A Proposed Framework of Test Administration Methods

Nathan A. Thompson

Assessment Systems Corporation

Contact:
Nathan A. Thompson, Ph.D., CCOA
Assessment Systems Corporation
2233 University Ave., Ste 200
Saint Paul, MN 55114
nthompson@assess.com

Abstract

The widespread application of personal computers to educational and psychological testing has substantially increased the number of test administration methodologies available to testing programs. Many of these mediums are referred to by their acronyms, such as CAT, CBT, CCT, and LOFT. The similarities between the acronyms and the methods themselves can be a source of confusion to testing professionals. This purpose of this paper is to suggest a hierarchical classification system for test administration methods and clarify what each entails while also briefly discussing the similarities and differences of each.

A Proposed Framework of Test Administration Methods

The use of personal computers as a medium of test administration has led to a growing number of specific approaches to administration that are usually referred to by their acronyms, including CAT, CBT, CCT, CMT, and LOFT. To a testing professional without formal psychometric training, these might seem confusing, all the more so because they overlap and nest within each other. The purpose of this paper is to suggest a categorization and definition of these approaches, as well enumerate the definitions by briefly reviewing them both with respect to practical advantages and disadvantages, and with respect to the sometimes subtle psychometric differences.

Broadly speaking, test administration is divided into three categories that all are familiar with, based on medium: live performance testing, paper-and-pencil testing (often abbreviated PPT, PNP, or P&P), and computer-based testing (CBT; Folk, March, & Hurst, 2006). Each has their advantages and disadvantages in terms of costs, logistics, reliability, and validity. For example, "real-life" performance tests are often claimed to have higher validity because the examinee is being asked to do exactly what the test is supposed to assess. However, performance tests often face higher costs, logistical problems, and reduced reliability. Moreover, as reliability is a prerequisite component of validity, the validity is not necessarily stronger.

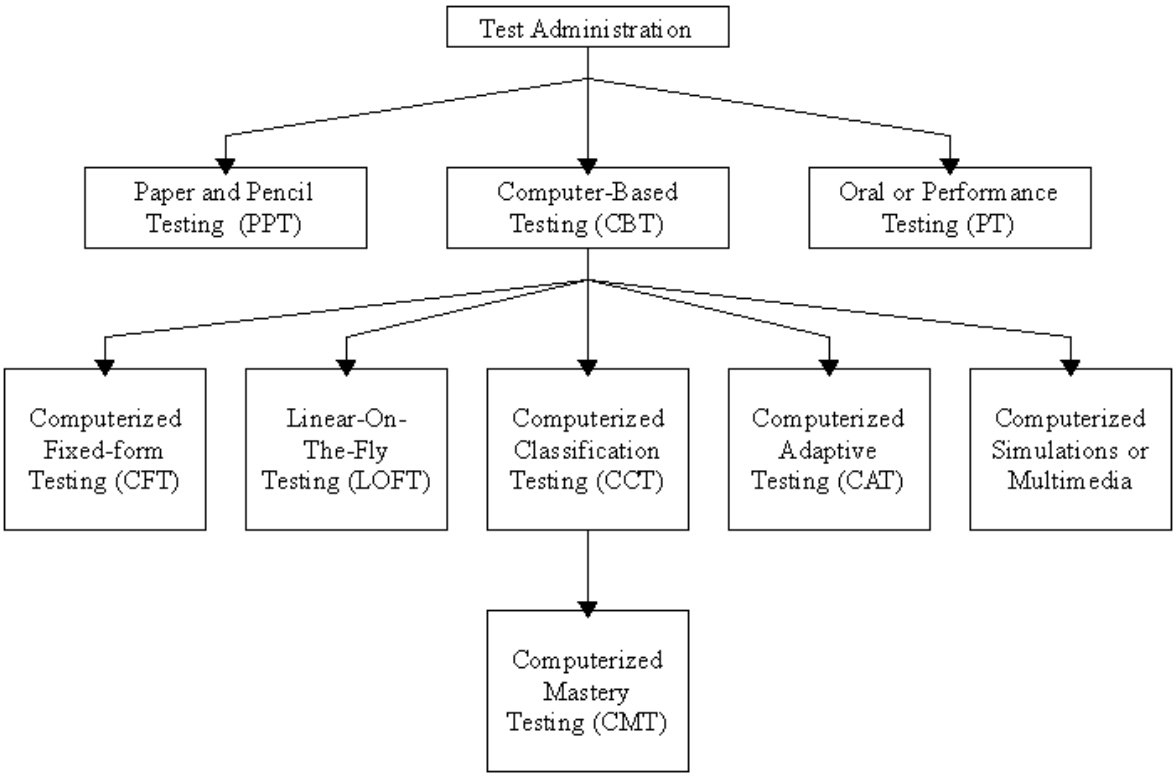
PPT and performance testing have existed for millennia. CBT, on the other hand, has obviously existed only since the advent of computers, and not become common until the availability of personal computers made it economically feasible. CBT, given its name, is a broad topic that encompasses *all* tests administered by computer, and therefore technically subsumes all other computer-related approaches (Figure 1). This list not only includes psychometric technologies such as computerized adaptive testing (CAT; Weiss & Kingsbury, 1984), computerized classification testing (CCT; Parshall, Spray, Kalohn, & Davey, 2002), computerized mastery testing (CMT; Sheehan & Lewis, 1992), linear-on-the-fly testing (LOFT; Stocking, Smith, & Swanson, 2000), and but also computerized performance simulations (Pucel & Anderson, 2003).

