical shifts in the instruction process.

Of course, a test can provide results immediately only if the computer is able to score all of the items presented. Whether or not this is possible depends on the sort of items administered. Any item type for which students select or otherwise indicate their responses (e.g., clicking, highlighting, dragging and dropping) can easily be scored by computer.

Writing samples can sometimes be scored by computer (Attali & Burstein, 2006). However, some items or tasks remain dependent on human raters for scoring. A CBT that includes such items or tasks will therefore not be able to report full results immediately. A compromise approach where the computer reports what it can and full results follow after human ratings are produced remains a possibility.

Integrated data management systems. Testing on computer can allow scores to be entered automatically into classroom-, school-, district-, or state-level databases. Once there, various individual and aggregate reports can easily be produced to summarize and track the performance of individual students and defined groups.

Diagnostic assessment and integration with instructional software. Self-proctoring, immediate scoring, and easy data management makes CBTs — adaptive CBTs in particular — ideal for diagnostic or formative assessment. Consider the problem of assessing a student's pattern of strengths and weaknesses across a fairly broad content domain. An adaptive CBT can begin with a brief overview of the domain to determine the student's overall level of proficiency. This is akin to searching a dark room with a relatively dim, but wide-beamed flashlight. The locations of large objects can be mapped but details would not be visible. Interesting objects are best examined more closely with a brighter, more narrowly focused beam. Certain adaptive CBTs could be specifically designed to switch continually between these roles and thus would be uniquely suited for this kind of search.

A further advantage on the diagnostic front is the ability to connect the scores output from a CBT directly to instructional software. This can allow the diagnosis-remediation cycle to proceed much more quickly and easily than might be possible with paper-based tests.

Flexible scheduling. Because CBTs can be self-proctored and self-scored, they can allow testing to take place when schools and/or students find it convenient rather than according to some imposed schedule.

Reach and speed. Although CBTs are sometimes given in fixed sites dedicated to test administration, they can theoretically be delivered anywhere and anytime a computer is available. It is also possible to get a CBT packaged and distributed much faster than a paper test can be formatted, printed, boxed, and shipped. This situation can allow tests to change rapidly in order to keep up with fast-changing curricula or subject matter.

Preference. Most surveys reveal that students overwhelmingly prefer testing on computer to testing on paper (Cassady & Gridley, 2005). The extent of preference is likely to grow with successive generations of students, whose exposure to and use of computers will be increasingly widespread, at ever-younger ages.

Although the potential advantages of CBT are clear, fully realizing these benefits is contingent in large part on choosing the right design for implementation. The following section will outline some of the design options that are available and sketch out a process for choosing amongst them.

The design of any test is therefore necessarily a compromise, ideally one that properly reflects the values of its developers, the preferences of examinees and administrators, and the needs of its score users.

Computer-Based Test Design Options

In a perfect world, every test would be short, reliable, secure, convenient to administer, cheap to develop, and easy to maintain, and would offer immediate and detailed summative and formative scores. Unfortunately, in the real world, most of these desirable characteristics stand in direct opposition to others. It is difficult for a short test to report detailed and reliable scores. It is difficult for a test that includes enough alternate forms or a large enough item pool to be secure under repeated administration to be cheaply developed and easily maintained. And it is very difficult for a test to provide both formative and summative information. The design of any test is therefore necessarily a compromise, ideally one that properly reflects the values of its developers, the preferences of examinees and administrators, and the needs of its score users.

Finding the right compromise requires designers to sort through the various properties that a test can exhibit and decide which are essential and which can be partially or wholly sacrificed. The list of preferred properties can then be compared to the pattern of strengths and weaknesses inherent in the available CBT administration models so that the best fitting model can be determined.

The designer's task is then to identify the model whose strengths best match priorities.

We begin below with a discussion of five test properties that would typically be considered during the design process. This is followed by a description of five alternative CBT models. Each of the key properties essentially defines a scale along which the competing administration models can be roughly ordered by the extent to which each model is likely to manifest the property. The designer's task is then to identify the model whose strengths best match priorities.

