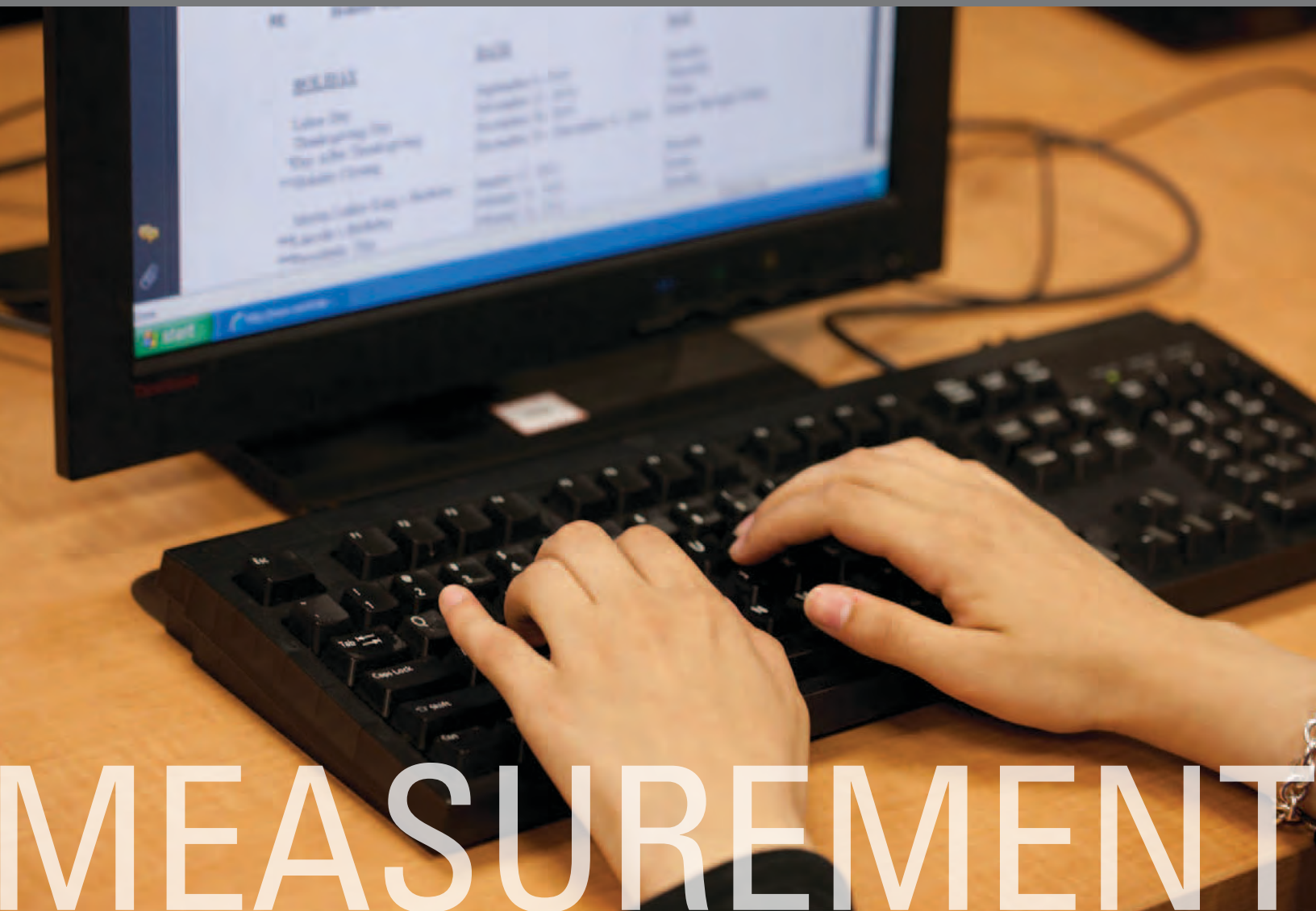
RESEARCH REPORT 2011-12

CollegeBoard

A Review of Models for Computer-Based Testing

By Richard M. Luecht and Stephen G. Sireci



MEASUREMENT

Richard M. Luecht is professor of Education Research Methodology at the University of North Carolina at Greensboro.

Stephen G. Sireci is Professor of Education and Co-Chairperson of the Research and Evaluation Methods Program and Director of the Center for Educational Assessment in the School of Education at the University of Massachusetts Amherst.

Mission Statement

The College Board's mission is to connect students to college success and opportunity. We are a not-for-profit membership organization committed to excellence and equity in education.

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of more than 5,900 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators and schools.

For further information, visit www.collegeboard.org.

For more information on College Board research and data, visit www.collegeboard.org/research.

Contents

Tables

Figures

Executive Summary

Over the past four decades, there has been incremental growth in computer-based testing (CBT) as a viable alternative to paper-and-pencil testing. However, the transition to CBT is neither easy nor inexpensive. As Drasgow, Luecht, and Bennett (2006) noted, many design engineering, test development, operations/logistics, and psychometric changes are required to develop a successful operational program. Early research on CBT almost exclusively focused on theoretical issues such as improving measurement efficiency by achieving adequate levels of test score reliability using as few items as possible. However, it was soon evident that practical issues — such as ensuring content representation, making sure all examinees have sufficient time to complete the test, implementation of new item types, and controlling the degree to which items were exposed to examinees — needed to be addressed, too. In the past few years, research on CBT has focused on developing models that achieve desired levels of measurement efficiency while simultaneously satisfying other important goals, such as minimizing item exposure and maintaining content validity. In addition, there has been a growing awareness among practitioners that basic CBT research using small samples or simulation studies needs to be vetted using cost-benefit analysis, as well as engineering design and implementation criteria to ensure that feasibility, scalability, and efficiency are evaluated in more concrete ways than by merely reporting a reduction of error variances for theoretical examinee scores (Luecht, 2005a, 2005b).

Today, CBT is a broad-based industry that encompasses a large variety of assessment types, purposes, test delivery designs, and item types appropriated for educational accountability and achievement testing, college and graduate admission testing, professional certification and licensure testing, psychological testing, intelligence testing, language testing, employment testing, adult education, military use. The delivery of CBT has also undergone many changes from the early days of "dumb" terminals connected to a mainframe or minicomputer. CBTs can be administered on networked PC workstations, personal computers (PCs), laptops, netbooks, and even hand-held devices such as smart phones and tablet computers. Testing locations or sites include dedicated CBT centers, classrooms or computer labs in schools, colleges, and universities; temporary CBT testing facilities set up at auditoriums, hotels, or other large meeting sites; and even personalized testing in the privacy of one's home, using a PC with an Internet connection and an online proctoring service. The items or assessment tasks available for CBT include many variants of simple multiple-choice or selected-response formats, constructed- or extended-response items, essays, technology-enhanced items using novel response-capturing mechanisms involving a mouse or another input device, and complex, computerized performance exercises that may simulate real-life tasks using synthetic challenges made engaging through virtual realism. Test assembly and delivery formats also vary widely and may include preconstructed test forms (i.e., where everybody sees the same items), test forms constructed in real time, or many varieties of computer-adaptive tests (CATs) that tailor the difficulty of each test form to the proficiency of every examinee. Simply put, CBT is not constrained to a particular technology platform, item type or test design; instead, it is a growing collection of technologies and systems that serve many different test purposes, constituencies, and test-taker populations (Luecht, 2005a; Drasgow, Luecht, & Bennett, 2006).

Modern CBT can be implemented in any of five ways: (a) on a stand-alone personal computer (PC); (b) in dedicated CBT centers; (c) at temporary test centers; (d) in multipurpose computer labs; or (e) using a PC, laptop, netbook, tablet, or hand-held device connected to the Internet, possibly remotely proctored. With the exception of using stand-alone PCs, CBT usually requires some level of connectivity, with the most successful implementations having the capability to link multiple computers to the test delivery software and item banks, and to

rapidly transmit test materials, results, scores, and other information where and when they are needed. The earliest CBTs consisted of testing terminals physically connected to a mainframe computer. The mainframe computer did all of the processing; the workstations merely displayed the information on screen and collected responses via a keyboard. The advent of personal computers and local area networks (LANs) made it possible to connect stand-alone microcomputers — that is, smart terminals capable of handling some or all of the processing — to centralized storage file servers and shared processing resources. Wide area networks expanded this connectivity principle to allow remote networks to be connected across multiple physical locations.

The development of the Internet TCP/IP (packet switching protocols) widened networking capabilities in the mid 1990s. TCP/IP offered an open networking architecture that could rather seamlessly communicate and share resources and data across operating systems and computing platforms. Since the late 1990s, the introduction of virtualization, cloud computing, advances in Internet browser capabilities, dramatic improvements in routing and switching technologies, and high-speed wireless connectivity have removed most practical barriers to networking and cross-platform computing anywhere in the world. However, that does not necessarily mean that CBT is now possible anywhere in the world, on demand.

A Brief History of CBT

One of the first large-scale computerized-adaptive testing programs to go operational was the College Board's ACCUPLACER® testing program, which in 1985 consisted of four tests: Reading Comprehension, Sentence Skills, Arithmetic, and Elementary Algebra. These examinations were introduced to assist in placing entering college students in English and mathematics courses. Thus, it was a relatively low-stakes test. The first high-stakes CAT was the Novell corporation's certified network engineer (CNE) examination. The CNE went online at Drake Prometric testing centers in 1990 and transitioned to online CAT in 1991 (Foster, 2011). The CNE was followed by Education Testing Service's (ETS) Graduate Record Examination (GRE), which was operationally deployed as a CAT at Sylvan testing centers across the U.S. as of 1992 (Eignor et al., 1993; Mills & Stocking, 1996). Two NCLEX examinations for nurse candidates were implemented using a CAT format at commercial testing centers in 1994 (Zara, 1994). A CAT version of the Armed Services Vocational Aptitude Battery (ASVAB) went online at Military Entrance Processing Stations (Sands, Waters, & McBride, 1997).

In addition to these programs, the Graduate Management Admission Council implemented a CAT version of the GMAT in 1997. The Architect Registration Examination (ARE) was also rolled out in 1997, offering interactive, computer-aided architectural problems within a custom graphical interface. The architect examination was followed by the United States Medical Licensing Examination (USMLE) transitioning to CBT in 1999. This exam incorporated highly interactive computerized patient-management simulations to the examinations (Clymer, Melnick, & Clauser, 2005; Dillon, Clyman, Clauser, & Margolis, 2002). Interactive accounting simulations were added to the Uniform CPA Examination in 2004 (Devore, 2002; Luecht, 2002a, 2002b) and implemented one of the first computer-adaptive multistage testing frameworks for large-scale applications (Luecht, Brumfield, & Breithaupt, 2006; Melican, Breithaupt, & Zhang, 2010; Breithaupt, Ariel, & Hare, 2010). These CBT programs are just examples of the numerous CBTs administered in the licensure and certification arena.

CBT is also used for many types of psychological and employment tests. Many states are now offering CBT options for end-of-grade, end-of-course, and high school graduation, with an even more expansive use of CBT within schools planned under the U.S. government's

Race to the Top (RTTT) initiatives (www2.ed.gov/programs/racetothetop). Examples of current statewide CBT testing programs can be found in Kansas, Oregon, Texas, and Virginia.

Many of the earliest computer-based tests quickly jumped on the computerized adaptive testing (CAT) bandwagon — capitalizing on faster computers and network technologies, and hoping to fulfill the promises of more accuracy with shorter tests (compared to paper-and-pencil versions). Other testing programs, such as the Physical Therapist licensure exam (https://www.fsbpt.org/ForCandidatesAndLicensees/NPTE/FAQs/index.asp), decided to use multiple, fixed test forms, essentially mimicking paper-and-pencil test forms. More recently, many organizations are considering a practical hybrid known as computer-adaptive multistage testing (ca-MST), which combines many of the quality control benefits of fixed test forms with the adaptive efficiencies of CAT. For example, in adult education, the multistage-adaptive Massachusetts Adult Proficiency Tests became operational in 2006 (Sireci et al., 2006). This report highlights some of the more concrete differences between these different CBT delivery models.

One thing is clear from the CBT research: There is no single CBT model that is ideal for all educational tests. Rather, all models have their strengths and weaknesses, and some are better suited to the characteristics of a particular testing program than others. Recent research has also shown the advantages and limitations of particular CBT models. The purpose of this report is to review the most popular CBT models that should be considered by the College Board as it moves toward computerization of all of its testing programs. Our selection of the specific CBT models reviewed here is based on our opinions regarding models that show the most promise and are most likely to be applicable to College Board exams. In the next section, we provide an overview of some of the major technologies used in CBT. This technological overview is intended to provide background information relevant to various features of the CBT models. We follow that with a presentation of eight CBT models. For each model, we provide a brief description of its critical features and we review the relevant research literature on its functioning.

Technological Overview

Models for delivering computer-based tests vary in their complexity. It is important to understand certain aspects of complexity in order to evaluate particular features of the eight CBT models presented in this report. This section briefly reviews five issues that help define what we mean by "complexity": (1) the degree and nature of test adaptation; (2) size and flexible units of test administration; (3) user interface issues; (4) automated test assembly and test form quality controls; and (5) security risks. We also summarize some of the relevant considerations and criteria for comparing CBT models with respect to each of these five issues.

Degree and Nature of Test Adaptation

A fundamental technology that distinguishes among many CBT models is the degree to which the test is made *adaptive*. The basic mechanism behind an adaptive test is relatively simple. An adaptive test tailors the difficulty of the test items to the apparent ability or proficiency of each examinee. The specific goal in a purely adaptive test is to maximize the test reliability (score precision) for every examinee, regardless of his or her score. Items that are too easy or too difficult for particular examinees add little to the reliability of their scores. By tailoring the difficulty of the items to the ability of a particular examinee, it can be shown that we are indeed maximizing the reliability of the test score. Tailored or adaptive testing therefore leads

to certain measurement efficiencies where a particular level of reliability can be achieved with fewer items. That is, score precision can be improved relative to a nonadaptive fixed-length test.

