Some Useful Cost-Benefit Criteria for Evaluating Computer-based Test Delivery Models and Systems

Richard M. Luecht
University of North Carolina at Greensboro

April, 2005

Acknowledgements: Aspects of this work were developed collaboratively with Dr. Steve Sireci (University of Massachusetts at Amherst) as part of a research study funded by The College Board. Any ill-conceived opinions, omissions, or errors are solely those of the author.

Abstract: Computer-based testing (CBT) is typically implemented using one of three general test delivery models: (1) multiple fixed testing (MFT); (2) computer-adaptive testing (CAT); or (3) multistage testing (MSTs). This article reviews some of the real cost drivers associated with CBT implementation—focusing on item production costs, the costs associated with administering the tests, and system development costs—and elaborates three classes of cost-benefit-related factors useful for evaluating CBT models: (1) real measurement efficiency; (2) testing system performance; and (3) provision for data quality control/assurance.

Introduction

Multistage tests (MSTs) and multiple-fixed tests (MFTs) are emerging as extremely useful alternatives to other test delivery models such as linear-on-the-fly (LOFT) and computerized adaptive testing (CAT). MST and MFT models tend to provide stronger quality controls over the test forms and data. They also reduce the complexity of test assembly and scoring algorithms needed by the test delivery software and appear better suited for incorporating multi-item problem sets and computerized performance simulations (CPS)

Obviously, no single CBT delivery model fits for every testing program. What should be clear, however, is that both benefits and costs need to be computed on common metrics, for purposes of comparative evaluation between competing models. For example, the argument typically offered in the theoretical psychometric literature in favor of CAT stresses 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 <u>not</u> the only relevant metric for comparing different CBT models. Other useful cost-benefit metrics are needed. For example, 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 perceived benefits in the absence of costs.

In this paper, I review some of the real financial cost drivers associated with CBT implementation and present three classes of cost-benefit-related factors that may prove useful in evaluating these CBT models: (1) real measurement efficiency; (2) testing system performance; and (3) provision for data quality control/assurance. Real measurement efficiency refers to reductions in test production and/or test administration costs. Testing system performance relates to technical performance of the CBT system and includes considerations of parsimony in the design, implementation, and maintenance of the subsystems and procedures needed to support ongoing testing operations. Finally, data quality control relates to manual and automated procedures that improve the overall yield of high quality items and tests that meet all relevant specifications and that accomplish their intended purpose, as well as improving the integrity of data moving within the CBT system. My argument is that, when evaluating and comparing different testing models, real financial cost drivers and practical benefits may need to supplant some of the more esoteric reliability or fit criteria offered in the psychometric literature.

A Brief Overview of Psychometric Testing Models

A comprehensive overview of the various psychometric test delivery models is beyond the scope of this paper. More encyclopedic reviews of the various models are available elsewhere (e.g., Sands, Waters, & McBride; 1997; Luecht & Nungester, 1998, 2000; Patsula & Hambleton, 1999; Parshall, Spray, Kalohn, & Davey, 2002; Folk & Smith, 2002; Jodoin, Zenisky, & Hambleton, 2002;

van der Linden, 2000; Luecht, 2000, 2005a, 2005b, in press). Nonetheless, it is useful to provide a brief overview and highlight come of the similarities and differences of these models. Advantages and disadvantages of these models are addressed further in the context of the costs and benefits described in this paper.

Multiple Fixed Tests

Multiple fixed tests (MFTs) are characterized as parallel, preconstructed, intact test forms that are administered by computer to large numbers of students (Parshall, Spray, Kalohn, & Davey, 2002). Large numbers of MFTs can be constructed simultaneously, using automated test assembly (ATA) to ensure that every test form meets that same common set of statistical and content specifications. MFTs are directly analogous to having a large number of fixed-item paper-and-pencil test forms. A special case of MFTs are linear-on-the-fly tests (LOFT—see Gibson & Weiner, 1998; Folk & Smith, 2002). A LOFT constructs each test form in real-time or immediately prior to testing. Obviously, when MFTs or LOFTs are preconstructed, there are opportunities for subject-matter experts (SMEs) to review each for match to statistical and test content specifications.