Test Properties

The most important test properties can be organized into five categories: measurement efficiency, test security, item development requirements, design complexity, and cost. Each of these is described and briefly discussed below.

Measurement efficiency. All designers would like their tests to be both short and reliable. Short tests make students and teachers happy by taking less time away from other activities. Reliable tests make score users happy by allowing better inferences to be drawn from test scores. Unfortunately, test length and reliability are strongly related, with reliable tests tending to be longer and shorter tests tending to be less reliable. However, CBT designs differ considerably in *measurement efficiency*, which can be loosely defined as "reliability divided by test length." An efficient test, then, is one that offers more measurement precision per item or, perhaps more importantly, more precision per unit time.

Test security. A test score is useful only to the extent that it is a genuine measure of a student's capabilities. This would not be the case if a student achieved a score by copying from neighboring examinees or through prior exposure to the items on the test. Concerns regarding security go hand-in-hand with the consequences or stakes attached to a test's scores. Tests used for placement or formative purposes may be relatively immune because students would derive no benefit from cheating. However, admissions or exit tests are generally more consequential and hence require more attention to security.

CBTs are subject to many of the same security concerns that afflict conventional paper tests. However, there are a few differences worth pointing out. First, the administration environment for CBTs can greatly reduce the possibility of students copying from one another, at least absent organized collusion. Unlike answer sheets that sit open on desktops and are accessible to prying eyes, answers usually appear on the screen of a CBT only fleetingly before being replaced by the next item. CBT designs that vary the items administered or the order of their administration across students are even better in this regard. Items stored in encrypted files on a computer are also much better protected prior to administration than a box of booklets locked in a desk drawer or closet.

Students can also gain pre-exposure to items prior to testing due to administration policies. Allowing students to retest with the same test form with which they were originally tested is an obvious example. Even if students are not allowed to retest, repeated use of the same form (or item pool) over an extended period of time in a high-stakes environment may entice students (or teachers!) to promote widespread awareness of its content. Some CBT designs are just as vulnerable as conventional tests to retest and reuse policies. Although designs that administer to each student only a fraction of the items available for use can be a bit more secure, one must be careful not to overstate the benefits here.

CBTs are subject to many of the same security concerns that afflict conventional paper tests.

Consider, for example, a high-stakes conventional paper-based testing program that tests a large number of students on each of a handful of administration dates throughout the year. Security concerns might have historically dictated that each test form be used on only a single occasion and then discarded. This effectively eliminates the problem of item pre-exposure. Of course, this administration schedule and this form reuse policy can be applied under CBT as well. But providing large groups of students with computers is much more difficult than providing them with No. 2 pencils. It is therefore not uncommon to see CBT testing spread over administration "windows" covering days or even weeks. In fact, the conveniences noted earlier actually may make a CBT easier to offer on-demand than confined to a limited number of administration dates. Test designers must therefore carefully weigh the potential impact of CBTs on administration policies and the consequences those policies might have on test security.

Item development requirements. CBT designs differ considerably in the number of items that need to be developed to properly support administration. Some designs have requirements identical to paper testing. However, designs that present each student with only a portion of the items available may require that substantially more items be developed. The stakes attached to testing also play a large role in determining development requirements since security concerns dictate the frequency with which test content is replaced. It is also important to note that the innovative CBT items needed to change the nature of measurement may be much more difficult and expensive to develop than text-based multiple-choice items.

Design complexity. Complexity can be divided into three general areas that concern, respectively, the test administration model, the scoring methodology, and the mechanisms that must be in place to ensure that scores are comparable across time.

The test administration model determines which items are seen by which students. Under the simplest model, every student is administered the same items, perhaps even in the same order. Under more complex adaptive models, each student may be administered a unique combination of items.