Birnbaum (1968) introduced the concept of the "test information function" as a psychometric analysis mechanism for designing and comparing the measurement precision of tests in the context of item response theory (IRT). Under IRT, the conditional measurement error variance, $\mathrm{var}\left(E|\theta\right)$, is inversely proportional to the test information function, $I(\theta)$. That is,

$$\mathrm{var}\left(E|\theta\right) = I\left(\theta\right)^{-1} = \sum_{i=1}^{n} \frac{\partial P_i\left(\theta\right) / \partial\theta}{P_i\left(\theta\right)\ 1-P_i\left(\theta\right)}^{2\ \ -1} = \frac{1}{\sum_{i=1}^{n} I_i\left(\theta\right)} \qquad \text{(Equation 1)}$$

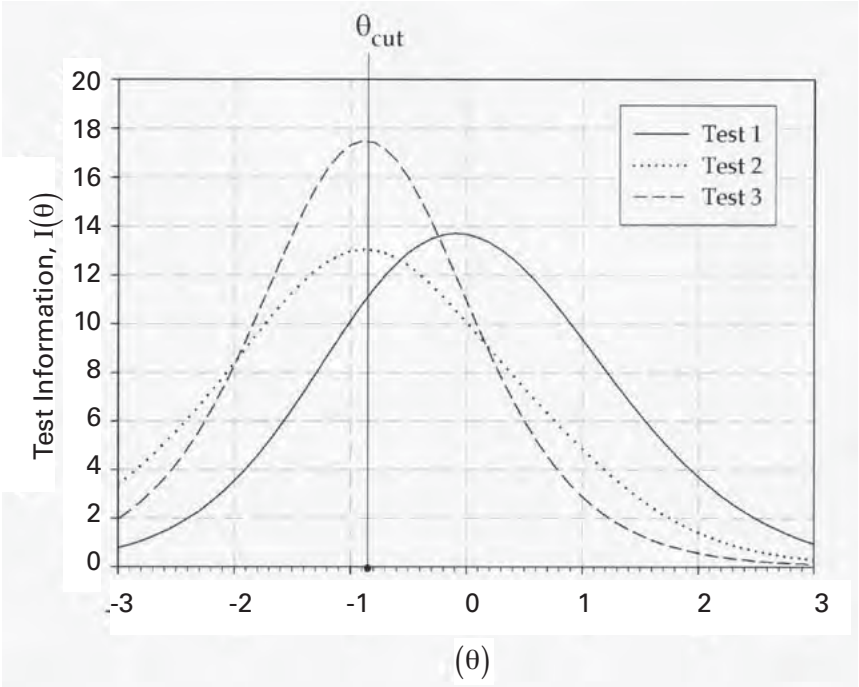
where $I_i\left(\theta\right)$ is the item information function at some proficiency score of interest, denoted as θ (Lord, 1980). The exact mathematical form of the information function varies for different IRT models (one-, two-, or three-parameter logistic, partial-credit, or graded-response models).

Equation 1 suggests two important aspects about measurement precision. First, **each item contributes some amount of measurement information** to the reliability or score precision of the total test. That is, the total *test information function* (TIF) is the sum of the item information functions. Second, by **increasing the test information function**, we correspondingly **reduce the measurement error variance** of the estimated θ score. Simply put, when test information is maximized, measurement errors are minimized. We can achieve maximum test information in two ways. We can choose highly discriminating items that provide **maximum** item information within particular regions of the proficiency scale or at specific proficiency scores. Or, we can merely continue adding items to increase the amount of information until a desired level of precision is achieved.

The amount of information varies at different levels of proficiency, as indicated by the three TIF curves shown in Figure 1. The location of the peak of each curve indicates where along the proficiency score scale, θ, the test form is most precise. The height of each curve indicates the amount of precision. For example, Test 2 and Test 3 both peak near $\theta = -1.0$ on the proficiency scale, but Test 3 has more information at that peak and for most proficiency scores. Test 1 has about the same amount of information as Test 2, but most of the measurement information is near $\theta = 0.0$ on the score scale. There is usually less information near the tails of the proficiency scale.

A certification or licensure test will typically have a passing score set along the scale. We would therefore like the peak of the TIF to be located near that passing cut score (see Test 1 in Figure 1). In contrast, an achievement test or any test that primarily reports scores along the entire scale might prefer to amass most of the measurement information either near the mean of the population — assuming that the greatest concentration of proficiency scores occurs in the vicinity of the mean — or more uniformly spread out across the scale.

Figure 1
Test information function (TIF) curves for three tests



Maximizing the test information at each examinee's score is tantamount to choosing a customized, optimally reliable test for each examinee. Lord (1977) introduced the idea of using the IRT maximum information criterion as an item selection mechanism to tailor a test for every examinee. The adaptive process involves randomly or by some defined mechanism selecting a starting item or a small number of items to administer to an examinee. The examinee answers the item(s), the computer scores the items, and a provisional score is estimated. The computer then uses an IRT model to select the most informative item at the examinee's current provisional proficiency score. Each item adds to the information accumulated about the examinee's proficiency score (see Equation 1). The adaptive test proceeds until a particular stopping rule is satisfied. Two standard stopping rules for adaptive tests are: (1) a fixed test length has been met or (2) a minimum level of score precision has been satisfied.[1]

Since the 1970s, a plethora of somewhat more complex adaptive strategies have emerged, including adapting on item sets or modules (Adema, 1990; Sheehan & Lewis, 1992; Luecht, Nungester, & Hadadi, 1996; Mills, Potenza, Fremer, & Ward, 2002; van der Linden & Glas, 2000; Wainer & Kiely, 1987; Wainer & Lewis, 1990, Wainer, Kaplan, & Lewis, 1992); using sophisticated item selection heuristics for balancing content and other test features (Kingsbury & Zara, 1991; Stocking & Swanson, 1993; van der Linden & Reese, 1998; van der Linden, 2000, 2005), using stratification schemes to block on item characteristics — effectively using less discriminating items earlier in an adaptive test (Chang & Ying, 1997, 1999; Chang & van der Linden, 2000; Chang, Qian, & Ying, 2001), selecting items with information proportional to the error variance of provisional proficiency scores (Luecht, 1995); stochastically controlling the exposure of test materials within the population (Sympson

1. For pass/fail mastery tests that are typically used in certification and licensure testing, a different stopping rule can be implemented related to the desired statistical confidence in the accuracy of the classification decision(s).

& Hetter, 1985; Stocking & Lewis, 1995, 1998, 2000; Davey & Parshall, 1995), and even simultaneously estimating multiple abilities or proficiencies using multidimensional IRT adaptive testing algorithms (Segall, 1996, 2010; Luecht, 1996).

Despite the increased sophistication of modern adaptive algorithms, the central goal of the algorithm is still to either maximize the test information function or to achieve a targeted amount of information at various points of the proficiency scale. If we seek the most information possible, we call that "maximizing" the information. If we seek to achieve a prescribed amount of information, we call that "targeting." This latter targeting approach is used for some types of multistage adaptive tests and for nonadaptive tests, including computerized fixed tests.

Measurement Efficiency Through Adaptive Testing

Two standard psychometrically oriented considerations associated with computerized adaptive tests are: (1) relative efficiency; and (2) reductions in test length. These considerations are related to one another.

Relative efficiency refers to a proportional improvement in test information (score precision) and can be computed as the ratio of test information functions or reciprocal error variances for two tests (see Equation 1; also see Lord, 1980). Furthermore, this relative efficiency metric can be applied to improvements in the accuracy of proficiency scores or to decision accuracy in the context of mastery tests or certification/licensure tests. For example, if the average test information function for a fixed-item test is 10.0 and the average test information function for an adaptive test is 15.0, the adaptive test is said to be 150% as efficient as the fixed-item test.

Relative efficiency depends on two factors: The first factor is the "baseline" test information function (see Equation 1) being used for comparison. The baseline test information function may be computed from an existing fixed-item test form. Optionally, a test information baseline could also represent the *maximally* informative test that can be drawn from a particular item pool. The second factor is the location along the proficiency scale where greater efficiency is desired. A test that is more efficient in one region of the proficiency scale may be less efficient elsewhere. When adaptive tests are compared to fixed-item tests (e.g., see Figure 1), most of the efficiency gains are realized near the tails of the proficiency distribution, where the fixed-item test has little information.

Measurement efficiency is also associated with **reductions in test length**. For example, if a 20-item adaptive test can provide the same precision as a 40-item nonadaptive test, there is an obvious reduction in the amount of test materials and testing time needed (assuming, of course, that a shorter test would take substantially less time than a longer test). Much of the early adaptive testing research reported that typical fixed-length academic achievement tests used could be reduced by half by moving to a computerized adaptive test (Wainer, 1993). Unfortunately, that early research ignored the perceptions of some test users — especially in high-stakes testing circles — that short adaptive tests containing only 10 or 20 items could not adequately cover enough content to make valid decisions or uses of scores. Modern adaptive tests typically avoid such criticism by using either fixed lengths or at least some minimum test length to ensure basic content coverage. As a result, reported measurement efficiency gains tend to be less dramatic than we once believed possible.

Although adaptation is clearly important as a psychometric criterion, it is easy sometimes to overstate the real cost-reduction benefits that can be specifically attributed to gains in measurement efficiency. For example, measurement efficiency gains from adaptive testing

are often equated with reduced testing time. However, any potential savings in testing time may prove to be unimportant if a computer-based examination is administered at commercial CBT centers. That is, commercial CBT centers typically charge fixed hourly rates per examinee and require a guaranteed [minimum] testing time. Therefore, if the CBT test center vendor negotiates with the test developer for a four-hour test, the same fee may be charged whether the examinee is at the center for two, three, or four hours.

Size and Flexible Units of Test Administration

A single test question (item) is often thought of as the fundamental unit of test administration (i.e., one item = one test administration object). However, test administration units can be small or large. For example, sets of items assigned to a reading passage or to a particular problem scenario can also be packaged to present as a single unit or module. In fact, any cluster of items can be preconstructed and packaged as a unique test administration unit or module. By extension, modules can also be grouped into fairly large, discrete test administration units (e.g., subtests or test sections).

The terms "testlets" or "modules" are two common ways to refer to preconstructed sets of items that are packaged and administered as intact units. Although the term "testlets" is sometimes explicitly associated with layered-problem item sets that may involve internal branching (Wainer & Kiely, 1987) or sometimes with computerized mastery tests using multi-item modules (Sheehan & Lewis, 1992), there is nothing precluding the use of that label in almost any type of CBT model to describe a preconstructed test module that includes everything ranging from a cluster of items to a set of computerized performance exercises (Luecht & Nungester, 1998; Luecht, 2002a; Luecht et al., 2006). The items in each modularized unit can be linked to a common passage, graphic, or scenario; grouped by content; or merely formed by organizing some number of discrete items into a cluster. In addition to providing better control over test content, modules allow for more authentic assessments in many contexts such as reading passages or problem-solving vignettes.

As unique test administration units, testlets can be selected and administered to examinees by a variety of mechanisms, including random selection from a pool of testlets, sequentially from a list, or by an adaptive algorithm. Some of the multistage adaptive test models described further on specifically combine the modular features of testlets with adaptive selection and scoring mechanisms.

In terms of examinee perspectives, surveys of examinees have consistently reported that examinees unilaterally prefer being able to navigate through an examination section, rather than being forced to answer a single item at a time, with no chance to review and change previously answered items, or to skip ahead (Wise, 1996; Hadadi et al., 1998; Parshall, Spray, Kalohn, & Davey 2002). This would argue for larger, rather than smaller, testlets. However, there are also timing and pacing issues to consider. For example, in field studies related to computerized versions of the United States Medical Licensing Examination, Hadadi et al. (1998) indicated that lower proficiency examinees may be penalized for their lack of pacing skills on lengthy test sections and may be able to better pace themselves on moderate-sized modules. There is a growing body of evidence suggesting strong examinee preferences for controlling their test-taking by reviewing and changing answers (Zenisky, Hambleton, & Luecht, 2010; Melican, Breithaupt, & Zhang, 2010). In fact, examinee preference to be able to review and change answers may be one reason why the GRE is transitioning to a computer-adaptive multistage testing model in 2011-12 and abandoning CAT (ETS, 2011).

From a database control perspective, creating uniquely identified, hierarchically related "structured data objects" (test forms, testlets, or modules) is an efficient way to manage test data. Modern CBT requires enormous amounts of data to be moved, usually on a near-continuous basis. For example, 10,000 examinees taking a 50-item computer-based test will generate 500,000 response records (item answers, response times, etc.). Despite the tremendous improvements in data encryption, transmission, and database management technologies over the past decade, there is always some potential for errors related to data distortion and corruption, broken or faulty data links, or general programming faults in the data management system(s). Eliminating errors is the ultimate goal, however, the ideal (completely error-free data) cannot be achieved in practice. The point is that numerous quality control and quality assurance procedures are necessary at different points in time to either reduce the likelihood of data errors (prevention) or at least to identify errors when they occur (detection). In virtually any database management situation, *structure reduces error*! If more structure can be imposed on the data, fewer errors are likely because preventative measures are easier to implement. And when errors do occur, it is easier to detect them in highly structured data than in less-structured data (Luecht, 2000, 2005a, 2005b, 2005d).

Using larger units is also an advantage in terms of navigation and presentation. That is, commonly used item-rendering properties can be stored at the structured module level and inherited by the individual test items (or other subunits) within each module[2]. This leads to improved efficiency and accuracy in rendering test materials. The integrity of the test materials and subsequent response data are also easier to manage with structured units (modularization) because the test unit data can be checked against known control parameters. By contrast, the data for computerized tests constructed in real-time® (randomly selected "on the fly" tests or adaptive tests) cannot be easily checked for integrity or reconciled to any known units because each test is a unique creation.

It is important to realize there is no magical size that qualifies as the optimal test administration unit size. Intermediate test administration units such as testlets are indeed easier to handle from a data management perspective, and examinees seem to prefer them. However, some amount of flexibility and mobility is always sacrificed through consolidation. The trade-offs largely depend on the choices of costs and benefits, some of which may be indirect and even intangible (e.g., perceptions of fairness, trust, or integrity by the test users). In any case, a CBT model that restricts the test administration unit size to a single item or a fixed-size testlet or module may be overly restrictive in terms of future flexibility.

User Interface Issues

User interface issues include considerations about the design of the test or software that simply affect how examinees take the test, aspects of the software or test design that have some direct effect on examinee performance, and test or software design factors that directly affect examinee perceptions (but not necessarily their performance). Although there are a myriad of issues to consider, we have elected to focus on two that seem germane to this review: (1) timing or pacing issues and (2) navigation.

Timing or Pacing Issues

Speededness is one of the more problematic aspects of any time-limited, standardized test (Cronbach & Warrington, 1951). The fact that some examinees either fail to reach certain items on speeded tests, or are induced to engage in "rapid guessing" behaviors, is troubling to most test developers and psychometricians, and is certainly a serious concern for most

2. Modern object-oriented databases and programming languages deal specifically with hierarchically related "objects" in a very straightforward manner.

examinees. Unlike paper-and-pencil tests, where it is impossible to distinguish between intentionally omitted and not-reached items, the speededness problem is easily detectable under CBT (Hadadi & Luecht, 1998; Bridgeman & Cline, 2004; van der Linden, Breithaupt, Chuah, & Zhang, 2007).

Under CBT, where the tests are often administered at commercial test centers, the speededness problem can be exacerbated by the fact that seat time directly contributes to the cost of the examination. Where time equals money, policies may be made that tend to minimize rather than maximize the time allotted for examinees to take the examination.

Empirical (real data) studies of speededness using actual item-by-item response times are rare (e.g., Swanson, Featherman, Case, Luecht, & Nungester, 1997; Hadadi & Luecht, 1998; Bridgeman & Cline, 2004; van der Linden et al., 2007). For tests, with highly restrictive time limits, there tends to be a moderate to high degree of correlation between the time spent on test items and the difficulty of the items (van der Linden, Scrams, and Schnipke, 1999). As a result, tests that provide differentially difficult items for low, medium, and high proficiency examinees — which include most types of adaptive tests — may be speeded for higher-ability examinees (van der Linden, 2000). At the same time, empirical studies conducted with medical students have shown that higher ability examinees may pace themselves better than lower-ability examinees (Swanson et al.,1997; Hadadi & Luecht, 1998).