Computer-Adaptive Tests

The traditional approach to computer-adaptive testing (CAT) is well known. Under CAT, items are adaptively selected to maximize score precision for each examinee (Lord, 1977, 1980; Kingbury & Zara, 1989). After administering a few start-up items, a provisional score is computed. The CAT

item selection algorithm is activated to select the next item. A provisional score is recomputed and the process continues until one of several stopping rules is satisfied: (1) a fixed test length has been reached; (2) a pre-specified level of measurement precision (standard error or error variance of the provisional score) is reached; or (3) in criterion-referenced testing situations, such as in licensure or certification testing, it is clear that an examinee's proficiency is probabilistically above or below a specific threshold, such as a passing score. Most modern CAT item selection algorithms simultaneously attempt to balance content and account for the overexposure of the most informative test items (e.g., Hetter & Sympson, 1997; Stocking & Lewis, 1998; Revuela & Ponsoda, 1998; Robin, 2001). More recent variations on the CAT theme include constrained CAT using "shadow tests" (van der Linden & Reese, 1998; van der Linden, 2000) and stratified CAT (Chang & Ying, 1999; Chang, Qian, & Ying, 2001).

Multistage Tests

Multistage tests come in several varieties, including computerized mastery tests (Lewis & Sheehan, 1990; Adema, 1990); computerized adaptive testlets (Wainer & Keily, 1987); and preconstructed, computer-adaptive multistage tests (Luecht & Nungester, 1999; Luecht, 2000; 2003). These multistage testing models all involve clustering items into pre-assembled units called "testlets" (Wainer & Keily, 1987). However, MSTs differ in subtle ways in terms of the measurement properties of each testlet (mean difficulty, location of maximum IRT information), how content balancing and other test assembly

constraints are handled (e.g., held constant for all tests, balanced in a compensatory fashion over testlets along a prescribed pathway), how scoring is carried out (real-time IRT scoring versus use of number-correct look-up tables), and whether or not testlets are pre-packaged into "panels" (Luecht & Nungester, 1998).

Missing Indicators: the Real Financial Costs and Operational Benefits of CBT

Many of the critical indicators that could inform policy decisions or choices regarding various CBT models are seldom considered in research studies reported in the psychometric and educational measurement literature. These missing indicators often relate to real costs of testing, or perceived benefits that actually make a difference to examinees and testing organizations. As psychometricians, we sometimes focus on academic criteria like "reductions in standard errors of estimate" without serious regard as to how much it will cost to implement and maintain a particular CBT model as part of an ongoing testing enterprise.

The undeniable fact is that most testing programs moving to CBT have experienced the need to substantially increase the testing fees charged to their candidates. These substantial cost and fee increases were never predicted by the earlier advocates of CBT or CAT—in fact, most of the research suggested that testing would be less expensive under CBT and CAT. However, the reality is that testing fee increases from 200 to 500 percent over paper-and-pencil testing (PPT) costs are not at all uncommon and policy makers are faced with justifying

the cost by nebulous "value-added" arguments. In addition, many testing organizations are shocked to discover that they must invest tens or hundreds of thousands of dollars—sometimes millions of dollars—of reserve operating funds to redesign and re-engineer most of their software systems and procedures to support the transition to CBT (e.g., Mills 2004; Luecht, 2005, in press). These investments may not be recovered for many years to come.

The more honest reality is that test delivery models like CFT, CAT, or MST can do very little to defray the most serious costs of transitioning to CBT and maintaining the testing enterprise over time—however, we can try. Three of the greatest costs associated with computer-based testing are: (1) the cost of item production and associated test development to support continuous or near-continuous testing; (2) the cost of test administration; and (3) the cost of redesigning/re-engineering systems and procedures for continuous or near-continuous CBT. To put some perspective on the subsequent discussion of CBT delivery models costs and benefits, it may be important to explain these costs in more depth and, where possible, suggest several indices that may directly or indirectly prove useful in the discussion to follow.