The scoring methodologies applied under various test designs can similarly differ in complexity. Simply counting up the number of right answers sits at the uncomplicated end of the scale. However, adaptive tests that administer different tests to each student need much more complex scoring schemes to ensure that scores are comparable.

The third aspect of complexity concerns the statistical mechanisms required to ensure comparability not just across examinees, but across time as well. Many testing programs periodically develop and introduce new test forms or replenish item pools. The frequency with which this is done is dictated both by security concerns and by how "perishable" the item content is. Replacement of test forms or item pools should have no impact on the interpretation or comparability of reported scores. But statistical procedures must be in place to ensure this. In the case where one test form replaces another, these procedures are called *equating* methods. These methods have been well-researched and extensively applied in operational practice. However, replacing or changing an item pool requires a different, more complicated set of methods to ensure continuity.

All other things equal, simpler test designs are preferable to more complex designs. Simpler designs, with less to go wrong, are more robust and more resistant to unanticipated problems. Simpler designs also are generally cheaper to develop and maintain. Complex designs can be more efficient and more secure, but are likely to impose higher item development requirements and maintenance costs.

Cost. Although saved for last on this list of considerations, cost usually exerts a very strong influence on test design. The drivers of cost include item development, administration logistics, score reporting, and the statistical work needed to effectively maintain a testing program over time (although this last component generally is tiny in comparison to the others). Simple test designs making judicious use of innovative item types and featuring less-frequent administration dates or windows will come in at the more economical end of the scale. Conversely, complicated adaptive test designs that push the limits of innovative measurement, accommodate students with frequent or continuous administrations, and operate under strict security policies are likely to strain the budget of even well-heeled test sponsors.

CBT Administration Models

The most important decision made by a CBT designer is the choice of the *test administration model*, which controls the items with which a student is presented and the order in which they are presented. The administration model strongly impacts all of the test properties discussed above, largely determining the efficiency, security, item development needs, complexity, and cost of the testing program.

Five distinct test administration models are described below. Three of these can be considered adaptive in that the testing process can change in response to each student's performance. The description of each model is supplemented by a table highlighting strengths and weaknesses relative to the five test properties.

Linear or fixed form. The simplest type of CBT essentially replicates the administration model of conventional paper tests. Each student is presented with the same set of items, either in the same order or in a randomly scrambled order. Fixed form CBTs are constructed and scored like conventional tests as well, with scores computed either by totaling the number of correct answers or through item response theory (IRT) methods. (See Table 1.)

Table 1

Linear or Fixed Form Computer-Based Tests

Efficiency	A fixed form CBT has the same measurement efficiency as a conventional paper test.
Security	A fixed form CBT is more secure than a conventional test with respect to students copying from one another or inadvertent disclosure of test content. The fixed form model is as secure as a conventional test in terms of item disclosure through repeated administration of the same test form. Reuse must be limited to maintain security in medium- and high-stakes settings. Multiple, parallel test forms are needed to maintain security if frequent administrations are offered.
Item development requirements	A fixed form CBT requires the same number of items as a conventional test. However, development cost may be greater to the extent that innovative item formats are used.
Complexity	As simple as a CBT gets.
Cost	The lowest of all the CBT models.
Comments	(1) Fixed form CBTs can and should allow students to freely navigate the test, skipping forward or returning to check previously answered items. As noted below, this freedom is often denied to examinees under adaptive test models, which usually require that the current item be answered in order to move to the next and prevent return to previously answered items. (2) Before implementing the scrambled-order variant of a linear CBT, it should be confirmed that item order does not affect student performance.

Random form. Under this model, each student is presented with a set of items drawn from a pool containing more items than necessary to construct a single test form. Items are usually drawn from the pool to satisfy specified substantive or statistical rules. These rules are imposed to ensure that the different forms drawn for different students each measure the same content and are parallel in difficulty and reliability. Although test scores can be computed by totaling the number of correct answers, IRT methods may be preferred. (See Table 2.)