Pacing aids can range from online clocks to "pacer mechanisms" to designing the test in a way that facilitates pacing for most examinees. For example, Hadadi, Luecht, Swanson, and Case (1998) empirically showed that using reasonably sized modules as the basic test administration units facilitated all examinees and specifically helped the lower ability examinees whose pacing skills on a timed, multiple-choice CBT appeared to be underdeveloped. Automated test-assembly procedures (see the Automated Test Assembly and Test-Form Quality Control section in this report) can also be used to reduce speededness (van der Linden, 1998; van der Linden, Scrams, & Schnipke, 1999).

Navigation

Navigation relates to how the examinee moves around in a test. There are two aspects to navigation: (1) visual style of the navigation control and (2) blocking review and/or changing answers to previously seen items.

The design and visual style of navigation software controls differs across test delivery drivers (and sometimes across test delivery system platforms). Every test has some navigation mechanisms. Some CBT test delivery drivers merely use "forward/next" and "back" keys or mouse-clickable buttons to move item by item. Other CBT test drivers include a "jump" control that allows the examinee to enter in the number of a test item and immediately go to that item. Some CBT test delivery drivers provide a full-page "review screen" to display all of the items, the examinee's item response (if any), and any flags left by the examinee indicating that he or she wants to possibly review the item later. Many of the recent genres of CBT graphical user interfaces now provide an "explorer" or "helm" style of navigation control that continuously shows an ordered list of the test items in a narrow scrollable window within some segment of the display screen. This ordered list format is particularly helpful to examinees who want to skip items and go back to them later, if time permits.

The style of the navigation control can affect performance and/or examinee perceptions in positive or negative ways. For example, some of the early CBT review screens completely covered up the test item display when the examinee selected the on-screen "review" button. Inexperienced CBT examinees sometimes panicked because they believed that the computer

had lost their test. Panic during an already anxiety-provoking testing session is definitely an undesirable emotion to induce.

The blocking of review and/or answer changes is typically implemented in terms of the navigation control(s). Review simply means that items can navigate to previously seen materials or skip ahead to view upcoming materials. The "no review/no changing answers" aspect of item-by-item adaptive testing is strongly criticized by most examinees (Wise, 1997; Pommerich & Burden, 2000). However, if examinees are allowed to revise their answers to previous items in a CAT, there are ways they can game the system (Wainer, 1993). When review and answer changes are precluded, examinees report that they felt overly restricted in their test-taking strategies. Some of the multi-item, modular test designs, discussed further on, attempt to avoid that criticism by allowing the examinees to freely navigate and change answers within the testlet or module. The examinees are required to electronically submit their testlet or module (analogous to physically turning in a paper-and-pencil test booklet and answer sheet at the end of a test section). Once a testlet or module is submitted, the examinee is prohibited from revisiting it. What is not clear from the research is how small the testlets or modules can be before examinees perceive that their choices in test-taking strategies are limited.

Automated Test Assembly and Test-Form Quality Control

Test assembly implies selecting items (usually by some principled means) to comprise individual test forms and composing all of the associated item and test data needed to administer the items within a computer-based test delivery system. For CBT, these data are stored in a test bank. At one extreme, all aspects of test assembly must be performed before the test administration takes place. In short, the test bank contains intact test forms. At the other extreme, all test assembly is performed in real time, at the testing center, either immediately before or while the examinee is taking the examination. In this latter case, the test bank merely contains the item components and data needed for assembly. The online test delivery software must handle the actual test assembly. The more trust we place in the test assembly software the more we are "automating" the process. In the context of CBT, automation is often a matter of degree.

Although measurement precision is a primary psychometric goal, the quality of test forms in terms of content validity and other criteria are often equally important. Any professionally developed examination has a table of specifications or "blueprint" that defines the content areas and other relevant attributes that must be covered on every test form (e.g., 10 to 15 items in intermediate algebra). These specifications may be broad or very specific and can include multiple coded taxonomies (levels of a content outline, cognitive levels, settings, themes, etc.).

Unfortunately, these types of content validity goals often compete with the measurement efficiency goals, given the availability of items in the pool (Stocking & Swanson, 1993; Mills & Stocking, 1996; Luecht, 2000; van der Linden, 2005). As a result, trade-offs are required. The trade-offs become more severe as the demand increases for items meeting critical content goals or critical measurement efficiency goals increases, especially under continuous or near continuous testing. To help deal with the trade-offs in a systematic way, most testing programs now employ automated test assembly (ATA) heuristics or algorithms as a core technology within their test development software systems and operations.

ATA item selection mechanisms involve the use of formal mathematical optimization procedures that include linear programming algorithms, network flow algorithms, and various

greedy-type heuristics (van der Linden & Boekkooi-Timminga, 1989; van der Linden & Adema, 1989; Luecht & Hirsch, 1992; Armstrong, Jones & Wu, 1992; Swanson & Stocking, 1993; van der Linden, 1998; Sanders & Veerschoor, 1998; Luecht, 1998, 2000; Armstrong, Jones & Kunce, 1998; Berger, 1998; van der Linden, 1998; van der Linden & Reese, 1998; Stocking & Swanson, 1998; Timminga, 1998; van der Linden, 2005). Using these algorithms or heuristics, ATA seeks to satisfy a particular mathematical goal, called the "objective function," subject to any number of constraints on content and other test item attributes. For example, ATA algorithms can build a test form to achieve a particular level of test difficulty (e.g., every test form should have an average difficulty of 65% correct, subject to also meeting almost any number of content requirements or other test specifications such as minimum test length, word counts, reading levels, statistical impact on minority and majority groups, DIF, enemy items that cannot appear on the same test form, etc.). The latter are called "constraints" in the ATA literature.

The real power of ATA is realized for large-scale test production enterprises. For example, what if we want to generate 100 test forms, each containing 25 items that jointly meet exactly the same content constraints? This could take weeks or months for human test editors to accomplish and the test forms may or may not uniformly achieve an acceptable level of statistical and content comparability. Simply put, when hundreds or thousands of tests need to be generated from a fixed resource — the item pool or test bank — ATA becomes a necessity.

The technical aspects of ATA are fairly well developed for preconstructed test units (fixed-length test forms, testlets or modules). Test assembly heuristics like the weighted deviations model (Swanson & Stocking, 1993) and the normalized weighted absolute deviations heuristic (Luecht, 1998b, 2000) have also been used for linear-on-the-fly tests (LOFT) at Prometric Inc. and for CAT at the Educational Testing Service (ETS) for some time. As ATA has been popularized and demonstrated at national and international assessment conferences, and with the introduction of van der Linden's (2005) book on the subject, an increasing number of operational CBT delivery systems have begun to integrate ATA capabilities into their operational test development and production systems — at least for preconstructed test forms. The use of advanced mixed integer optimization algorithms or other formal ATA heuristics in real time to meet complex objective functions and test specifications for online, live testing is far less prevalent given the lack of sustainable, high-speed connectivity via the Internet. These connectivity and transmission speed limitations have undoubtedly stymied the operational use of some promising CBT models that require a high degree of interactive computing between a local network or workstations and a central processing facility. For example, one of the models discussed further on, called "shadow testing" (van der Linden & Reese, 1998; van der Linden, 2000, 2005), has not been operationally implemented to date because of a lack of integrated test delivery software. In large part, this is due to the high costs and time to design and build the needed integrated ATA computer subsystems. That may change in the future. But for now, most existing, real-time CBT delivery systems are only able to engage in a very rudimentary content balancing (e.g., selecting items randomly from within discrete content categories).

The quality control (QC) and quality assurance (QA) aspects of test form composition and production are nontrivial issues for many test development experts. Even using ATA does not guarantee that an absolute quality standard is met for every test form. ATA can certainly help satisfy the tangible test specifications that can be coded or computed, stored in a database, and quantified for purposes of solving a particular test construction optimization problem. However, ATA cannot deal well with qualitative considerations, aesthetics, or fuzzy

specifications that human test content experts may consider in addition to the formal test specifications.

In the world of paper-and-pencil testing, many testing organizations make extensive use of expert content committees to conduct a thorough quality control review and approve the final items on every test form. This can be very costly in terms of bringing the committees together to review one or two test forms. Furthermore, problems still arise, even following extensive human review (e.g., miskeyed answers, missing or incorrect pictorial materials associated with items, typos). In the CBT world, where there may be hundreds or thousands of intact test forms produced, carrying out test committee reviews for every test form is impossible. Worse, the potential for errors may become exponential. This is especially true for CBT models like linear-on-the-fly and computerized adaptive tests that rely entirely on real-time item selection and test assembly during the live examination. However, if a problem such as a miskeyed item is found, it can be immediately fixed without the need to reprint test forms.

Although it is not feasible to employ much of any physical QC review for tests generated in real time, there at least need to be QA procedures in place. This may involve building QA acceptability models to flag and discard potentially problematic items and test forms, before they are administered. Some organizations use simulated test administrations (i.e., computer-generated examinees and IRT model-based responses that fit a particular model) as a type of QA. However, those types of simulations fall short insofar as catching common typographical, referencing, and other test packaging errors. The empirical research on effective QA in large-scale CBT is conspicuously sparse.

Preconstructed, computerized fixed tests have a distinct advantage in terms of QC because every form can be checked or at least sample audited. Some adaptive CBT models, like computer-adaptive multistage testing (Luecht & Nungester, 1998; Luecht, 2000; Melican et al., 2010; Zenisky et al., 2010; Sireci et al., 2006), preconstruct and prepackage all of the pieces of a multistage adaptive test beforehand. By preconstructing and prepackaging any adaptive test, it is possible to engage in formal QC data checks and audit reviews — up to a 100% QC audit of all test forms before release.

From a QA/QC perspective, a key element of test assembly is where the item selections and test assembly take place. If test units can be preconstructed, more quality control is possible. Conversely, if test assembly is performed in real time, using ATA algorithms or heuristics that are incorporated into the test-delivery software, quality control may be largely nonexistent. Theoretically, if the test-bank or item pool is thoroughly checked before it is activated and if the computerized test delivery software and associated algorithms are fully tested and found to be robust under all potential problem scenarios, and if all data references for interactions between the examinees and the items are logged without error, additional QC may not be necessary. However, few if any CBT programs consistently meet these conditions on an ongoing basis and many QC/QA errors probably go undetected, altogether.

Security Risks

One of the most important CBT implementation issues for high stakes examinations is item-pool exposure (Haynie & Way, 1994; Stocking, 1993). The inherent flexibility of offering CBT on-demand or over a wide range of test dates potentially exposes the item pools to both small-scale and large-scale efforts aimed at cheating. That is, realizing that item pools may remain active over an extended period of time, examinees can conspire to memorize items, their intent being to reconstruct as much of the pool as possible to advantage retakers or to share with future first-taker examinees. The ease of communications over the Internet

further widens the potential scope of efforts aimed at cheating. Examinees need not even be in the same city or country to share information about the test. Luecht (1998a) termed these "examinee collaboration networks"; that is, collaborative groups formed for the sole purpose of recovering and sharing a large portion of active item pools for high-stakes examinations. Test developers have only recently begun to take measures to deal with some very real threats to the integrity of their testing programs posed by item-pool exposure collaboration and other forms of cheating.

The risks to the security of computer-based tests are somewhat analogous to the cheating threats faced by gambling casinos or lotteries. Given any type of high stakes (e.g., entrance into graduate school, scholarships, a coveted course placement, a job, a license, a professional certificate), there will be some group of cheaters intent on "beating the odds" (of random chance or luck) by employing well-thought-out strategies that provide them with any possible advantage, however slight that may be.

There are four general methods for dealing with risks in high-stakes CBT: (i) using randomization schemes to scramble items and other "test units" as much as possible; (ii) increasing the size of active item pools; (iii) rotating item pools (intact or partially) over time; and (iv) specifically controlling item exposures as part of the computerized test assembly process (e.g., using item-level exposure control mechanisms, see Hetter & Sympson, 1997; Sympson & Hetter, 1985; Stocking & Lewis, 1995, 1998; Revuela & Ponsoda, 1998). These methods each deal with particular types of risks, often in fundamentally different ways. For example, a randomly selected fixed-item test form has exposure controls implicitly built into the item selections. In contrast, an adaptive test typically requires more elaborate exposure controls to counteract the tendency to consistently choose the same highly discriminating items. Preconstructed test forms, including some of the multistage adaptive testing models, actually build the exposure controls into the test assembly process by controlling the amount of item overlap allowed across test units and by creating many test units. Simple random sampling is then used to select the preconstructed test units from a larger set of available units.

In high-stakes testing environments, the best test-delivery models are those that minimize the greatest number of risks and simultaneously reduce the magnitude of specific security risks, all without requiring extensive sacrifices or trade-offs elsewhere and without substantially adding to overall costs. In the next section, we review current models for delivering computer-based tests.

A Review of Computer-Based Test-Delivery Models

In this report, we distinguish among eight CBT models. These eight models differ primarily with respect to their use of adaptive algorithms, the size of the test administration units, and the nature and extent to which automated test assembly is used. In reviewing these models, we evaluate them with respect to several criteria including measurement efficiency, ability to ensure content balance and other test form quality aspects, risk considerations related to data management, item-pool usage, ease of implementation, and performance within large-scale, secure testing networks (including Web-based testing).

Preassembled Parallel, Computerized Fixed-Test Forms

This category of computer-based tests includes preconstructed, intact test forms that are administered by computer to large numbers of students (i.e., preassembled test forms). Different examinees may see different forms of the test, but all examinees administered the

same form will see exactly the same items (i.e., the items are fixed for each form). Parshall et al. (2002) describe these models as *computerized fixed tests* (CFT). One example of a CFT is the Physical Therapist Licensing exam. In the typical implementation of this model, several [or many] test forms (of the same fixed length) are available for administration and one is [randomly] selected for each examinee. The different forms are parallel with respect to test content and are either formally equated for difficulty (using classical test theory or item response theory) or are assumed to be randomly equivalent. A CFT is directly analogous to having fixed-item paper-and-pencil test forms (PPT). Some CFTs allow overlap among the items in different forms, although this strategy increases the exposure of those items.

One advantage of a CFT over a PPT is that the presentation sequence for the items may be scrambled (i.e., randomly ordered). Scrambling the item presentation sequence prevents certain types of cheating (e.g., coming in with an ordered list of illicitly obtained answers for the test). For multiple-choice questions, distractors may further be scrambled within each item as an added measure of security. However, scrambling creates a rather minor data management challenge because the scrambled test items (or components of test items) must be unscrambled or otherwise dealt with as part of the test form scoring process.