An important financial indicator related to item production costs is the **average cost-per-item** (ACPI). ACPI is commonly used in test development and includes the cost of initial item authoring and editing, experimentally trying out the item, and ultimately, publishing each item. ACPI typically runs from several hundred to more than fifteen hundred dollars per item. Some computer

simulations may cost thousands of dollars per item to develop. These types of item development costs accrue from paying item writers, editors, and test publication specialists, as well as administrative, processing, and analysis costs associated with item tryouts. Different types of items may have different costs, depending on the extent of development and analysis work involved. For example, complex computer-simulation items usually require substantially more work to design and author the final items, and more empirical response data to develop the final answer keys. Item costs should also include differences in the length of time items are active for use (e.g., average time to item attrition). Item attrition can occur because of loss during tryout—usually signaling problems in item development, because of dated content, or because of security risks associated with easily memorized test materials that are quickly distributed on the Internet through examinee collaboration networks. Simply put, classes of items or content areas with higher item attrition will tend to cost more per item.

The importance of ACPI stems from the fact that substantial increases in item production will result if more tests are administered more often under CBT than under PPT (Luecht, 2005a; 2005b). For example, Mills and Stocking (1996) estimated that test item pools need to be four to ten times larger under CAT than under PPT. Accordingly, if a testing program needs 200 items per year at a cost of $500 per item, the annual test development costs would be $100,000. If Mills'

and Stocking's estimates are correct[1], then test development costs would rise to between $400,000 and $1,000,000 to maintain the same testing program using a CAT delivery model.

A second useful financial index is the **cost per testing event** (CPTE). The CPTE index includes the testing seat time (usually negotiated at hourly rates), plus fixed per examinee fees for registration, test driver usage, and other administrative services. Some test delivery models promise to reduce testing time, but the actual cost reduction may end up being miniscule. That is, reduced testing time does not always result in a lower average CPTE value for two reasons. First, as part of the contract with each testing organization, commercial testing vendors often negotiate a fixed or minimum number of hours for testing each candidate. Therefore, even if testing time can theoretically be reduced, the testing organization (and, by extension, the examinee) may still pay for the minimum testing time. Second, hourly rates and fees at testing centers are often based upon various volume factors. For example, testing fees for a four-hour examination can usually be guaranteed at a slightly lower rate per hour than for a one-hour examination. Accordingly, a testing program may be able to reduce the CPTE but then be required to pay more per hour for testing seat time.

A third financial indicator is the accumulated, nonrecurring costs associated with redesigning and re-engineering the multitude of software[2] and

[1] These estimates may be dated. However, they still appear reasonable, today.
[2] Software licensing costs associated with item authoring and item banking may be negotiated at a fixed rate, on a time-limited licensing arrangement, or some combination.

human-based procedural systems for item development and banking, test assembly, composition, and publishing, examinee registration and scheduling, test delivery; and post-examination processing (see, for example, Luecht, 2001, 2002, in press). These **systems design costs** (SDC) can range from thousands to millions of dollars. Tough choices must often be made by testing organizations as to whether to contract for, lease, or purchase expertise, systems, and services to develop those same capabilities in house. The extent of design work and upfront costs required to develop or redesign/re-engineer in-house testing support systems is usually greater than contracting for, leasing, or purchasing existing services from testing vendors and other contractors. On the other hand, the potential flexibility in customizing the systems, as well as considerations of long-term recurring lease/licensing costs may make the investment in in-house systems an attractive option. Most testing organizations opt for a blended model that includes developing some systems in-house (e.g., item authoring, registration, psychometric analysis, reporting), licensing other systems (item banking, test assembly, etc.), and contracting for still other systems (e.g., scheduling, test delivery).