Table 2

Random Form Computer-Based Tests

Efficiency	A random form CBT has the same measurement efficiency as a fixed form CBT or a conventional paper test.
Security	Like a fixed form CBT, the random form model is more secure than a conventional test with respect to students copying from one another or inadvertent disclosure of test content. But because each student sees only a fraction of the items available, the random form model can be slightly more secure than fixed form CBTs or conventional forms with respect to item disclosure due to repeated administration. Larger item pools add incrementally to test security. However, unless pools are quite large (e.g., the equivalent of 10 or more test forms) they will not remain secure long in even medium-stakes settings. Accordingly, pools will need regular replenishment or replacement if frequent administrations are offered.
Item development requirements	A random form CBT requires more items to establish a pool than fixed form CBTs or conventional tests require to construct a single test form.
Complexity	A random form CBT is more complicated than fixed form CBTs or conventional tests in several ways. First, the random form administration software must construct a test form for each student, as well as deliver it. Depending on the number and nature of the test construction rules or specifications, this can be far from straightforward. Second, IRT-based scoring may be required to ensure that the different test forms administered to different students produce comparable scores.
Cost	More expensive than the fixed form model, due to increased item development requirements, more complicated (IRT-based) scoring rules, and, perhaps, more complicated test administration software.
Comments	(1) Random form CBTs can allow students to freely navigate the test, skipping forward or returning to check previously answered items. (2) Random form CBTs are psychometrically riskier than fixed form tests because each student takes an essentially unique form, requiring that different, less-robust equating methods be employed. (3) Random form CBTs also are riskier than fixed form or conventional tests from a content perspective because the unique forms presented to each student must be constructed by the computer and cannot be reviewed by a test developer prior to administration.

Multi-stage. The first, and simplest, of the adaptive administration models is the *multi-stage test* (MST). An MST begins by presenting each student with a first-stage or *routing* test, which will typically contain 10 or more items. Once the student completes the routing test, a score is produced and a decision is made. This decision is to choose among two or more second-stage tests by determining which is most appropriate given performance on the routing test. Following the standard principles of adaptive testing, students who performed well on the routing test are assigned a second-stage test composed mainly of more difficult items, while students who struggled are administered an easier second-stage test. Upon completion of the second-stage test, a two-stage MST ends and a final score is produced that aggregates performance across both the routing and second stages of the test. However, more elaborate branching designs also are possible, with additional decisions and a third or fourth stage following the second. This process is illustrated in Figure 2 for a simple two-stage test.

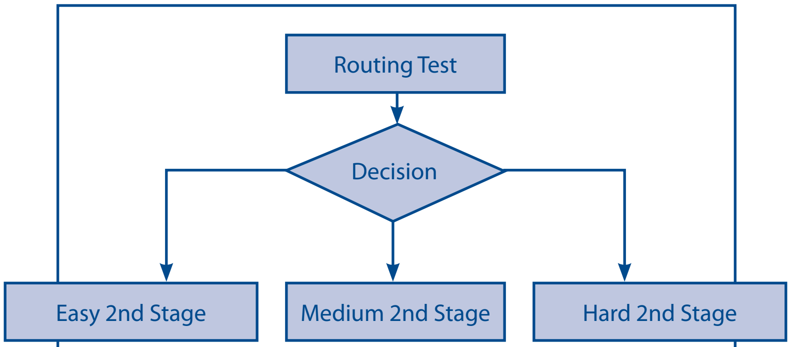

Figure 2. Branching designs for multi-stage tests

Each stage of an MST can be and often is presented as an intact, separately timed section. MSTs are constructed essentially as miniature tests (and so sometimes called *testlets*), in accord with detailed content and statistical specifications. The routing test is assembled to broadly sample the content domain, focusing on items of middle difficulty. Second-stage tests also are selected to represent the content domain, but differ from one another in item difficulty.