Advantages

When properly implemented, the CFT model has several attractive features. First, automated test assembly (ATA) procedures usually need only deal with a single target test information function (see Figure 1; also see Luecht, 2006) and a constant set of test specifications (e.g., constraints on content). Second, the test forms can also be constructed simultaneously with item-overlap controls explicitly used to control item exposure across test forms. If each test form is assigned randomly to each examinee from a larger pool of test forms (screening out previously seen forms for retakers), security risks can be minimized. Third, the data management associated with CFT are minimal because the number of relationships is limited to the number of examinees times the number of test forms (not items). Fourth, because the test assembly is done beforehand, extensive quality control procedures can be implemented to check any and all test forms for content balance and other critical features, before release. That is, using preassembled test forms allows for standard content and technical reviews of test forms prior to test administration. Fifth, the model is simple to implement because it does not require the real-time or online test-delivery software to perform any special type of item selections. Furthermore, only very simple software procedures are needed if items are to be [randomly] scrambled within each form. Sixth, because there is usually no real-time item selection or scoring being performed by the test delivery software, the performance of the system is usually optimal in any testing network or Web-based environment and the online/on-site test bank does not need to store item statistics or other data. There are hidden security advantages in that respect (i.e., less data at risk on the Web or at test centers). Finally, this model and its variants provide all of the general advantages of CBT such as flexible test administration schedules, automatic score reporting, and the use of novel item formats.

Disadvantages

The major limitation of the CFT model is that it is not efficient from a measurement perspective. Because a CFT is nonadaptive in nature and items are not "optimally" selected for individual examinees about twice the number of items may be required to achieve the precision of measurement associated with a purely adaptive test (e.g., Wainer, 1993; also see the example provided under the Computer Adaptive Testing section). A second disadvantage relates to exposure risks. That is, unless ATA is used to mass-produce many simultaneous

CFT forms, with overlap controlled, there could be serious security risks. Having only a limited number of test forms could pose serious security risks for testing programs over time, especially if the examinees are allowed to continuously take (and retake) the test. In fact, the Physical Therapist exam, mentioned earlier, has had several instances of coordinated cheating efforts to reproduce items. Examinees do memorize and share items over time on high-stakes tests. A practical limitation on number of test forms may be imposed by the size of the active item bank; that is, the item bank may simply have an insufficient number of items to build large numbers of unique forms. Unless large item banks are available, item exposure will be large, relative to the proportion of examinees taking the test. For these reasons CFTs are typically limited to those situations where new tests are developed frequently to reflect major content changes (e.g., information technology certification exams) and low-stakes testing situations where measurement accuracy and test security are less important.

Linear-on-the-Fly Testing

A variation of preassembled test forms is *linear-on-the-fly testing* (LOFT), which involves the real-time assembly of a unique fixed-length test for each examinee (Folk & Smith, 2002). Like CFT, classical test theory or IRT can be used to generate randomly parallel LOFT test forms (Gibson & Weiner, 1998). There are at least two variations of the LOFT model: a large number of unique test forms can be developed far in advance of test administration (which is merely a special case of CFT, where ATA is employed, as noted above) or test forms can be generated immediately prior to testing (i.e., in real time). The primary advantage of developing the test forms in advance is that content and measurement experts can review each form. According to Jodoin et al. (2002), the securities industry has administered about 100,000 LOFT exams a year for more than 15 years.

Advantages

The primary advantage of the LOFT model is that numerous forms can be developed in real time from the same item pool. Furthermore, there is typically some overlap of items allowed across the test forms. When test forms are assembled just prior to administration, the current exposure levels of the items can be considered in the test assembly algorithm. At-risk items can be made unavailable for selection. For real-time LOFT, explicit item exposure controls can be used to limit the exposure of particular items (e.g., Sympson & Hetter, 1985), in addition to the random sampling scheme. The benefits of LOFT include all those associated with CFTs with the addition of more efficient item-pool usage and reduced item exposure.

Disadvantages

The disadvantages of LOFT are similar to those of CFTs (i.e., decreased measurement efficiency and exposure risks if test banks are relatively small, limiting the number of forms that can be produced). In addition, real-time LOFT may limit or altogether preclude certain quality controls such as test content reviews and data integrity checks. If exposure controls are implemented, there can be a very subtle interaction between the availability of items in the item bank, the content constraints, any statistical targets used, and the choices of control parameters used[3]. Although some quality assurance can be integrated into the live test assembly algorithm, doing so tends to complicate the functionality of the test-delivery system and introduces additional data management challenges (e.g., reconciling examinee records). This latter problem can be slightly reduced in terms of risks to the integrity of the data by creative database management (e.g., using system generated test form identifiers for every LOFT form).

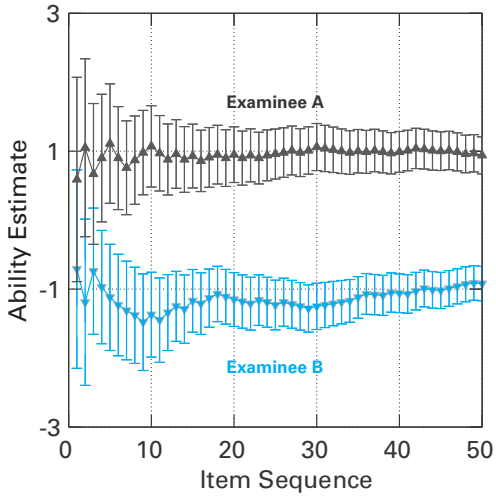
3. In reality, the interaction between the item "supply" (i.e., characteristics of the item bank), the "demands" (test assembly constraints and statistical targets), exposure control mechanisms, and item selection algorithms will impact any type of test assembly, and therefore, all CBT delivery models.

Computer-Adaptive Tests

As we discussed earlier (see Degree and Nature of Test Adaptation), a CAT adapts or tailors the exam to each examinee. Under the purest form of CAT, this tailoring is done by keeping track of an examinee's performance on each test item and then using this information to select the next item to be administered. Thus, CATs are sequentially developed item-by-item in real time by the test-delivery software. The criteria for selecting the next item to be administered to an examinee can range from simply choosing items that maximize the reliability of each examinee's score to complex ATA heuristics. However, the primary item-selection criterion in CAT is to maximize the test information function (Equation 1) and minimize the measurements error of the examinee's score.

Figure 2 shows what happens to the provisional proficiency scores and associated standard errors (the square root of the error variance from Equation 1) for two hypothetical examinees taking a 50-item CAT. The proficiency scale is shown as the vertical axis (–3.0 to +3.0). The sequence of 50 adaptively administered items is shown on the horizontal scale. Although not shown in the picture, initially, both examinees start with proficiency estimates near zero. After the first item is given, the estimated proficiency scores immediately begin to separate (4 for Examinee A and 5 for Examinee B). Over the course of 50 items, the individual proficiency scores for these two examinees systematically diverge to their approximate true values of +1.0 for Examinee A and –1.0 for Examinee B. The difficulties of the 50 items selected for each examinee CAT would track in a pattern similar to the symbols plotted for the provisional proficiency scores. The plot also indicates the estimation errors present throughout the CAT. The size of each error band about the proficiency score denotes the relative amount of error associated with the scores. Larger bands indicate more error than narrower bands. Near to the left side of the plot the error bands are quite large, indicating fairly imprecise scores. During the first half of the CAT, the error bands rapidly shrink in size. After 20 items or so, the error bands tend to stabilize (i.e., still shrink, but more slowly). This demonstrates how the CAT quickly reduces error variance and improves the efficiency of a test. Obviously, these two examinees have very different proficiencies and deserve tests of different difficulty. If you imagine trying to design a single test that would be appropriate for both examinees, the efficiency benefits of a CAT become apparent.

Figure 2
Proficiency scores and standard errors for a 50-Item CAT for two hypothetical examinees



Presently, there are numerous examples of successful, large-scale CAT testing programs such as the ACCUPLACER postsecondary placement exams (College Board, 1993), the Graduate Record Exam[4] (Eignor et al., 1993), the Armed Service Vocational Aptitude Battery (Sands et al., 1997), the Measures of Academic Progress (Northwest Evaluation Association, 2005) used in K-12 settings, and several licensure or certification tests such as the Novell certification exams and the licensure exam for registered nurses (Zara, 1994).

The idea of using the computer to match the difficulty of an item to the proficiency of an examinee was initially proposed by Lord (1977, 1980). Lord's idea was to begin a test administration by presenting an item of moderate difficulty to an examinee. If the examinee answers the question correctly, a slightly more difficult item is administered. If the examinee answers the question incorrectly, a slightly easier question is administered. This iterative process continues until a sufficient number of items is administered for confident estimation of the examinee's score. The adaptive nature of a computerized-adaptive test is controlled by the item-selection heuristic. As described previously, a key goal of the algorithm is to match item difficulty to examinee proficiency. Obviously, the proficiency level of an examinee is not known at the time of testing. Therefore, estimates of examinee proficiency must be used throughout the test session. At the beginning of the test, the proficiency estimate for an examinee is typically set just below the average of the population of all test takers (this estimate is usually selected based on extensive pretesting of the examinee population). A value slightly below the average is used to reduce the chance that the first item on the test will be particularly difficult for an examinee. After each response to an item, the proficiency estimate for the examinee is updated. In addition to matching item difficulty to examinee proficiency and determining when a test ends, a CAT item-selection algorithm also selects items to maximize test information (i.e., reduce measurement error) and may control several other factors such as content representation and item exposure.

There are several types of item-selection algorithms (see van der Linden & Pashley, 2010, for a more complete description). Traditional approaches that are cited in the psychometric literature include maximum information item selection (Lord, 1977), maximum information item selection with the Sympson-Hetter (unconditional) item exposure control procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985), maximum information and Stocking and Lewis (conditional) item exposure control procedure (Stocking & Lewis, 1995, 1998), and maximum information and stochastic (conditional) exposure control procedure (Revuela & Ponsoda, 1998; Robin, 1999, 2001).

Computerized-adaptive testing almost always relies on IRT in selecting items and scoring examinees. IRT posits several mathematical models that characterize items and examinees on a common scale. In IRT, the scale that indicates the difficulty of an item is the same scale that is used to assign scores to examinees. Thus, an item of average difficulty would have the same value on the scale as the value assigned to an examinee of average proficiency. There are several attractive features of IRT, including the ability to provide scores on a common scale for examinees who take different items. A more detailed account of IRT is beyond the scope of this paper. Readers desiring more specific information are referred to the excellent textbooks in this area (e.g., Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968).

4. In 2011-12, the GRE is transitioning to a computer-adaptive multistage test. The CAT version of the GRE was successfully operating from 1992 to 2010.

Options for Ending a CAT Session

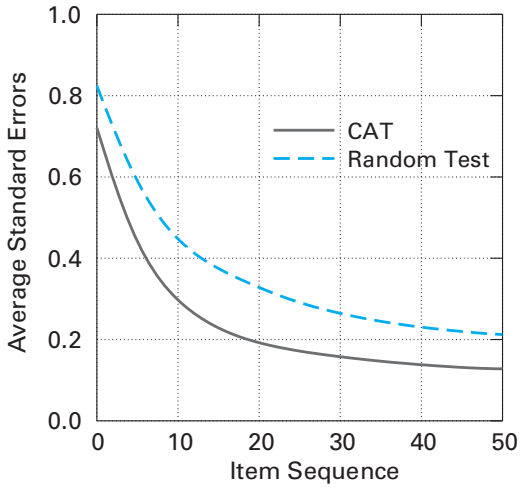
There are several different methods for ending a computerized-adaptive testing session. In some situations, fixed-length CATs are used, where all examinees are administered the same number of items, regardless of the measurement error associated with their score (e.g., College Board, 1993; Northwest Evaluation Association, 2005). However, some testing programs use a variable-length CAT procedure in which the test session ends when some pre-specified level of measurement precision is reached. Test stopping rules for variable length CATs typically use one of two methods, depending on the testing context. In a norm-referenced context, where no performance standards are set on the test, a minimum standard error criterion is typically used. In this situation, an examinee's test ends when the measurement error associated with her or his score dips below a pre-specified level (Lord, 1980). This criterion assures that the scores for all examinees meet a minimum standard of reliability. In criterion-referenced testing situations, such as in licensure or certification testing, a test session ends when it is clear that an examinee's proficiency is above or below a specific threshold, such as a passing score. Further discussion of this approach is presented in a subsequent section of the paper (see Multistage Computerized Mastery Testing).

Advantages

As suggested earlier, computerized-adaptive testing offers improved testing efficiency, which means we can obtain more confident estimates of examinees' performance using fewer items than are typically required on nonadaptive tests. This gain in efficiency stems directly from the CAT item-selection algorithm, which avoids administering items that are too easy or too hard for an examinee. Therefore, CATs are often significantly shorter than their paper-and-pencil counterparts — typically about half as long as a parallel nonadaptive test (Wainer, 1993).

Figure 3 shows the efficiency gains of a hypothetical CAT, compared to a test for which the items were randomly selected. The plot shows the **average** standard errors of the proficiency estimates (the square root of the error variance from Equation 1) over 50 items (horizontal axis). The standard errors are averaged for examinees having different proficiency scores. Also note that the item characteristics used to generate the test results for Figure 3 are rather typical of most professionally developed achievement tests.

Figure 3
Average standard errors for a 50-Item CAT vs. 50 randomly selected items



In Figure 3, we can more specifically see how the errors decrease over the course of the two tests. It is important to realize that the errors decrease for a randomly selected set of items, too. However, CAT clearly does a better job of more rapidly reducing the errors. For example, at 20 items, the CAT achieves nearly the same efficiency as the 50-item random test; at 50 items, the average standard error for the CAT is approximately half as large as for the random test.

Another widely cited benefit of computerized-adaptive testing is a reduction in test anxiety for many examinees (Gershon & Bergstrom, 1991). The assumption is that, in traditional testing, some examinees may "freeze" when presented with an item that is much too difficult for them to answer. The speculation offered by Gershon and Bergstrom is that such examinees may find taking an adaptive test less anxiety provoking. However, other research suggests that a reduction in test anxiety due to the adaptive nature of the test may only apply to examinees of relatively low proficiency (Wise, 1997).

Disadvantages

The item-selection algorithm that governs a CAT requires technical sophistication in several areas. First, stable estimates of item parameters are necessary, and these estimates traditionally need to be gathered from large numbers of examinees in a nonadaptive format. Second, as was learned from the ETS/Kaplan incident (Celis, 1994), if item exposure is not closely monitored, examinees of similar proficiency will see similar items, which compromises test security. Content representation must also be ensured in the item-selection algorithm. When content representation and conditional item exposure control are incorporated into the algorithm, much of the measurement efficiency gains associated with a CAT may be greatly reduced.

In describing this dilemma, Davey and Parshall (1995) noted that CAT item selection algorithms typically strive to meet four types of objectives: (a) measurement precision, (b) content balancing, (c) test security, and (d) efficiency of test administrations. Unfortunately, these objectives are often at odds with each other (Swanson & Stocking, 1993; Luecht, 1995; Robin, 2001). For example, achieving a higher level of test security generally results in lower measurement precision or greatly increased numbers of test items of high quality.