As suggested earlier, it is difficult to directly link particular CBT delivery models to reductions in the any of these [ACPI, CPTE, or SDC] financial indicators. It is equally difficult to concretely link choices between particular CBT delivery models, or their features, to real or perceived benefits experienced by examinees and testing organizations. For examinees, perhaps the most tangible

documented benefit of CBT involves the faster turn-around of scores and related test results. Obtaining their results more quickly allows the examinees to more rapidly pursue career or educational opportunities. It also makes it easier for those examinees that fail the examination, or otherwise wish to attempt to improve their scores, to schedule a retest. For testing organizations, there may be three apparent benefits of moving to CBT. One perceived benefit is that testing programs can implement adaptive testing models like CAT or MST that are not practical to implement using paper-and-pencil. This allows testing organizations to claim that they are "on the cutting edge" or "technologically savvy". A second benefit is more tangible and covers the many data management and examination processing efficiencies that can be realized, once an appropriate CBT infrastructure has been implemented. For example, the physical mail shipment and reconciliation of test booklets and scanning of thousands of answer sheets each examination cycle is completely replaced by the almost instantaneous and secure electronic transfer of test data between test delivery centers and test processing facilities, with far less human intervention and opportunities for lost or stolen test materials. Similarly, cumbersome and time-consuming test development and test analysis procedures that once relied heavily upon human intervention must, of necessity, are streamlined and at least semi-automated to support an ongoing CBT enterprise. This streamlining process often leads to better documentation of the processing steps in test development and test analysis and may help identify high cost, redundant, and

otherwise inefficient or error-prone procedures and system components. This is one area where the various CBT models differ. Some models add subtle complexity to the software systems and/or reduce opportunities for quality controls to be implemented. As a result, system performance or data quality may deteriorate or be difficult to even investigate. A final benefit is that CBT allows testing organizations to implement computerized performance exercises and simulations meant to tap skills that cannot be easily assessed using multiple-choice and other objective item response formats.

Possibly Overstating the Importance of Maximum Information/Reliability

Maximizing measurement information or the reliability of a test is put forth as the primary goal of test construction, especially among CAT advocates. Indeed, most of the comparisons among different CBT models—certainly those comparing CAT to other models—present differences in test information curves, standard error curves, or false-positive and false-negative decisions errors.

Two of the most common indicators used to quantify the costs or benefits of particular testing models are: (1) relative efficiency and (2) reductions in test length. **Relative efficiency** (RE) refers to the proportional improvement in test information (score precision), relative to some baseline test. RE is computed as the ratio of test information functions or reciprocal error variances for two tests (Lord, 1980). The RE index can further 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 17.0, the adaptive test is said to be 170% as efficient as the fixed-item test. Relative efficiency depends on two factors. The first factor is the baseline test information function 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, 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 25-item adaptive test can provide the same precision as a 100-item non-adaptive test, there is a obvious reduction in the amount of test materials needed and less testing time needed (assuming, of course, that a shorter test ought to take substantially less time than a longer test). Early adaptive testing research reported that typical fixed-length academic achievement tests used could be shortened by half by moving to a CAT (Wainer, 1993).

These two indices are often used as the primary motivations for moving to some type of adaptive testing. However, I would argue that RE and reductions in test length are sometimes vastly overstated in terms of their importance for real-life testing[3]. On the surface, it may seem that any reduction in test length or testing time would reduce testing costs. Yet, that is seldom the case in practice. The reality is that many test developer specialists and test users—especially in high-stakes testing circles—seriously question whether short adaptive tests containing only 10 or 20 items are able to adequately cover enough of the content to make valid decisions or uses of scores. Their argument is that validity may be seriously impacted at the cost of any added reliability, if a test is shortened too much. Neither do improvements in measurement efficiency relate directly to financial cost reductions in ACPI or CPTE. That is, if a computer-based examination is administered at commercial CBT center, there is usually a fixed hourly rate per examinee and testing organizations are required to guarantee a minimum amount of testing time. For example, 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. Real cost savings are only realized if <u>significant</u> reductions in testing time can be demonstrated (e.g., moving from two days of testing to one day of testing).