There are three basic challenges to scoring a multi-stage test. The first is to produce a score on the routing test that properly informs the selection of the correct second-stage test. The second is to combine a student's results across the two (or more) stages into a coherent total score. The last is to ensure that students who took different second-stage tests receive scores that are directly comparable to one another.

Several approaches can be taken to meeting each of these three goals. But the simplest, most robust, and least likely to run into technical problems and complications is to use number-right scoring and equating procedures of the sort that have been successfully employed with conventional tests for decades. Although it may seem paradoxical (and perhaps even a bit old-fashioned) to apply conventional scoring procedures to an adaptive test, there are compelling reasons for doing so. As expanded on below, the most important is to avoid making the strong assumptions that IRT-based scoring requires. Students actively engaged in the learning process and learning at different rates do not always behave as these strong assumptions presume. The theoretical benefits promised by IRT scoring are therefore unlikely to be realized in practice. The conservative solution is therefore the best solution under these circumstances.

The three scoring challenges of an MST are met by number-right scoring as follows. First, a number-correct score is computed on the routing test to inform the branching decision. The rule for choosing a second-stage test is simple, implemented by dividing the range of the routing score into bins or intervals. The second-stage test is then determined by the bin into which a routing score falls.

Combining the results of the routing and second-stage test is also straightforward: The scores are simply added together. However, the aggregated number-right scores are not directly comparable if examinees have taken different second-stage tests. Clearly, scores for examinees who have taken the routing test in conjunction with the easier second-stage tests are not on the same scale as the scores of examinees who took the routing test and a hard second-stage test. All adaptive tests face this same problem, but multi-stage tests can solve it in a particularly simple way by equating the different test forms (combinations of routing and later-stage tests) by robust and time-tested methods. (See Table 3.)

Table 3

Multi-Stage Tests

Efficiency	An MST can easily exceed the efficiency of a conventional, fixed, or random forms test by 20% or more. Under most circumstances, the efficiency of a MST is on par with that of the item-adaptive model, described below.
Security	Like all CBTs, MSTs are more secure than a conventional test with respect to students copying from one another or inadvertent disclosure of test content. Depending on the number and the arrangement of its stages, an MST is roughly on par with the random forms model in terms of item disclosure due to repeated administration. More "routes" or unique combinations of the stages and testlets within stages produce the same effect as increasing the pool size of a random forms test and add incrementally to test security. However, a single MST cannot long remain secure in even medium-stakes settings. Fortunately, MSTs are relatively easy to construct, at least compared to the item-adaptive model. Periodic replacement of an MST with a new version is therefore not a terribly onerous task. In fact, frequent replacement of test content is a common feature of a number of operational implementations of MSTs (Breithaupt & Hare, 2007).
Item development requirements	An MST CBT requires more items to populate the various stages and testlets than a fixed form CBT or conventional test. Item development requirements are roughly equal to those of the random form model, but less than those of the item-adaptive model.
Complexity	Although the MST is quite a bit more complicated than fixed form CBTs or conventional tests, it is by far the simplest of the adaptive models. This is particularly true if number-right scoring is used for routing decisions and final scoring.
Cost	More expensive than any of the nonadaptive models, but much cheaper than the item-adaptive model.
Comments	(1) MSTs can allow students to freely navigate within a stage, skipping forward or returning to check previously answered items in any given testlet. This is particularly effective if each stage is administered in a separately timed test section. (2) MSTs permit complete control of test content and composition. Importantly, all possible MSTs can be reviewed prior to administration. This is in contrast with the random forms or item-adaptive models, where the unique forms constructed by computer for each student cannot be reviewed prior to administration. Instead, the computerized item selection routine must be trusted to assemble an acceptable form in real time. The difference between the two approaches is particularly stark in an achievement test setting where proper control of test content is of paramount importance.