A second limitation of CATs is that they provide less control over the psychometric characteristics of the specific tests that are administered to examinees, relative to other CBT models and to paper-based tests. For example, subject matter experts and psychometricians are unable to review a test "form" before it is administered to an examinee. Also, there are some data that suggest some examinees who take a CAT will receive sub-optimal tests with respect to measurement precision (Robin, 2001). As Robin described:

> "… with no mechanism to ensure a minimum level of test information, it is possible that some examinees are not provided with an adequate opportunity to demonstrate their ability, despite acceptable average test information levels for the target population (Davey & Fan, 2000). This problem is likely to be prevalent for highly constrained tests (i.e., tests that include complex and restrictive content and exposure specifications) assembled from item pools of limited sizes and/or quality." (p. 6)

A third limitation of CATs is that examinees are unable to skip test items or review the answers to previous items. Wainer (1993) pointed out that if examinees are allowed to skip and change answers to questions, they may be able to "trick" the algorithm into administering them the easiest possible set of test questions and subsequently bias their scores upward.

For this reason, the American Council on Education's *Guidelines for Computerized-Adaptive Testing* (1995), recommended against allowing examinees to change their answers. However, this prohibition is resented by many examinees (Legg & Buhr, 1992; Vispoel, 1998).

A fourth potential limitation of CATs is that their [typical] implementation requires that virtually all of the information about every item be stored in the live test bank (item text, answer keys, attribute codes, item statistics, etc.). This poses a certain amount of security risk for any testing program that distributes and stores the live test bank to local test site servers. Even with encryption layers imposed on the data, entire test banks can be stolen if all of the data are "out there."

The final limitation relates to computer system performance issues specific to implementing CATs, especially for large-scale, high-stakes applications. A CAT is a very data intensive application and requires a fairly high degree of computation during the live test administration. On a single computer, system performance issues are usually trivial. That is, most modern personal computers and notebooks have more-than-adequate speed and storage capabilities. The limitation arises when a CAT program is deployed in Web-based or Internet-based testing networks, or in wide area network environments, where a central computer server needs to score and implement an item selection heuristic after every item and for every examinee. Most of the early CBT delivery vendors were unaware of these performance issues because they downloaded the entire test bank and testing software to the local test site and deployed the CATs from local servers. (That approach, of course, created other somewhat hidden security risks, as noted in the previous paragraph.) As testing programs move toward using the Internet, system performance issues will become more prominent.

Constrained Adaptive Testing Using Shadow Tests

Like other CBT researchers, van der Linden (2000, 2002, 2010) conceptualizes the goals of CAT item selection algorithms as a multiple optimization problem. As van der Linden (2002) describes:

> "The objective of the optimization problem is to select the test items so that the statistical information in the test on the ability of the examinee is maximized. At the same time, the selection of the items has to meet a usually large number of constraints to guarantee that the test realizes the same set of content specifications across examinees. … the goal of the test is to maximize its reliability; constraints are necessary to maintain its content validity" … (p. 95).

To meet these constraints and achieve measurement efficiency van der Linden and Reese (1998) introduced the concept of a "shadow test" as a method of achieving an optimal CAT in the face of numerous constraints (also see van der Linden, 2000, 2010) in real time (i.e., as the examinee takes the test). In this method, a complete test is reassembled following each item administration. This test, called the shadow test, incorporates all of the required content constraints, item exposure rules, and other constraints (e.g., cognitive levels, total word counts, test-timing requirements, clueing across items), and uses maximization of test information at the examinee's current proficiency estimate as its objective function. At any given point, the shadow test contains all the items already administered to the examinee, meets all the content constraints, and provides the optimal test information given the examinee's current proficiency estimate. Instead of selecting items from the item bank, an item is selected from the shadow test. After the examinee answers an item, the proficiency estimate is updated, all unused items are returned to the bank, and a new shadow test is

created. The shadow test model is an efficient means for balancing the goals of meeting content constraints and maximizing test information. It is seen as more efficient because the exposure of items can be tracked as examinees take the test and this information can be incorporated into the item selection for the shadow test.

Advantages

A shadow test is a special case of content-constrained CAT that explicitly uses ATA for each adaptive item selection. In that regard, this model blends the efficiency of CAT with the sophistication of using powerful linear programming techniques (or other ATA heuristics) to ensure a psychometrically optimal test that simultaneously meets any number of test-level specifications and item attribute constraints. Shadow testing can further incorporate exposure control mechanisms as a security measure to combat some types of cheating (van der Linden, 2000, 2010). It also does not require simulation studies to establish the item exposure parameters for the items before administering a test.

Disadvantages

Shadow testing is a mathematically elegant model for CAT that has not been implemented to date in a real CBT system. There is little dispute in that regard. Simulation research conducted with paper-and-pencil item banks from the Law School Admissions Test shows extreme promise (van der Linden & Reese, 1998) but is hardly conclusive. There is also a predictable complication with shadow testing that relates directly to system performance, especially with regard to Web-based testing (WBT). Shadow testing requires that a powerful linear programming software package be fully integrated as part of the test-delivery software driver (Diao & van der Linden, 2011). Although commercial linear programming software packages do exist (e.g., the *CPLEX Optimization Studio available from IBM*), they will be costly and complicated to integrate with the current class of test-delivery applications available throughout most of CBT world. Furthermore, even if implemented, the impact on system performance is unknown for WBT (or large-network installations) running most of the required computations and data management routines on the server side. Unless these pragmatic systems issues can be resolved and allow content-constrained CAT with shadow testing to gain widespread use, it may remain an elegant (and somewhat costly) solution that remains "on the shelf."

a-Stratified Computerized Adaptive Testing

a-stratified computerized adaptive testing (AS) (Chang & Ying, 1997, 1999) is an interesting modification on the adaptive theme. AS adapts the test to the examinee's proficiency — like a traditional CAT. However, the AS model eliminates the need for formal exposure controls and makes use of a greater proportion of the test bank than traditional CAT. The issue of test bank use is extremely important from an economic perspective. One of the more unappealing artifacts of an adaptive algorithm that maximizes the test information function for each examinee is that the most informative items are continually in high demand. This leads to overexposure of a relatively small portion of the entire test bank — typically on the order of 30% to 40% — unless item exposure controls are implemented (Sympson & Hetter, 1985). This often leads to requirements for even larger test banks over time, which ultimately increases by a significant amount the total cost of maintaining the testing program. Some very recent unpublished research has suggested that, even with the best item-exposure controls in place, many items in the test bank are still used only once under a traditional CAT algorithm.

a-stratified CAT partitions the test bank into ordered layers, based on statistical characteristics of the items (Chang & Ying, 1997, 1999). First, the items are sorted according to their estimated IRT item discrimination parameters.[5] Second, the sorted list is partitioned into layers (the strata) of a fixed size. Third, one or more items are selected within each strata by the usual CAT maximum information algorithm. AS then proceeds sequentially through the strata, from the least to the most discriminating strata. The item selections may or may not be subject to also meeting applicable content specifications or constraints.

Chang and Ying (1997, 1999) reasoned that, during the initial portion of an adaptive test, less discriminating items could be used because the proficiency estimates have not yet stabilized. This stratification strategy effectively ensures that most discriminating items are saved until later in the test when they can be more accurately targeted to the provisional proficiency scores. (A similar rationale with a different heuristic strategy was suggested by Luecht, 1995.) In short, the AS approach avoids wasting the "high demand" items too early on in the test and makes effective use of the low demand items that, ordinarily, are seldom if ever selected in CAT.

Chang, Qian, and Ying (2001) went a step further to also block the items based on the IRT difficulty parameters. This modification is intended to deal more effectively with exposure risks when the IRT discrimination and difficulty parameters are correlated with each other within a particular item pool.

Advantages

The stratified CAT model appears to have most of the efficiency advantages of a traditional CAT. Other advantages of this method relate primarily to its simplicity in controlling exposure "naturally" and in making better use of the entire test bank. Test banks are expensive to produce. If a large portion of an expensive resource is not used effectively, the unused or minimally used resource is wasted. By systematically using the least discriminating items early, this stratification method uses the entire test bank. Also, by exposing the items more uniformly, AS naturally implements a type of exposure control.

Disadvantages

This method has most of the same limitations discussed relative to CAT, especially regarding quality control, data management, and item review. (Although more than a single item can be administered within strata, effectively making this a multistage "on the fly" type of adaptive test.) The major limitation of this method is that it has never been implemented in even a field test situation. As a result, we have virtually no operational experience with the model.

Testlet-Based CATs

To address the practical shortcomings of CATs, Wainer and Kiely (1987) introduced the concept of a "testlet" to describe a subset of items or a "minitest" that could be used in an adaptive testing environment (see also Wainer & Lewis, 1990). A testlet-CAT involves the adaptive administration of preassembled *sets* of items to an examinee, rather than single items. Examples of testlets include sets of items that are associated with a common reading passage or visual stimulus, or a carefully constructed subset of items that mirrors the overall content specifications for a test. After completing the testlet, the computer scores the items within it and then chooses the next testlet to be administered. Thus, this type of test is adaptive at the testlet level rather than at the item level. This approach allows for better

5. See Lord (1980) or Hambleton and Swaminathan (1985) for a more detailed description of IRT item parameters for multiple-choice questions and related objective response items.

control over exam content and can be used to allow examinees to skip, review, and change answers within a block of test items. It also allows for content and measurement review of these sets of items prior to operational administration.

Sometimes, the testlets are assigned to stages in a variation of multistage testing, which is described in the next section. At the first stage, examinees are administered a *routing test* that determines the difficulty level of the test they will take at the second stage. Their performance on the second stage of the test determines the test they will take at the third stage (if there were one), etcetera. The difference between a testlet-CAT and a multistage test is that with the latter, the minitests administered at each stage can be much larger than a typical testlet, and the number of stages is relatively small, with two or three stages being most common. In practice, multistage tests may differ with respect to several factors such as the numbers of modules administered, the number of items within a module, branching rules used, amount of item overlap across modules, and item exposure levels (Luecht & Nungester, 1998; Jodoin et al., 2002; Zenisky et al. 2010).

Advantages

It should be clear that testlet-based CATs are only partially adaptive because items within a stage (testlet) are administered in a linear fashion. However, both the multistage adaptive and testlet-based CAT models offer a compromise between the traditional, nonadaptive format and the purely adaptive model. Advantages of multistage testing include increased testing efficiency relative to nonadaptive tests; the ability of content experts and sensitivity reviewers to review individual, preconstructed testlets and subtests to evaluate content quality; and the ability of examinees to skip, review, and change answers to questions within a testlet or stage.

Disadvantages

One disadvantage of testlet-based CAT, relative to item-level adaptive tests, is that formation of the testlets sacrifices some amount of measurement precision insofar as the items are not individually targeted to the examinees' proficiency scores. However, as discussed subsequently, recent research (Zenisky et al., 2010) suggests that the loss in efficiency may be minor, particularly in the context of classification testing (e.g., placement, licensure, certification). A second disadvantage is that testlets cannot contain any item overlap. That is, testlet-based CAT requires the testlets to be unique because it is combinatorically not feasible to track testlet enemies (i.e., mutually exclusive testlets). This requirement may severely restrict the number of testlets that can be produced and slightly increase exposure risks. A third limitation is that testlet-based CAT, despite the use of ATA to build the testlets, may yield test forms that do not meet all of the test-level specifications when various testlets are combined. Provided that all of the test specifications can be distributed at the testlet level, this is not a serious problem. However, various programs attempting to implement this model have encountered serious test form quality problems.

Multistage Computerized Mastery Testing

The literature in this area includes discussions of *adaptive mastery testing* (Kingsbury & Weiss, 1983; Weiss & Kingsbury, 1984; Adema, 1990; Kingsbury & Zara, 1989) and *computerized mastery testing* (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992). Parshall et al (2002) use the term *computerized classification tests* in discussing this literature, because many of these models are seen exclusively in classification contexts such as licensure and certification testing. In reviewing these models Pitoniak (2000) distinguished between adaptive mastery testing and computerized classification tests.

As Pitoniak (2000) described, adaptive mastery testing and computerized classification testing differ along two major dimensions: how items are selected and how the classification decision (e.g., pass/fail decision) is made. In computerized classification testing, items are selected to maximize information around the cut score. In adaptive mastery testing items are selected to maximize information around an examinee's current proficiency estimate. With respect to how the classification decision is made, the most popular approaches use either the sequential probability ratio test or Bayesian confidence intervals.

The sequential probability ratio test (SPRT) was developed by Wald (1947) to serve as a quality control test for products. Using the binomial distribution, Wald's test involved two competing hypotheses (product is of sufficient quality or is not of sufficient quality). The hypotheses were evaluated by the number of deficiencies discovered in a sample. SPRT is typically implemented in CBT using IRT, which does not assume that all items (products) are of equal importance. Reckase (1983) used IRT-based SPRT to devise an adaptive test that focused on whether an examinee's proficiency estimate was above or below a specific threshold (cut score). The important feature of this approach is that the confidence interval used for making classification decisions is formed around a specific cut score, rather than around an examinee's proficiency estimate.

The Bayesian approach for making the classification decision, as implemented in adaptive mastery testing, forms a confidence interval around an examinee's current proficiency estimate. If the cut score is above or below the interval, and the criterion of number of test items or sufficient test information is reached, testing stops and the examinee is classified accordingly. If the cut score is contained within this interval, testing continues, unless the maximum number of items or maximum testing times has been reached. In such cases, the examinee is classified based on whether the current proficiency estimate is above or below the cut score (Folk & Smith, 2002; Vos & Glas, 2010).

Lewis and Sheehan (1990) proposed an adaptive testlet-based procedure to balance the goals of measurement efficiency and content constraints. Their original model focused on computerized mastery testing, although extensions to the nonmastery situation have also been proposed (Smith & Lewis, 1995). In the original design, testlets are randomly selected from a pool of parallel testlets and cut-score thresholds are established. After a minimum number of testlets are completed by an examinee and scored by the computer, loss functions associated with "pass," "fail" or "continue testing" are calculated.

In addition to the general benefits of testlet-based CBTs, there are several advantages to the Lewis-Sheehan approach. These benefits include computational efficiency (the examinee's proficiency estimate does not need to be determined after each testlet), use of random testlets within a testing stage (which simplifies the test administration and may provide better item exposure control), and a priori construction of content-balanced testlets that differ systematically in difficulty.

Advantages and Disadvantages

The advantages and disadvantages are virtually identical to those described for testlet-based CATs in the previous section.