---

[3] "Real-life testing" implies administering real tests to real examinees, in contrast to research studies that involve running thousands of computer simulations conducted with fictitious examinees sampled from some theoretical distribution, who always fit the response generating model, who never have to take bathroom breaks, and who rarely, if ever, complain about their testing experiences.

Nor do apparent reductions in test length necessarily lead to lowered item production costs or reduced overall test development costs. Consider that, although various adaptive testing models may indeed reduce the test length for individual test takers (or reduce testing time, or improve the RE relative to MFT or fixed-length paper-and-pencil test), the overall cost of developing and maintaining large item banks to support CAT under continuous or near-continuous testing can actually increase the overall costs of testing by an order of magnitude.

My point on this topic is that relative efficiency and test length reductions often produce, at best only trivial real cost-reduction benefits in real-life testing. That is not to imply that we should ignore improvements in our IRT test information function targets or discredit apparent reductions in standard errors or decision errors. I am merely emphasizing that we should not put those psychometric indicators on a pedestal as the ultimate criteria for evaluating CBT delivery models.

Some Other Useful Criteria for Evaluating CBT Models

As noted in the Introduction, there are three classes of cost-benefit-related factors that would seem useful in evaluating these CBT models: (1) real measurement efficiency; (2) testing system performance; and (3) provision for data quality control/assurance. These are described below and discussed in the context of the various CBT delivery models.

Real Measurement Efficiency

Reductions in the average cost per item (ACPI) and cost per testing event (CPTE) are real financial benefits that should be considered when evaluating competing test delivery models. However, the comparisons need to be carefully done.

Consider that many standardized tests allow approximately one to two minutes to complete a multiple-choice (MC) item. Average time allocations for other item types may be more or less. If we are able to compute the amount of IRT test information at critical decision points on the score scale per unit of time, or in terms of reliability, we should be able to demonstrate how achieving a particular level of precision reduces CPTE. Unfortunately, as noted earlier, the CPTE indicator may be subject to minimum testing time thresholds set by the contract negotiated between the testing organization and the CBT vendor. Still, it is possible to use CPTE; however, we need to increase the size of the time units employed. For example, if the critical cost or pricing unit for testing seats is an hour, then a legitimate indicator of measurement efficiency would be the proportional reduction in the number of testing hours under one model versus another (not minutes or items). Of course, this rather gross CPTE indicator puts a great deal of responsibility on each testing model to demonstrate substantial gains in efficiency. For example, if the MC items take, on average, one minute to complete, a CAT or adaptive MST that demonstrated a real improvement in CPTE would need to demonstrate the equal or better reliability or decision

accuracy than a MFT (or prior paper-and-pencil test) while reducing the test length by at least 60 items.

Reductions in ACPI are likely to be somewhat more difficult (or at least more indirect) to demonstrate. As noted, continuous or near-continuous testing has created huge demands for new tests to deal with item exposure, efforts by examinees to collaborate (cheat), and other types of security breaches over time. Therefore, the total cost of item production may go up regardless of the CBT delivery model chosen. However, some types of MSTs allow creative configurations of testlets to be used to minimize the demand for items. For example, consider the 1-3-3 computer-adaptive MST design shown in Figure 1. There are seven pre-assembled, differentially difficult testlets in the panel, each assigned to one of three stages. Routing within the panel is controlled for prescribed "pathways" (A+B+E, A+B+F, etc.). Additional details on this type of MST panel design are presented by Luecht and Nungester (1998) and elaborated by Luecht (2000, 2003).