While these terms were not originally developed to be mutually exclusive, this categorization suggested by this paper proposes they be so for clarification reasons, with the exception of CMT, which is a subset of CCT. CCT and CMT – both sometimes referred to as sequential testing – are considered separately from CAT because they differ from CAT with respect to the two fundamental algorithms in dynamic testing, item selection and termination criterion. CAT utilizes different algorithms because the purpose of many of these tests is precise point estimation rather than classification.

However, CBT also subsumes the simple approach of a traditional fixed-form test similar to PPT, but administered via computer. Because the psychometrically advanced methods are very specific cases, the term "CBT" is often used to refer to only a fixed-form test administered on computer, since it is the "leftover" approach without its own name. For clarification, this paper will refer to this approach as computerized fixed-form testing (CFT), with CBT reserved for the broad category of all tests based on computers. Additionally, performance testing of professions that are computer based, such as software developers, will be considered as computerized simulations because there is no difference between a computerized test and a "real life" test.

Figure 1 presents a framework for test administration methods based on these considerations. The foundation of this framework is that administration method can be categorized on a broad level by the medium used: paper, computer, or real-life. A secondary level of categorization is characterized by the algorithms utilized in computerized delivery: fixed-form, LOFT, CCT, CAT, and simulations.

Figure 1: Categorization of test administration methods



Testing organizations launching a new testing program are all faced with the choice of administration method, while some testing programs that are currently active might be better served by a different administration method or a combination. The purpose of this paper is to elucidate the differences between the methods of administration in Figure 1, and how these differences translate into advantages or disadvantages for testing programs. In many cases, the advantages will vary depending on the program itself; CAT and LOFT, for instance, realize most of their advantages only with large scale testing of thousands of examinees. First, background is provided for important considerations. Then, administration methods are reviewed in ascending order of sophistication with respect to psychometrics and logistics. PPT represents the simplest form of examinations, while computerized simulation tests are the most advanced.

Important Considerations

There are many issues to consider when evaluating options for test administration. One that all are familiar with is practical considerations including cost and logistics,

psychometrics, test development, and stakeholder relations issues are all vital to the success of a testing program. Each of these issues is discussed in the next section.

Practical considerations

The two primary practical issues are cost and logistics, which are often intertwined. Take a performance test for example in which human raters are needed onsite. In this instance subject matter experts might have to be brought to a central location to serve as raters. Besides the logistics of arranging flights, accommodations, meals, and schedules, such a test also has the extensive direct costs involved in the same facets. Moreover, there are also indirect costs including the time these subject experts are required to be away from their families and jobs.

Costs undoubtedly increase with logistical requirements, but it also tends to increase with sophistication. The psychometric and test development work needed to launch an adaptive test for 1,000 examinees per year is more substantial than the work needed to launch a single fixed form for the same volume of examinees. However, such a test might reduce the logistic-associated costs with fixed-form paper and pencil testing in specific calendar windows or test security issues.

Psychometric theory

Tests must produce *valid* scores that support intended uses and interpretations, which require adherence to standards of test development, administration, and psychometrics. These professional expectations are outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), which not only provides a guide but also addresses practical concerns applicable to many testing programs. Psychometrics refers to the statistical aspects of ensuring that a test demonstrates evidence to support valid interpretations, as opposed to more practical aspects such as standardized delivery (e.g., *Standard 5.1*) and score reports sent to examinees (e.g., *Standard 5.10*).

Test scores demonstrates evidence of reliable when they provide consistent measurement, that is, we can expect an examinee to get approximately the same score each time, within an acceptable amount of random measurement error. Score interpretations for a test are valid if they provide the meaning they intend and are supported by sufficient evidence that is designed to evaluate those intended uses. For instance, if a test claims to measure basic knowledge reflecting competency in a profession, then high scores should indicate high levels of knowledge and, by extension, competency.