Item-adaptive. Essentially the extreme case of the MST, the item-adaptive model computes a score following each item, and makes a decision as to what to present next. However, the item-adaptive model is far less deterministic than the MST, instead drawing items from a pool like that supporting the random forms model. Items are selected from the pool based on the performance level a student has demonstrated on items administered earlier in the test. Item selection is usually intended to best meet some or all of three overarching goals.

The first goal is to maximize test efficiency by measuring students as precisely as possible with as few items as possible. In practice, this typically means selecting an item that challenges but does not overwhelm the student. As pointed out above, we learn the most about a student's level of performance by asking questions that are neither too easy nor too difficult for that student.

The second goal is to construct for each student a test that is properly balanced in terms of item substance or content. The intent is to have adaptive tests follow the same sort of content-driven test assembly process that has been used with conventional tests for decades. The substantive meaning of test scores has long been dictated by creating test specifications that spell out in detail the sort of items a test includes and in what proportions they are included. The challenge, of course, is in meeting these standards when the computer is assembling a test on the fly as it proceeds. But we can be certain that the scores of different students who took different test forms are substantively comparable only when test assembly standards are properly expressed and are reliably met by the item selection algorithm that drives the item-adaptive testing process. That standards are reliably met is acutely important given that there is no opportunity for a test developer to review the test form with which any examinee was presented prior to it being administered.

The third item selection consideration is to protect certain items from overexposure and to encourage the use of other, less-popular items. Without such protection, some items will be administered to a large proportion of examinees, while others will be used rarely or not at all. Items used frequently can threaten test security as they become known to students; items administered too seldom are a waste of resources (Davey & Nering, 2002; Mills & Steffen, 2000).

Once an item is selected and administered, the response to it is used to refine an ongoing estimate of the student's level of performance. This estimate is necessarily rough early on, but improves as the test continues. Performance is estimated and test scores are computed according to an IRT model that is assumed to accurately characterize the interactions of students with items.

Researchers have developed scores of specific, competing methodologies for each component of the item-adaptive testing process. These methods determine how items are selected, how test content is most reliably balanced, how attractive items are most appropriately protected from overuse, and how test scores are best computed. Even an overview of these methods is beyond the scope of this paper. Those seeking a more complete description of the available options and the relative advantages each conveys are referred to Davey and Pitoniak (2006). (See Table 4.)

Table 4

Item-Adaptive Tests

Efficiency	An item-adaptive test can be more efficient than an MST, although in practice the differences are small. The item-adaptive model can easily exceed the efficiency of a conventional, fixed, or random forms test by 25% or more. However, as noted above, most item-adaptive tests do not single-mindedly pursue the goal of maximum efficiency. Instead, the drive for efficiency is tempered by the competing priorities of ensuring that content requirements are met and that attractive items are suitably protected against overexposure. Imposing strict content standards is therefore likely to lower test precision by forcing the selection of items with less-optimal measurement properties. Similarly, strongly protecting against the over-administration of items with exceptional measurement properties will depress efficiency as well.
Security	Like all CBTs, item-adaptive tests are more secure than conventional tests with respect to students copying from one another or inadvertent disclosure of test content. If properly configured, the item-adaptive test potentially offers the best protection of any of the adaptive-test models against item overexposure and attendant security difficulties. However, properly configuring an item-adaptive test is neither easy nor common in operational practice. Doing so requires that test efficiency be greatly sublimated to the goal of protecting attractive items and equalizing exposure rates across items. A variety of exotic methods for building such protections into the item-selection process have been developed, many of which are described in Davey and Nering (2002) and Davey and Pitoniak (2006).
Item development requirements	As commonly implemented, item-adaptive tests impose the highest item development requirements of any CBT model. As would be expected, requirements are strongly affected by security concerns. Strong protection of items against overexposure must be supplemented by large item pools. High-stakes settings also require that item pools be frequently replenished or replaced.
Complexity	Item-adaptive testing is by far the most complex of the CBT models. This is particularly true if security concerns or stiff content requirements force the use of exotic and difficult-to-implement test administration methods and software. Such complexity can impose costs beyond the financial, greatly complicating test development and long-term operational maintenance.
Cost	As commonly implemented, the most expensive of the CBT models. This is largely the result of operational complexity and high item development requirements.
Comments	(1) Item-adaptive tests typically do not allow students to freely navigate by skipping forward or returning to check previously answered items. Skipping forward is particularly problematic, given that the response to the current item is needed in order to properly select subsequent items. (2) Item-adaptive is the most flexible of the CBT administration models, providing a wide variety of options and allowing designers to explicitly address the delicate balance among efficiency, content control, and test security. (3) Because an item-adaptive test unfolds in real time as items are successively selected and administered, the process is inherently unpredictable and, therefore, difficult to completely control. Some small number of students will be poorly served, either by receiving a test that does not fully conform to content specifications or by receiving a score that does not precisely capture their ability.