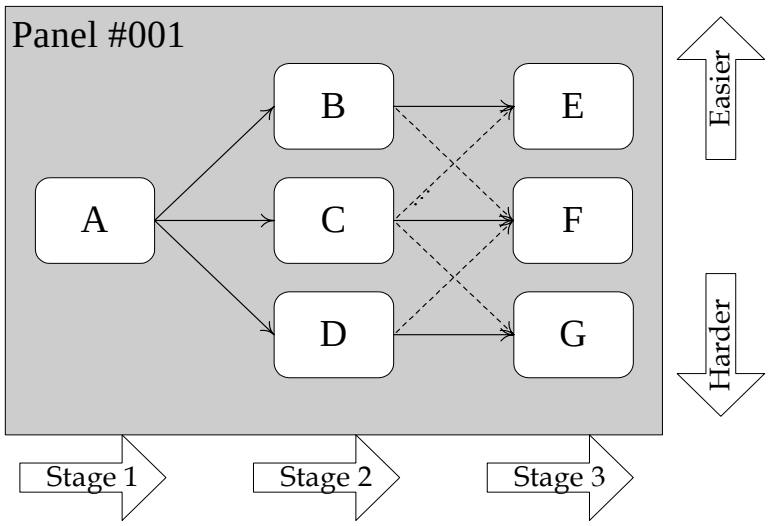

Figure 1. A 1-3-3 Computer-Adaptive Multistage Testing "Panel" Configuration

It is relatively easy to show that, by reducing the size of the testlets in the second and third stages, we can reduce the overall item demands for each panel. For example, if each testlet has constant size of 20 items (i.e., a 60-item test across three stages), each panel would require 140 items. We would therefore need an item pool of at least 700 items to build five non-overlapping panels. In contrast, by changing the size of the testlets to 40, 10, 10 for Stages 1, 2, and 3 respectively—again producing a 60-item test—each panel would require only 100 items, leading to a reduction of 200 items for each five-panel build. There are, of course, other considerations. But showing a reduction of 200 times an ACPI of $500 represents a cost savings of $100,000—a substantial savings by almost any yardstick. Furthermore, there have been some very recent applications of linear and mixed integer programming techniques to develop optimizations models for item inventory control, within the context of various MST designs (Breithaupt & Hare, 2005; Belov and Armstrong, 2005). These

optimizations strategies are extremely promising approaches for conducting evaluative comparisons relative to these types of cost reductions.

Testing System Performance

Despite technological growth and the significant improvements in testing software and database management over the past decade or so, the system performance of computers has not advanced to a point where throughput, network flow, and other performance considerations are trivial considerations. System performance relates to technical performance of the CBT system. 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, "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.

One of most effective performance-enhancing strategies is to reduce the load or demand.

Highly interactive CBT delivery models like CAT or LOFT create huge real-time demands on a test delivery file server. Consider that, after each item, the test delivery driver must engage in: (a) some type of scoring computations (e.g., computing an IRT ability estimate); (b) selection of the next item—which might involve sophisticated calls to an automated test assembly algorithm (van der Linden & Reese, 1998; van der Linden, 2000); and (c) the real-time composition of the rendered items and navigational controls seen by the examinee. If an Internet-based testing (IBT) platform is employed, other complex and often unpredictable constraints are placed on the system. In contrast, models like computer-adaptive MST employ highly structured tests comprised of pre-assembled testlets that are then prepackaged into "panels." The test delivery driver only needs to: (i) randomly select a panel, screening out any previously seen or inactive panels; (ii) administer the first testlet in the panel; (iii) compute a simple number-correct score; and (iv) look up the next testlet to administer based on a scoring table (Luecht, 2003). The use of testlets and the structured test data incorporated into a panel, as well as the simplification of the scoring and selection computations under MST, can significantly reduce computational loads. For example, using testlets has many advantages in terms of navigation and presentation. 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 testlet. This leads to improved efficiency and accuracy in rendering test materials.