Computer-Adaptive Multistage Testing

Luecht and Nungester (1998) introduced *computer-adaptive multistage testing* (ca-MST) as a framework[6] for managing real-life test construction requirements for large-scale CBT applications (also see Luecht et al., 1996; Luecht, 2000; Melican et al., 2010; Zenisky et al., 2010). Functionally, ca-MST is a preconstructed, multistage adaptive test model. The model uses a manufacturing-engineering paradigm that incorporates multistage adaptive technologies and automated test assembly (ATA) in a way that allows test developers to maintain a greater degree of control over the quality of test forms and data. It can be used for adaptive testing applications or mastery testing applications. ca-MST is adaptive in nature and is therefore more efficient than a CFT or LOFT. Yet, ca-MST provides explicit control over content validity, test form quality, and the exposure of test materials. The many practical advantages of ca-MST are reasons why programs like the Uniform CPA Examination and Graduate Record Examination have adopted a ca-MST model instead of CAT.

ca-MST uses the fundamental building block unit — termed a **module** — as the basis for test construction and test delivery. Modules are preconfigured sets of items which may range in size from several items to well over 100 items. More recently, some ca-MST descriptions have used the term "testlets" in place of "modules" as a matter of convention. Certainly, modules may include discrete items or items that share a common stimulus (e.g., sets of 10 to 12 items, each associated with a particular reading passage). These modules or testlets are usually targeted to have specific statistical properties (e.g., a particular average item difficulty or level of precision) and all content balancing is built into the construction of the module. In turn, the test modules are activated as part of a "panel" and are assigned to a particular stage of testing within the panel. This approach of assigning items to modules and modules to panels makes adaptive testing viable under ca-MST and further provides a concrete way of controlling exposure of items and/or modules over time, via the reuse or overlap rules associated with panels.

From an examinee's perspective, ca-MST appears to function as a multistage linear test. After each stage, a scoring and routing process is initiated. The scoring and routing process may involve test adaptation or mastery decision-making, but is largely invisible to the examinee. From a psychometric perspective, each series of modules (the "test form" actually seen by the examinee) needs to meet a specific statistical target, which will be operationally defined as a prescribed level of measurement precision within a particular region of the score scale (i.e., an IRT test information target). From a test development perspective, each "test form" must also meet a variety of categorical test specifications, including the content and other specifications. Automated test assembly (ATA) typically must be used to preconstruct all of the modules so that they individually meet all relevant statistical and categorical specifications.

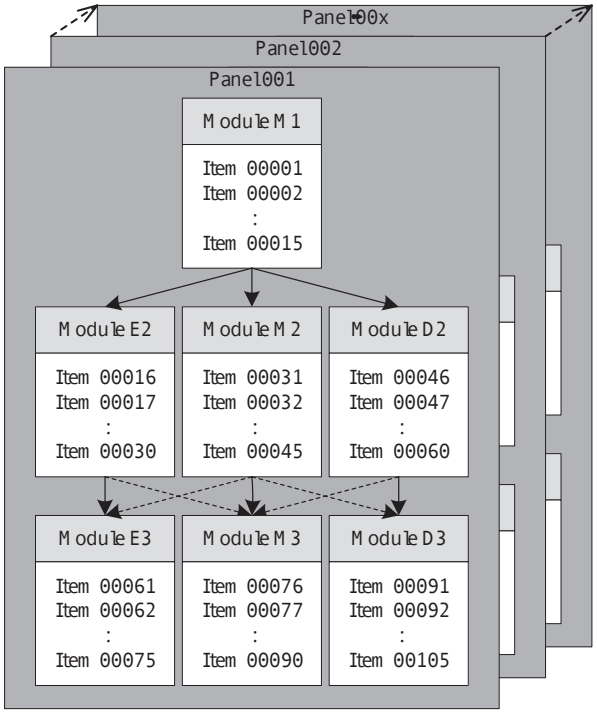
The modules are assigned to a **panel** by stage and difficulty level. The panels themselves, are preconstructed, self-contained adaptive units that are assigned and administered to each examinee by random assignment. A panel is therefore a test administration unit in much the same way that a test form is a unique test administration unit. Having multiple forms of a test helps reduce risks to the integrity of the examination program by lowering the likelihood that some examinees that may have had prior knowledge of some of the test items will actually

6. The original implementation of ca-MST was called computer-adaptive sequential testing (CAST). That label was later modified to ca-MST. Another comparable label is "multistage adaptive testing" (Melican et al., 2010). The inclusion of the "computer-adaptive" qualifier is important to convince policymakers that the test is fully adaptive — just not a CAT.

see those items on their test forms. Similarly, multiple ca-MST panels can be randomly assigned to examinees to enhance test security.

There is a natural hierarchical arrangement in that panels have multiple modules and modules have multiple items. The modules are assigned to distinct stages within the panel. The number of stages and amount of adaptation possible in a panel are indicated by a simple sequence of integers called a *panel configuration*. For example, a two-stage panel with one module at stage 1 and two modules for adapting the test at stage 2 would be denoted as a 1-2 panel configuration. A three-stage panel with three levels of adaptation at stages 2 and 3 would be called a 1-3-3 panel configuration. A five-stage panel with incremental layers of adaptation at the later stages might be denoted as a 1-2-3-4-5 panel configuration. Figure 4 shows a 1-3-3 panel configuration with 15 items per module.

Figure 4
Sample replications of a 1-3-3 panel configuration



Active ca-MST panels are randomly assigned to examinees — filtering out for retesters any panels with previously seen items. The panel then administers itself to the examinee by any of several "score-and-select-next-module" mechanisms that allow the examinee's cumulative performance to dictate the particular **route** through the panel. There are seven plausible routes for the 1-3-3 panel configuration shown in Figure 4. Everybody assigned a particular panel sees the same stage 1 module (M1). Based on the examinees' performances on M1, they may be routed to the E2 (easier), M2 (moderately difficult), or D2 (difficult) module. From E2 an examinee can only be routed to E3 or M3 at stage 3. From M2 he or she can branch to E3, M3, or D3, and from D2 branching is only allowed to M3 or D3. These restrictions on the routes can be set as a matter of policy[7]. In Figure 4, the solid lined arrows denote the primary routes; the dashed line arrows denote secondary routes. Most examinees will follow the primary routes through the panel, although the secondary routes do provide some opportunity for "recovery" in the later stages. The routing is the adaptive part of ca-MST. It is accomplished by cumulatively scoring each examinee through a completed stage and then selecting the most informative module at the next stage. The module selections are done using essentially the same adaptive mechanisms that are used for computerized adaptive testing (CAT), but can be simplified through test design and automated test assembly to provide a maximally precise test for every examinee and simultaneously control quite exactly the proportion of examinees who will see a particular module in the panel (Luecht, 2003; Luecht & Burgin, 2003).

7. Examinees generally should not make extreme jumps (e.g. E2 to D3) unless they are cheating or otherwise exhibiting extremely aberrant response patterns. Allowing the examinees to jump to extreme modules in the panel could benefit cheaters who have prior access to one or more modules.

As noted earlier, each module can include any number of items. The 1-3-3 example shown in Figure 4 shows exactly 15 items per module[8]. The modules, perhaps, should be more formally called **module item lists** (MILs) because each form has *data object* status, where modules are hierarchically related [below] to the item bank (item IDs) and [above] to a particular panel. Each of the panels shown in Figure 4 would require 105 items assigned to one of the seven modules assigned to a given panel. It is therefore relatively easy to plan for inventory by envisioning a fixed number of panels with nonoverlapping modules[9]. A test form (or **test-form list**, TFL) is determined by the route an examinee takes through the panel. Each of the 1-3-3 panels in Figure 4 has seven possible routes: M1+E2+E3, M1+E2+M3, M1+M2+E3, M1+M2+M3, M1+M2+D3, M1+D2+M3, and M1+D2+D3, and each of these routes represents a TFL as a union of the MILs, and, by extension, the items associated with each MIL.

The difficulty of each module and route through the panels is controlled through automated test assembly (ATA), using IRT test information functions (see Equation 1) to target the difficulty of each testlet to a specific region of the relevant score scale (Luecht & Nungester, 1998; Luecht, 2000, 2007; van der Linden, 2005; Melican et al., 2010). For example, there might be seven explicit target test information functions underlying the above 1-3-3 panel configuration depicted in Figure 4. In practice, the IRT test information targets used to construct each 1-3-3 panel would ideally place maximum information where it is most needed to reduce measurement or decision errors.

Once constructed, each panel therefore becomes a formal "data object" for purposes of test administration. That is, a panel is randomly selected and "knows" how to adaptively administer itself. Creating panels as formal data objects provides many operational system advantages in terms of security, quality control and data management. Test committees can review the content and quality of the "test forms" within each panel. Furthermore, trial runs can be made to make sure each panel is working properly, before activation in the live examination pool. From a security perspective, panels can be randomly assigned to examinees, the items can be randomly scrambled within testlets, and item overlap across panels can be explicitly controlled during the ATA builds as a means of controlling item exposure risks (Luecht, 2003; Luecht & Burgin, 2003). Finally, the panels concretely deal with retest issues — that is, previously seen panels can be precluded from selection.

In real time, scoring and routing of the examinees within each panel can be greatly simplified by including a *score routing table* for each panel. The score routing mechanism uses cumulative number-correct scoring and score look-ups to mimic the maximum information criterion used in CAT. The number of correct cut-offs for each authorized route within the panel can be pre-computed and packaged as part of the panel data. For example, the 1-3-3 design shown in Figure 4 requires exactly ten score look-up values (M1→E2, M1→M2, M1→D2, M1+E2→E3, M1+E2→M3, M1+M2→E3, M1+M2→M3, M1+M2→D3, M1+D2→M3, and M1+D2→D3). This feature simplifies the operational scoring and routing functionality needed by the test-delivery driver and potentially could improve performance of the test-delivery driver (i.e., involve less complex data processing and computational steps — especially in a Web-enabled testing environment).

8. Item order can be scrambled within a module as an added measure of security.

9. Modules may be mixed-and-matched later on to create more variants of the panels. This has security benefits insofar as increasing to apparent number of panels and also creates connectivity among the panels to facilitate statistical equating of scores across panels.

It is important to reiterate that the 1-3-3 panel design shown in Figure 4 is merely an example of a ca-MST panel configuration. Virtually any panel configuration can be custom-designed to fit a particular assessment scenario by merely implementing a *template* for the desired configuration (number of stages, number of difficulty levels per stage, size of the testlets or modules within each stage, etc.). Some of the more common panel configurations are the 1-2 and 1-3 (two-stage) designs, the 1-2-2, 1-2-3, 1-3-3, 1-3-4 (three-stage) designs, and the 1-2-3-4, 1-3-3-3, and 1-3-4-5 (four stage). More stages add to the adaptive flexibility. Luecht (2000) presented a number of practical design strategies and ATA considerations for implementing ca-MST designs.

Advantages

The ca-MST test-delivery model is essentially a compromise solution between preconstructed fixed forms and CAT that affords some degree of adaptation, while ensuring adherence to the content specifications for every examinee as well as limiting any overexposure of the test items. Perhaps the greatest asset of ca-MST is that it recognizes the practical limitations of current CBT data management and networking systems. There are many other advantages to ca-MST as well. First, research has shown that examinees like the ability to review the items within testlets (Hadadi & Luecht, 1998; Melican et al., 2010). Second, the adaptive nature of ca-MST capitalizes on many of the same measurement efficiencies as CAT, especially for longer tests or tests having severe content and other constraints (Luecht & Nungester, 1998; Luecht, 2000, 2007). Third, ca-MST simplifies some of the needs for developing, testing, and implementing costly new software systems. In fact, many of the largest commercial CBT test-delivery software vendors have already incorporated the essential functionality for ca-MST in their systems. Fourth, the design of each ca-MST panel fixes the amount of score precision (i.e., uses an absolute target test information function — see van der Linden, 2005, for a discussion of relative versus absolute targeting) where it is desired and reproduces or replicates that "information structure" across the panels. This represents a somewhat different philosophical perspective on the use of information than a computer-adaptive test (CAT). Rather than selecting items to simply maximize information within some region of the score scale, ca-MST instead targets the precision in terms of location and the amount of information provided. Fifth, ca-MST makes strong use of ATA as a front-end process, eliminating the need to implement ATA (i.e., constrained adaptive testing or "shadow testing") in a real-time test-delivery engine. The simultaneous construction of multistage panels using ATA has already been shown to be entirely feasible (see Luecht, 2000; Luecht et al., 2006, van der Linden, 2005; Luecht, 2007; Melican et al., 2010). Sixth, because the panels can be preconstructed, they can also be reviewed for quality of every "test form." Where human review is not entirely feasible, QC software mechanisms can be constructed to flag potentially problematic panels. Seventh, the object-oriented design of the panels and the simultaneous construction of multiple panels, using ATA, provides some very powerful ways of precisely controlling item exposure and managing related examination security risks, including: (a) precise control within and across panels of item overlap by placing appropriate constraints on the ATA test construction model; (b) specific reuse of testlets on various panels proportional to the risk of exposure for the different panel routes (a "mix-and-match" capability that allows multiple panels to be systematically constructed from an initial "parent" set of panels); (c) precise control over the presentation of pretest materials (i.e., who sees them and when); (d) capability to randomly scramble item presentation order within each testlet; and (e) the capability to randomly assign panels to examinees as "test objects", screening out

previously seen panels for test retakers. The final advantage relates to a somewhat technical data management issue. The ca-MST panel framework follows a formal "object-oriented design" (OOD) schema, which greatly facilitates how tests are stored, processed, and checked for quality. The technical advantages of OOD, in terms of data management, quality control, test delivery, and operational processing of test forms are too numerous to list and describe here.

Disadvantages

The only minor disadvantage of ca-ST demonstrated to date is that it is slightly less efficient in a statistical sense than an item-level CAT. However, the differences are miniscule, from practical perspectives (Luecht et al., 1996; Luecht & Nungester, 1998; Luecht, 2000; Hambleton & Xing, 2002; Jodoin, Zenisky, & Hambleton, 2002; Patsula & Hambleton, 1999). Furthermore, the increased control over content and overall test form content, the simplification of the unit selection and scoring functions that need to be built into the test-delivery software, and many other operational advantages that accrue from better quality control seem to offset the very minor efficiency improvements under CAT.

Summary of CBT Models

It should be evident that there is no single CBT model that is best for all testing situations. Each model has its advantages and disadvantages. Before we review research that has evaluated and compared CBT models, we present a summary of some of the strengths and limitations of each model. This summary is presented in Table 1.