Computerized classification. The computerized classification test (CCT) is the least common and most widely overlooked member of the adaptive CBT family. This is largely because the CCT is a special-purpose testing model that pursues objectives very different from those of other CBTs. The CCT does not seek to assign each student a precise numeric score, but rather attempts to classify students into groups. Pass/Fail, Master/Nonmaster, and Basic/Proficient/Advanced are examples of common groupings. Groups are defined by one or more cut-points or *classification thresholds*, which are points along the performance scale that separate the students assigned to one group from those assigned to the next.

Of course, it is both possible and common to assign students to one of several defined groups by comparing the numeric scores they receive on a test to the various classification thresholds. However, doing so sacrifices both efficiency and classification accuracy in comparison to what a CCT is capable of. A CCT can quickly and accurately classify students because it is not interested in drawing distinctions among students that are not essential to classification. In practice, this means that students assigned to the same classification group are considered as having performed equivalently on the test. Although students in one group can be differentiated from students in another group, they cannot be differentiated from one another.

Although the inability to draw distinctions between students with the same classification may be seen as too high a price to pay for the efficiency of a CCT, there are many testing situations where classification is, in fact, all that is needed. Licensure and certification testing provides a variety of examples. On the academic front, low- and medium-stakes placement, formative, and diagnostic tests may be appropriately administered as CCTs.

In its purest form, a CCT has no fixed test length. Instead, the test continues until either a classification decision is made or some maximum number of items has been administered (Lewis & Sheehan, 1990; Spray & Reckase, 1996). That test length should differ across students is best understood through the following example. Student A has a "true" performance level very near a classification threshold judged as dividing those who have mastered a concept from those who have not. Substantial evidence must therefore be collected before it can reliably be determined whether this student is just above or just below the threshold. Consider now Student B, whose performance level is far above the threshold. This fact will be clear based on performance on relatively few items, and reliable classification can be quickly made. This student is like a high jumper who easily cleared the bar that defined the threshold.

A CCT is pool-based, like the random forms and item-adaptive models. However, items are selected from the pool neither randomly nor to match the performance level of the student being classified. Rather, a CCT chooses an item that best targets the threshold most crucial to a given student's classification. Consider an example where students are to be classified as Basic, Proficient, or Advanced. For a student who is performing well on early items, the critical threshold would be that which divides Proficient from Advanced. This threshold would then be targeted by items chosen in the latter part of the test, until a classification can be reliably made. Conversely, the last items for a struggling student would focus on the threshold between Basic and Proficient. (See Table 5.)