When aggregated over tens of thousands of examinee-test transactions, the potential improvements in system performance may be dramatic—especially under an IBT platform. One final point is relevant in that regard. These types of system performance issues need to be evaluated under large-scale operational loads, not just under isolated trials.

Provision For Data Quality Control/Assurance

Quality control (QC) and quality assurance (QA) procedures are an integral part of any testing system, with implications for test development and forms production as well as data management, in general. The quality control and quality assurance aspects of test form composition and production are non-trivial issues for many test development experts. Even using automated test assembly (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. Under paper-and-pencil testing, many testing organizations make extensive use of committees composed of subject-matter content experts 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.

Although it may not be feasible to employ any type of quality control (QC) review for tests generated in real-time, there at least need to be quality assurance (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, since every form can be checked or at least sample audited. Some adaptive CBT models like computer-adaptive sequential testing (Luecht &

Nungester, 1998; Luecht, 2000) preconstruct and prepackage all of the pieces of a multistage adaptive test, beforehand. By preconstructing and prepackaging the adaptive test, it possible to engage in formal QC data checks and audit reviews—up to a 100 percent 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 were performed in real-time, using ATA algorithms or heuristics that are incorporated into the test delivery software, quality control may be largely non-existent. Theoretically, if the test bank or item pool were thoroughly checked before it is activated and if the computerized test delivery software and associated algorithms were 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 were 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.

The integrity of the test materials and subsequent response data are also easier to manage with structured units because the test unit data can be checked against known control parameters. In contrast, the test results data for computerized tests constructed in real-time—e.g., randomly selected LOFTs or CATs—cannot be easily checked for integrity or reconciled to any known units, because each test is a unique creation. Metrics such as the number of lost or

corrupted test result records per 1,000 examinees should be used when evaluating data integrity. Mean-time-to-detection is also a useful indicator to compare different test delivery models. In general, MFT and MST models should fair better in terms of this latter index since corrupted or improper data streams can be readily checked against the expected data for a test form or panel.

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. My 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.

Conclusions

This paper discussed some of the real costs and benefits associated with CBT and argued that, as psychometricians, perhaps we need to focus more on the actual costs of developing item pools and tests than on IRT measurement information and related psychometric criteria when evaluating different CBT models. By employing preconstructed testing units (test forms, testlets, panels, etc.) MFT and MST models may offer some apparent benefits in terms of system performance and opportunities for QC/QA. While adaptive testing may offer improved measurement efficiency, I argued that measurement efficiency indicators used in comparative studies should ideally reflect real cost savings, either in terms of reduced item/test production costs or in significant reductions in the cost per testing event.

References

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

Belov, D. I.; & Armstrong, R. (2005, April). *A Monte Carlo Approach to Design, Assemble and Evaluate Multi-Stage Adaptive Tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec. (Draft)

Breithaupt, K.; & Hare, D. R. (2005, April). *Optimal Inventory Design for High Stakes Performance Testing Programs*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec. (Draft)

Chang, H. H. & Ying, Z. (1999). A-stratified multi-stage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

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

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, New Jersey: Lawrence Erlbaum.

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.

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

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.

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. (2003, April). Exposure control using adaptive multistage item bundles. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2005). Computer-adaptive testing. *Encyclopedia of Statistics in Behavioral Science*. London: Wiley.

Luecht, R. M. (2005). Computer-based testing. *Encyclopedia of Social Measurement*. Amsterdam; London: Elsevier/Academic Press.

Luecht, R. M. (in press). Operational issues in computer-based testing. In D. Bartram and R. K. Hambleton (Eds.), *Computer-Based Testing and the Internet: Issues and Advances*. New York, NY: Wiley & Sons.

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

Mills, C. N. (2004, February). *It Cost How Much? Writing the Checks*. Symposium presentation at the Annual Meeting of the Association of Test Publishers, Indian Well, CA.

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

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.

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

Sands, W. A., Waters, B. K. & McBride, J. R. (Eds.). (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Washington, DC: American Psychological Association.

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.

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

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