Table 1

Summary of CBT Delivery Models

Model	Strengths	Limitations	Selected References	Currently Used By
Computerized Fixed Tests (CFT)	Can review test forms before administration. Examinees can skip and change answers to items. No item selection algorithm needed.	No improvement in measurement efficiency. Inefficient use of item pool. Poor control of item exposure (if few forms).	Parshall et al. (2002)	Microsoft and other IT certification exam agencies, Physical Therapist, Physical Therapist Assistant licensure exams
Linear-on-the-Fly Tests (LOFT)	Better use of item pool and improved item security relative to CFT. Simple item selection algorithm.	QA or review of operational test forms is impossible. Less efficient than any adaptive test.	Folk & Smith (2002)	Securities industry
Computerized Adaptive Tests (CAT)	Most efficient with respect to measurement precision and number of items used.	Content constraints and item exposure reduce efficiency. Requires complex item selection. Test form QA is difficult to implement. Requires large item banks. Poor use of entire item pool, even with exposure controls.	Sands et al. (1997). Wainer et al. (2000). Davey & Pitoniak (2006). Segall (1996, 2010) van der Linden & Pashley (2010).	ACCUPLACER, ASVAB, GRE, Measures of Academic Progress, Novell
a-Stratified Computerized Adaptive Testing (AS)	Uses more of the entire item pool. Nearly as efficient as a CAT. Simple random exposure control and content balancing possible.	QA of test forms is impossible.	Chang & Ying (1997, 1999); Chang & van der Linden (2000); Chang, Qian & Ying (2001)	None. Model-based simulations done with GRE data.
Content-Constrained CAT with Shadow Tests	Maximizes information while handling content and other constraints efficiently in real time, using linear programming optimization.	Not used operationally. System-level performance issues for large-scale CBT.	van der Linden & Reese (1998). van der Linden (2000, 2002, 2005, 2010)	None. Model-based simulations done with ASVAB and LSAT data.

Table 1 (continued)

Summary of CBT Delivery Models

Model	Strengths	Limitations	Selected References	Currently Used By
Testlet-Based CAT and Multistage Computerized Mastery Tests (*combined*)	Adaptive selected modules or testlets. Examinees can skip and change answers to items. Adaptive component improves measurement precision relative to fixed tests.	Less efficient measurement precision than pure CAT. Content balance at the test form level not guaranteed.	Weiss & Kingsbury, (1984); Kingsbury & Zara, (1999); Lewis & Sheehan (1990); Sheehan & Lewis, (1992)	Podiatry licensing exam Gibley (1998)
Computer-Adaptive Multistage Testing	Preconstructed content-balanced modules with targeted test information and built-in item/module exposure controls. QA of tests is possible. Simplified real-time scoring and routing (score tables). Adaptive component improves measurement precision relative to fixed tests.	Less efficient measurement precision than pure CAT.	Breithaupt & Hare (2007); Luecht (2000); Luecht & Nungester (1998, 2000); Luecht et al. (1996); Luecht et al. (2002); Sireci et al. (2008)	NBME (USMLE Field Tests), AICPA (Uniform CPA Examination), ETS (GRE), State of Oregon (ELPA), Massachusetts Adult Proficiency Tests

Some Empirical Studies of CBT Models

Several studies appear in the recent CBT literature that may be helpful for evaluating the different options available for delivering CBTs. Some studies have looked at practical issues in CBT, such as allowing examinees to review or change answers. In this section, we present a summary of selected studies in this area.

Comparing Linear, Adaptive, and Mastery Test Models

Luecht et al. (1996) and Luecht & Nungester (1998) compared ca-MST to CFT and CAT, using panel designs appropriate for the United States Medical Licensing Examinations® (the USMLE is copyrighted by the National Board of Medical Examiners and the Federation of State Medical Boards). A series of simulation studies were conducted that demonstrated that (for long tests like the USMLE Steps), CAST was approximately 98% as efficient as CAT and far more efficient than a randomly selected fixed-item test. Researchers at the National Board of Medical Examiners also carried out several large field-test studies in 1996 (Step 2) and in 1997 (Step 1) employing CAST designs with real medical students. The results confirmed the practical and psychometric benefits of CAST and also contributed to the knowledge base about examinee perceptions about navigation and item review, adapting on difficulty, etc.

In a series of studies, Hambleton and his colleagues (Hambleton & Xing, 2002; Jodoin et al., 2002; Patsula & Hambleton, 1999) evaluated various multistage testing models and compared them with fixed and random testing models. Patsula and Hambleton (1999) evaluated different options for designing multistage tests and compared them with linear tests and CATs. The multistage test design options investigated were the number of stages, the number of modules per stage, and the number of items per module. Using simulated data, they found that proficiency estimation accuracy using a three-stage test was similar to that obtained from a CAT when the number of stages was increased from two to three. The number of items per stage was relatively inconsequential with respect to the accuracy or efficiency of the multistage tests.

Jodoin et al. (2002) compared multistage tests to fixed-length tests within a licensure testing context by disassembling four fixed 60-item tests into a 240-item pool from which random and multistage tests were created. Three random 60-item tests and six adaptive, multistage tests were created. All tests were targeted to the average test information function calculated from the four operational tests. The first three multistage tests incorporated a 1-3-3 ca-MST-like panel design, which specifies three stages. In each stage, examinees were administered 20 items. The 1-3-3 design signifies one module of moderate difficulty at the first stage, and three modules at the second and third stages (i.e., easy, moderate, and difficult modules). The second three multistage tests dropped the third stage (i.e., 1-3 design). Jodoin et al. also systematically manipulated the use of highly discriminating items. In one analysis, test information was equated across stages. In another analysis, modules in the second and third stages yielded more test information. Because they were working within a licensure testing context, they investigated examinee classifications using three different passing scores (low, medium, high).

Jodoin et al. (2002) used three criteria for evaluating the models: correlation between true and estimated proficiency, decision accuracy, and decision consistency. Their results indicated that the LOFT and three-stage models produced results very similar to the operational tests (e.g., true/observed proficiency correlations were around 0.93 for these models). Interestingly, the two-stage models also displayed impressive levels of true/observed correlations (around 0.91). The decision consistency and accuracy results revealed only minor differences among all models. The three-stage designs had slightly higher decision accuracy than the LOFT

design and lower decision accuracy than the operational form. There were only minor differences among the three-stage, LOFT, and operational forms with misclassifications below 10% and correct classifications in excess of 90% for all designs at all three pass rates. Moreover, the two-stage results were slightly worse, but comparable, with misclassifications between 10% and 12% and correct classifications exceeding 88% across all scenarios. Finally, with respect to distribution of discriminating items (i.e., more or less test information at stage one), the results revealed no differences.

Jodoin et al. concluded that LOFT and two-stage adaptive tests may be practical alternatives for classification exams. They hypothesized that the three-stage exam may not have fared well due to the realistic content constraints that were imposed. The results of this study are encouraging for the development of short multistage tests that are designed to incorporate practical advantages into a CBT program such as allowing examinees to change answers (within a module), use the item pool efficiently, and maintain strict item exposure and content constraints.

Hambleton and Xing (2002) compared CAT, multistage test, and LOFT designs, also in a licensure context, to determine whether the results of Jodoin et al. (2002) would hold up if different test assembly strategies were used. They investigated two strategies — targeting the test information function to the region where most examinees were located or targeting the information to the cut score. With respect to decision consistency and decision accuracy, their results suggested that it made little difference. They concluded that although the CAT design performed best, the other designs were comparable. They also concluded that matching the test to the proficiency distribution of examinees, rather than to the passing score, led to slight improvements (an encouraging finding for improving item-pool usage).

Lewis and Sheehan conducted two studies to investigate different features of the CMT model. In Lewis and Sheehan (1990), they evaluated the characteristics of several loss functions and looked at the impact of allowing test length to vary, versus keeping it fixed. Parallel testlets containing 10 items each were constructed; each testlet covered two content areas. For the fixed-length condition, six testlets were administered (i.e., 60 items). For the variable-length condition, a minimum of two testlets (20 items) and a maximum of six testlets (60 items) were used. The prior probabilities of mastery and nonmastery were set to be equal at 0.5. Using simulated data, they found that a loss function that considered a false positive error (passing a nonmaster) to be twice the loss of a false negative error (failing a master) had the most desirable operating characteristics. Using this loss function, the variable-length tests were able to make mastery/nonmastery decisions with the same level of decision accuracy, but with many fewer items, than the fixed-length tests.

Sheehan and Lewis (1992) evaluated the degree to which randomly selected testlets could depart from parallelism without impacting decision accuracy. The "nonparallel" testlets each had the same number of items and were roughly equivalent with respect to content areas covered, but they were not constructed to ensure that likelihood functions of the number of correct scores would be equivalent across testlets. The object of the study was to determine whether using the same set of cut scores for all testlets would affect average test length, overall pass rate, and classification errors. Nonparallelism of testlets provided similar results to the use of parallel testlets with respect to all outcome variables (passing rates, test length, and errors rates). These results suggest that the Lewis-Sheehan model may not need strictly parallel testlets in some situations.

Several variations of multistage test designs were evaluated by Reese and her colleagues at the Law School Admission Council (Reese & Schnipke, 1999; Reese, Schnipke, & Luebke,

1999; Schnipke & Reese (1999). Reese and Schnipke compared the performance of a two-stage test design, a CAT design, and a linear design. Their results indicated that the two-stage approach provided precision similar to that of the CAT. The two-stage test provided slightly more information than the CAT in the middle of the ability distribution, but slightly less in the tails. In a follow-up study, Schnipke and Reese (1999) included two-stage designs in which the candidate could be rerouted within the second testlet, as well as designs with more than two stages of testing. As with the previous study, the CAT design had greater precision and less bias in the tails of the ability distribution. However, the multistage designs performed similarly to the CAT, particularly in the middle of the ability distribution.

Other researchers have focused on simplifying constraints within a CAT to take advantage of increased measurement precision without turning to a testlet-based model. For example, Guille, Lipner, Norcini, and Folske (2002) explored a simplified procedure for adding content constraints in a CAT using a stratified (by content area) random sampling of items. For their data, they achieved average conditional item exposure rates similar to those obtained using the Sympson and Hetter (SH) conditional item exposure approach (mean exposure was .15 versus .14 with the SH approach). To facilitate pool usage and item exposure control, Chang and Ying (1997, 1999), and Chang and van der Linden (2000) suggested stratifying the items within a pool by difficulty and discrimination (*a*-stratified CAT). Recently, Deng, Ansley, and Chang (2010) evaluated and compared three item selection procedures — one based on maximum information and the other two based on the *a*-stratified approach. They found maximum information had an obvious precision advantage when there were no constraints, but a refined *a*-stratified approach based on selecting more highly discriminating items was better in meeting constraints, and achieved similar precision to the maximum information strategy in most situations.

Test-Delivery Model Evaluations and Conclusions

As noted at the onset of this review, there is not a singular CBT model that fits for every testing program. There are direct and indirect benefits and costs associated with each of the eight CBT models presented. For purposes of a comparative evaluation, it is clear that both benefits and costs need to be computed on common metrics. For example, the arguments typically offered in the literature in favor of CAT stress the "efficiency gains," where efficiency is measured in terms of reductions in test length, reductions in errors, increases in IRT test information units, or improvements in reliability. However, efficiency is NOT the only relevant metric for comparing different CBT models. Other useful cost-benefit metrics are needed (Luecht, 2005b).

Continuing with the efficiency example, what are the *real* benefits (reported in dollar savings), if, on average, a CAT is demonstrated to be twice as "efficient" as a CFT or LOFT? Correspondingly, what are the costs of all associated test and system development, implementation, and maintenance? Virtually every testing program that has implemented CAT reports substantial increases in costs (item banking and computer system redesign, enormous R&D resource expenditures, item-pool production costs, etc.). It is not reasonable to evaluate benefits in the absence of costs. Four cost-benefit-related *metrics* that seem useful in evaluating these eight CBT models are: (1) parsimony; (2) system performance; (3) measurement efficiency; and (4) provision for quality control/assurance.

Parsimony implies simplicity in design, implementation, and maintenance. Unnecessary design features, complex implementation requirements, or needs for continual repair and

maintenance add to the costs and offset benefits. Luecht (2002c, 2005a, 2005b) discussed some of the enormous operational complexities and potential costs involving the [re]design, implementation, and maintenance of seven CBT functional subsystems: (a) item writing and development; (b) test assembly and composition; (c) examinee eligibility and registration management; (d) test-delivery software; (e) postexamination processing; (f) item and test analysis; and (g) final scoring, reporting, and communication. In this review, we have focused primarily on test assembly/composition, test delivery, and scoring. Regardless, the straightforward conclusion is that greater simplicity is viewed as being more cost effective.

System performance relates to technical performance of the CBT system. It does not require an advanced degree in computer science to predict that computational intensity, large-scale digital storage, and data transmission issues all impact various aspects of performance within a computer system. Computer users all-too-often complain that "the network is slow" or "the Internet seems jammed." In general, system performance is affected by anything that creates *load and/or demands* on the finite capacity system — which a computer system is. Included are factors such as increased numbers of computations by file servers and/or more complex computations, huge amounts of test material data and response data to be stored, and increased numbers of data transactions, all of which degrade to some extent the performance of a CBT system — especially in large-scale networks and Internet-based testing environments. Network flow optimization strategies and distributed processing paradigms can alleviate some load or demand factors; however, the problem will never completely disappear. The most effective strategy is to reduce the load or demand.

Measurement efficiency has been discussed extensively throughout this review. It has often been the sole criterion used in past CBT model evaluations. Consistent with Luecht (2005b), we suggest that the weight given to efficiency should be carefully reviewed and applied in terms of concrete financial benefits, and considered alongside the real associated costs of test material and systems design, implementation, and maintenance. For our present purposes, we have chosen to think about efficiency as measured in terms of reductions in test length and associated per-item costs.

Quality control (QC) ideally improves the overall yield of products (items and tests) and/or reduces waste or scrap. CBT is a large-scale production enterprise that requires the application of manufacturing-engineering principles (Luecht, 2000, 2002c). Some aspects of QC can be automated (e.g., computing tolerances and flagging outliers); other aspects require human review. Carrying out test-form quality checks can include a mixture of automated and human QC reviews. One thing, however, is clear: Increasing the number of opportunities for carrying out QC procedures and engaging in stronger quality controls are viewed as beneficial to the final product, a top-quality measurement instrument that accomplishes its intended purpose. As previously alluded to, quality assurance (QA) is different from QC. QA typically involves sampling products (sample audits) and using statistical models to detect potential problems. While quality assurance is clearly better than no quality check, it can be less beneficial than quality control, especially when evaluating sometimes fuzzy outcomes such as test-form quality.

Table 2 provides a comparison of the eight CBT models in terms of these four cost-benefit metrics. The ratings in the table (high, moderate, low) are not based on any absolute standard. In general, "high" indicates a positive or beneficial degree of parsimony, good-to-excellent system performance, large efficiency gains relative to a baseline CFT or LOFT, and strong provision for QC and/or substantial QA. A rating of "moderate" indicates a reasonable degree of simplicity, but some complexity — usually related to the need to add computational and

data management functionality to the test-delivery driver, somewhat strained-to-satisfactory system performance, moderate efficiency gains relative to a CFT baseline, and provision for solid QA (but not QC). Finally, a rating of "low" denotes fairly complicated computational procedures for real-time test assembly, scoring, and data management, increased demand computations and use of other system resources (e.g., increased storage demands), baseline measurement efficiency (CFT and LOFT), and limited provision for QA with no direct provision for QC of test materials or data. We acknowledge that others might assign different ratings, given their experiences and perspectives.