Table 5

Computerized Classification Test

Efficiency	The CCT is by far the most efficient model should nothing more than classification of examinees be required. A variable-length CCT can easily exceed the efficiency of even an adaptive test. It does so by not making finer distinctions among students assigned the same classification.
Security	Like all CBTs, CCTs are more secure than conventional tests with respect to students copying from one another or inadvertent disclosure of test content. Like the other pool-based models (random forms and item-adaptive) the resistance of a CCT to item overexposure is dependent largely on pool size. However, because a CCT targets items at relatively few performance thresholds, it is somewhat more prone to overexposure than the item-adaptive model.
Item development requirements	As commonly implemented, the CCT requires item pools of sizes comparable to those needed by the random forms model (but smaller than what the item-adaptive model requires). However, a CCT is most efficient when based on items that accurately target the classification thresholds. Because it is difficult to precisely control difficulty during item development, a CCT may reject as unusable far more items than any other model. The total development requirements for CCT are then generally at least on par with those of the item-adaptive model. As always, higher security requirements increase the frequency with which item pools must be replaced or refreshed, increasing item development needs accordingly.
Complexity	A CCT is roughly as complex to develop, administer, and maintain as the item-adaptive model. Complicated test administration logic makes development of the test delivery software difficult. Advanced psychometric methods also are needed to ensure that different item pools produce comparable classifications.
Cost	As commonly implemented, a CCT would be about as expensive to develop and deliver as an item-adaptive test. This is largely the result of operational complexity and high item development requirements.
Comments	A CCT typically would not allow students to freely navigate by skipping forward or returning to check previously answered items. Skipping forward is particularly problematic, given that the response to the current item is needed in order to properly select subsequent items.

Making an Informed Decision

The clearest lesson that can be drawn from the above discussion is that no single administration model is ideal for all tests and under all circumstances. Instead, the model that is best for a given testing program depends upon the unique characteristics of that program. The item types required to test the construct, the stakes attached to test scores, the characteristics of the examinee population, and the subjective values of the test's owners and score users are all relevant considerations.

Table 6 is an attempt to assist test designers by rating each of the CBT models with respect to each of the five important test properties. It is important to note that these ratings are based on typical or average implementations of each model and can be strongly influenced by the unique combinations of circumstances that define any testing program.

The model that is best for a given testing program depends upon the unique characteristics of that program.

Table 6

Properties of Computer-Based Test Models

	Efficiency	**Security**[2] (for single form/pool)	**Item development requirements** (for single form/pool)	**Complexity**	**Cost**
Fixed form	Low	Low	Low	Low	Low
Random form	Low	Medium	Medium	Medium	Medium
MST	High	Medium	Medium	Low	Medium
Item-adaptive	High	Medium/high	High	High	High
CCT	Very high	Medium/high	High	High	High

[2] Test security has multiple facets that involve such diverse characteristics as item content and format, test administration circumstances and policies, the frequency with which new test forms or item pools are developed and deployed, and the mechanisms by which alternative forms or pools are tied together so that scores achieved on any can be considered as comparable. As such, the ratings assigned to test designs here are holistic in nature and are largely driven by the probability of different examinees seeing many of the same items in common across their tests.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater*® v 2.0. *Journal of Technology, Learning, and Assessment*, 4(3). Available from http://www.jtla.org

Bennett, R. E. (1998). *Reinventing assessment.* Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1). Available from http://www.jtla.org

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, 67, 5–20.

Cassady, J. C., & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *Journal of Technology, Learning, and Assessment*, 4. Available from http://www.jtla.org

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripkey, D. R., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement*, 32: 397–415.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191). Mahwah, NJ: Lawrence Erlbaum.

Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M. (1996). *Writing test items to evaluate higher order thinking.* Needham Heights, MA: Allyn & Bacon.

Harmes, J. C., & Parshall, C. G. (2010). *A model for planning, designing and developing innovative items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice* 20(3), 16–25.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 367–386.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Mahwah, NJ: Lawrence Erlbaum.

Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Norwell, MA: Kluwer.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). New York, NY: Springer.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. C. (2001). *Practical considerations in computer-based testing.* New York, NY: Springer.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, 21, 405–414.

van der Linden, W. J., & Glas, C. G. (2000). *Computerized adaptive testing: Theory and practice.* Norwell, MA: Kluwer.

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.

Listening. Learning. Leading.®

www.ets.org