Table 2

A Comparative Evaluation of CBT Models Based on Four Metrics

CBT Model	Parsimony	System Performance	Measurement Efficiency	QA/QC Opportunities
CFT	High	High	Low	High
LOFT	High	Moderate	Low	Low–Moderate
CAT	Low–Moderate	Moderate	High	Low–Moderate
Shadow Test CAT	Low	Low	High	Moderate
a-Stratified CAT	Moderate	Moderate	High	Low–Moderate
Adaptive Testlets	Moderate	Moderate	High	Moderate–High
Multistage CMT	Moderate	Moderate	Moderate–High	Moderate
Ca-MST	Moderate	High	High	High

Table 2 is intended to highlight the potential benefits and costs of the eight models relative to one another. Although one could convert these qualitative ratings to numerical points, sum them, and then rank the models based on some total "score," we recommend against doing so. These four metrics clearly need to be weighted within the context of any organization's current examination programs, resources, and plans for systems changes in the future.

Perhaps conspicuously absent from our comparative evaluation of CBT models is a discussion of test-item bank size. The reason for its exclusion is simple. Large test-item banks are needed for virtually any type of high-stakes CBT that is administered on a continuous or near-continuous basis. The implication is that all eight CBT models are at serious risk if the test-item banks are too small. Neither do exposure controls help when a small test-item bank is exposed for an extended period. However, some models need larger item banks relative to others (e.g., CAT versus ca-MST).

Another missing consideration involves the use of innovative item types. That is, computerized testing has introduced many opportunities for new items types ranging from uses of multimedia stimuli (sound, video, tactile) to complex computer-based performance assessments involving simulated work environments (Bejar, 1991; Clyman, Melnick & Clauser, 1995; Luecht & Clauser, 2002; Drasgow, 2002; Devore, 2002; Drasgow et al., 2006). Given the potential for better fidelity measurements, any organization moving to CBT needs to consider the near certainty of using more than just multiple-choice questions and short answer items on tests. It is important to recognize that these new item types will have serious implications for test development, systems design and integration, security, exercise selection, presentation, timing, human-factors usability issues, and scoring. A thorough discussion of this topic is far beyond the scope of this paper (but see Sireci & Zenisky, 2006); nonetheless, their eventual use needs to remain a consideration. If a particular CBT model does not have the flexibility to easily incorporate new item types — including all necessary modifications to systems, data structures, and functions — it could preclude measurement innovations for years to come.

Validity Issues

Selection of a CBT design can affect the validity of scores from a testing program, so validity issues must also be considered in deciding on the best CBT model for a particular program. Messick (1989) stated that most validity issues could be described as stemming from construct underrepresentation or construct-irrelevant variance. As he put it, "Tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory, or else include something that should be left out, or both" (p. 34). CBT can improve the degree to which the items on a test represent the construct measured by allowing for more diverse item types than those available in a paper-based format. Thus, as mentioned earlier, the degree to which a test-delivery model can incorporate and score innovative items is an important issue in selecting a CBT model. However, CBT may also interfere with the construct measured if it somehow inhibits examinees from demonstrating their best performance. For example, if a particular CBT design is difficult for some examinees to navigate, it may slow them down or they may become frustrated. Also, adaptive testing may introduce or inhibit test anxiety for different types of examinees (Wise, 1996), so this issue deserves further study. On the other hand, CBTs may be able to better engage examinees through more interesting item formats, visuals, and even rewards.

CBTs may also promote validity by making tests more authentic. For example, if examinees (e.g., writers, computer programmers, accountants), typically do their work on a computer, putting the test on the computer provides a better match to how they complete their jobs in the real world. Computers can also solve the debate as to whether calculators should be allowed for math tests. Such decisions can be made on an item-by-item basis by simply making the calculator available for those items.

Another important validity issue in the 21st century is accessibility. Examinee populations are increasingly diverse, and tests are commonly being adapted to better serve individuals with disabilities or examinees who speak different languages. The ability of computers to address these needs in real time is just being realized. CBT systems should be able to provide choice with respect to several popular test accommodations such as alternate language versions, increased font size, screen-reading software, point-and-click interfaces, and encouragement. We imagine all of the models reviewed in this paper could accommodate such flexibility, but some could do so more easily. Nevertheless, designing CBTs that are accessible for all subgroups of an examinee population will be important as more tests become computerized.

Conclusions

In this review, we described the promises offered by computer-based testing and discussed the strengths and limitations of several models for delivering CBTs. These models can be evaluated using psychometric, cost, and practical criteria. There are no doubt additional issues and perspectives that we have left out. Our review has attempted to balance findings from academic research with operational experiences from existing CBT programs. We have further attempted to represent, at some level, views related to psychometrics, test development, computer science and information systems, and even finance. We hope that we have not misrepresented any of those perspectives. Clearly, each perspective should be used when selecting a design for delivering valid CBTs to best fulfill the purposes of a specific testing program.

References

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement*, *27*, 241–253.

American Council on Education. (1995). *Guidelines for computerized-adaptive test development and use in education*. Washington, DC: Author.

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*, 237–247.

Armstrong, R. D., Jones, D. H., & Wu, I-L. (1992). An automated test development of parallel tests. *Psychometrika*, *57*, 271–288.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology*, *76*, 522–532.

Berger, M. P. F. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement*, *22*, 248–258.

Birnbaum, A. (1968). Estimation of an ability. In F. M Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 423–479). Reading, MA: Addison-Wesley.

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2010). Assembling an inventory of multistage adaptive testing systems. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 247–268). New York: Springer.

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

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computerized-adaptive test scores. *Journal of Educational Measurement*, *41*, 137–148.

Celis, W. (1994, December 16). Computer admissions test found to be ripe for abuse. *The New York Times*, p. 3d.

Chang, H. H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing item b-blocking. *Applied Psychological Measurement*, *25*, 333–342.

Chang, H. H., & van der Linden, W. J. (2000, April). *A zero-one programming model for optimal stratification of item pools in* a*-stratified computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Chang, H. H., & Ying, Z. (1997, June). *Multistage CAT with stratification designs*. Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN.

Chang, H. H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based case simulations. In E. L. Mancall & Ph. G. Bashook (Eds.), *Assessing clinical reasoning: the oral examination and alternative methods* (pp. 139–149). Evanston, IL: American Board of Medical Specialties.

College Board (1993). *ACCUPLACER®: Computerized placement tests: Technical data supplement*. New York: Author.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *16*, 167–188.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for the item selection and exposure control with computerized adaptive testing*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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 Associates.

Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, *47*, 202–226.

Devore, R. (2002, April). *Considerations in the development of accounting simulations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*, 398–409.

Dillon, G. F., Clyman, S. G., Clauser, B. F., & Margolis, M. J. (2002). The introduction of computer-based case simulations into the United States Medical Licensing Examination. *Academic Medicine, Oct:77(10 Suppl.)*, S94–96.

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 1–35). Mahwah, NJ: Lawrence Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Washington, DC: American Council on Education/Praeger Publishers.

Educational Testing Service (2011). *GRE information and registration bulletin*. Princeton, NJ: Author. Retrieved from www.ets.org/s/gre/pdf/gre_info_reg_bulletin.pdf

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (RR-93-56). Princeton, NJ: Educational Testing Service.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum.

Foster, D. (April, 2011). Personal communication.

Gershon, R. C., & Bergstrom, B. (April, 1991). *Individual differences in computer adaptive testing: Anxiety, computer literacy, and satisfaction*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Gibley, C. W. (1998, Summer). The National Board of Podiatric Medical Examiners new testing methodology. *Clear Exam Review*, *9*(2), 29–32.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement*, *35*, 297–310.

Guille, R., Lipner, R. S., Norcini, J. J., & Folske, J. C. (2002, April). *Content-stratified random item selection in computerized classification testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Hadadi, A., & Luecht, R. M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine*, *73*, S47–50.

Hadadi, A., Luecht, R. M., Swanson, D. B., & Case, S. M. (1998, April). *Study 1: Effects of modular subtest structure and item review on examinee performance, perceptions and pacing*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Hambleton, R. K., & Swaminathan, H. R. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer.

Hambleton, R. K., Swaminathan, H. R., & Rogers, J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Hambleton, R. K., & Xing, D. (2002). *Comparative analysis of optimal and non-optimal computer-based test designs for making pass-fail decisions* (Center for Educational Assessment (Research Report No. 457). Amherst, MA: University of Massachusetts, School of Education

Haynie, K. A., & Way, W. D. (March, 1994). *The effects of item pool depth on the accuracy of pass/fail decisions for the NCLEX using CAT*. Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

Jodoin, M., Zenisky, A., & Hambleton, R. K. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

Kingsbury, G. G., & Zara, A. R. (1991). *A comparison of procedures for content-sensitive item selection in computerized adaptive tests*. Applied Measurement in Education, *3*, 241–261.

Kingsbury, G. G., & Zara, A. R. (1999, April). *A comparison of conventional and adaptive testing procedures for making single-point decisions*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice*, *11*, 23–27.

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

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*, 95–100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luecht, R. M. (1995, March). *Some alternative CAT item selection heuristics* (NBME Technical Report RES95031). Philadelphia: National Board of Medical Examiners.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404.

Luecht, R. M. (1998a, April). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R.M. (1998b). Computer assisted test assembly using optimization heuristics, *Applied Psychological Measurement*, *22*, 224–236.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high-quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2002a, February). *An automated test assembly heuristic for multistage adaptive tests with complex computer-based performance tasks*. Invited paper presented at the Annual Meeting of the Association of Test Publishers, Carlsbad, CA.

Luecht, R. M. (2002b, April). *From design to delivery: engineering the mass production of complex performance assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2002c, June). *Operational issues in computer-based testing*. Invited keynote address at International Test Commission Conference, Winchester, UK.

Luecht, R. M. (April, 2003). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2005a). Operational issues in computer-based testing. In D. Bartrum and R. Hambleton (Eds.), *Computer-based testing and the Internet*. New York: Wiley & Sons Publishing.

Luecht, R. M. (2005b). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Association of Test Publishers Journal*. Retrieved from www.testpublishers.org/journal.htm

Luecht, R. M. (2005d). Computer-based testing. *Encyclopedia of Social Measurement, 1*, 419–427.

Luecht, R. M. (2006). Designing tests for pass/fail decisions using IRT. In S. Downing & T. Haladyna (Eds.) *Handbook of test development* (pp. 575–596). Mahwah, NJ: Lawrence Erlbaum and Associates.

Luecht, R. M. (2007, October). *Multiple objective function, multiple constraint set optimization models for automated test assembly*. Invited paper presented at the Annual Meeting of International Conference on Advances in Interdisciplinary Statistics and Combinatorics, Greensboro, NC.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (April, 2002). *A testlet assembly design for the Uniform CPA Examination*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for the uniform CPA examination. *Applied Measurement in Education*, *19*, 189–202.

Luecht, R. M., & Burgin, W. (April, 2003). *Matching test design to decisions: Test specifications and use of automated test assembly for adaptive multi-stage testlets*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from http://www.psych.umn.edu/psylabs/catcentral

Luecht, R. M., & Clauser, B. E. (2002). Test models for complex CBT. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67–88). Mahwah, NJ: Lawrence Erlbaum.

Luecht, R. M., & Hirsch, T. R. (1992) . Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement*, *16*, 41–51.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229–249.

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multistage adaptive test: The Uniform CPA Examination. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–189). New York: Springer.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Mills, C. N., Potenza, M. T., Fremer, J., & Ward, W. C. (Eds.) (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*, 287–304.

Northwest Evaluation Association (2005, May). *Technical manual: For use with the Measures of Academic Progress and Achievement Level tests*. Lake Oswego, OR: Author.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Patsula, L. N., & Hambleton, R. K. (1999, April). *A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.

Pitoniak, M. J. (2000). *Testlet-based designs for computer-based testing in a licensure and certification setting* (Laboratory of Psychometric and Evaluation Methods Research Report No. 391). Amherst, MA: University of Massachusetts, School of Education.

Pommerich, M., & Burden, T. (2000, April). *From simulation to application: Examinees react to computerized testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–257). New York: Academic Press.

Reese, L. M. & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing (Law School Admission Council Computerized Testing Report 96–04)*. Newtown, PA: Law School Admission Council.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999, August). *Incorporating content constraints into a multi-stage adaptive testlet design* (Law School Admission Council Computerized Testing Report 97-02). Newtown, PA: Law School Admission Council.

Revuela, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311–327.

Robin, F. (1999, March). *Alternative item selection strategies for improving test security and pool usage in computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Québec.

Robin, F. (2001). *Development and evaluation of test assembly procedures for computerized adaptive testing* (Unpublished doctoral dissertation). School of Education, University of Massachusetts, Amherst, MA.

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement*, *22*, 212–223.

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

Schnipke, D. L., & Reese, L. M. (1999, May). *A comparison of testlet-based designs for computerized adaptive testing*. Law School Admission Council Computerized Testing Report 97-01, Newtown, PA: Law School Admission Council.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 57–101). New York: Springer.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, *16*, 65–76.

Sireci, S. G,, Baldwin, P., Martone, A., Zenisky, A., Hambleton, R. K., & Han, K. T. (October, 2006). *Massachusetts adult proficiency tests technical manual*. Amherst, MA: Center for Educational Assessment.

Sireci, S. G, Baldwin, P., Martone, A., Zenisky, A., Kaira, L., Lam, W., Shea, C., Han, K. T., Deng, N., Delton, J., & Hambleton, R. K. (April, 2008). *Massachusetts adult proficiency tests technical manual: Version 2*. Amherst, MA: Center for Educational Assessment.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of testing* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum.

Smith, R., & Lewis, C. (1995, April). *A Bayesian computerized mastery model with multiple cut scores*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report RR-93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995). *A new method for controlling item exposure in computerized adaptive testing (Research Report No. 95-25)*. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston: Kluwer.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277–292.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement*, *22*, 271–279.

Swanson, D. B., Featherman, C., Case, S. M., Luecht, R. M., & Nungester, R. J. (1997, April). *Relationship of response latency to test design, examinee proficiency, and item difficulty in computer-based test administration*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 177–186.

Sympson J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive tests*. Paper presented at the Annual Conference of the Military Testing Association, San Diego, CA.

Timminga, E. (1998). Solving infeasibility problems in computerized test assembly. *Applied Psychological Measurement*, *22*, 280–291.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195–211.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computer-adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer.

van der Linden, W. J. (2002). On complexity in CBT. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J. (2010). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 31–55). New York: Springer.

van der Linden, W. J., & Adema, J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics*, *14*, 279–290.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, *54*, 237–248.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in a multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computer-adaptive testing: Theory and practice*. Boston: Kluwer.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York: Springer.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259–270.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement*, *35*, 155–167.

Vos, H. J., & Glas, C. A. W. (2010). Testlet-based adaptive mastery testing. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 389–407). New York: Springer

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*, 15–20.

Wainer, H., Bradlow, E. R., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*, 243–251.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1–14.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Wise, S. L. (1997, April). *Examinee issues in CAT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Zara, A. R. (1994, March). *An overview of the NCLEX/CAT beta test*. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Zenisky, A., Hambleton, R. J., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York: Springer.

CollegeBoard

11b-4434