Psychometric analysis during the test development cycle is essential to evaluating evidence of reliability and validity, and the theoretical approach to psychometrics provides an important distinction between test administration methods. There are two dominant paradigms in psychometric theory: classical test theory (CTT) is based on evaluation of psychometric characteristics using the collection of items in a test, while item response theory (IRT) is based on mathematical models that focus on individual item characteristics and often requires larger sample sizes. A common rule of thumb regarding IRT sample size is that the one-parameter model requires at least 100 examinees, while the three-parameter model requires at least 500 examinees (Yoes,

1995). In-depth discussion of the psychometric differences can be found in many sources (e.g., Fan, 1998; Embretson & Reise, 2000; Stage, 2003).

Test development

For the purposes of this paper, test development refers to the operational processes directly related to the test, including blueprint development, item writing, item revision, form assembly, and score reporting. These processes, combined with psychometric analysis, comprise the technical steps required to develop a test (Downing, 2006).

Perhaps the largest difference between test administration methods in terms of test development is that live performance tests and task simulations have a completely different conceptualization of the test "item." Each item represents a task, and is often much more involved than a traditional item format such as multiple choice, fill-in-the-blank, or true/false. This in turn affects the number of tasks; a simulation with only 5 tasks might takes as long to complete as a test of 100 multiple choice items. The development of these tasks for assessment purposes is also much more involved. It can take extensive time and effort to break down complex professional processes into simple, scorable steps.

For more traditional item formats, such as multiple-choice, the largest difference is between form-based and pool-based methods. Form-based methods administer specific sets of items, with the sets chosen prior to publication to meet content, format, and psychometric constraints. CAT, CCT, and LOFT are pool-based methods, where the test is published as a relatively large pool of items, and the set to be administered to an examinee is selected at the time of the test.

Stakeholder relations

On top of all considerations internal to a testing program, the relationship with stakeholders must never be forgotten (Jones, 2000). It is necessary to justify and explain the testing process, and this is easier if it is accepted by the stakeholders. A candidate population that is not particularly technologically savvy might not wholly accept tests administered via computer, especially if the tests are complex simulations. There are many such things to which examinees might object. With LOFT and CAT, every examinee might see a different set of items. With CAT, most examinees will answer approximately half the items correctly, but the resultant scores can differ widely. With a fixed-form test scored using pattern scoring with the three-parameter IRT model, it is possible for an examinee to answer fewer items correctly than another candidate, but receive a higher score. These situations are all psychometrically justifiable, but might generate confusion and protests from the examinee population.

Test Administration Methods

The mode of test administration is an essential consideration of a testing program. Unfortunately, there are few simple answers in the process of determining which method to use. Some programs might be able to fully utilize advanced methods, while others

might be better served by traditional PPT even if they have the sample size for advanced methods.

Paper and pencil testing (PPT)

Practical considerations

PPT represents the least sophisticated medium of test administration, which is also its primary advantage, as it often translates into the lowest cost and greatest control over the security and exposure of items. However, a significant drawback to PPT is the frequent necessity of test administration windows and centralized administration. If a test is administered on computer via a nationwide vendor, it can be delivered to examinees on-demand at hundreds of locations around the world, even for smaller testing programs. With large testing programs, PPT can also be administered feasibly at a large number of locations, but the number of locations is limited for small testing programs. A program with only several examinees in each state might, for example, deliver their test only at an annual industry conference. It would not be economically feasible to arrange an administration in each state for only a few examinees. The primary advantage of low cost is eliminated as the number of examinees at each site decreases.

Furthermore, these administrations are limited by the necessity of windows; often, a given set of PPT forms will be used for a short period of time, such as one month or even one day. The test is then not administered while the results are being processed, until the next window occurs. On the other hand, window testing is not as much of an issue if examinees are prepared in corresponding windows; if candidates graduate from school in December and June every year, then windows in January and July are appropriate.

Another significant drawback is that score reporting can be substantially delayed. After the test administration window, answer sheets are returned to the testing program for processing. The results must then be reviewed, usually including psychometric analyses such as an item analysis (to determine if certain items should be eliminated) and equating (to determine comparable scores on multiple forms). This leads to the well-known stressful period, as long as six or eight weeks, where examinees wait to receive their scores on which their career can depend. Computerized administration, on the other hand, potentially allows for the test to be scored immediately and a detailed score report to be printed out and taken home.

Psychometrics

A major manifestation of the simplicity of PPT is scoring; while not always utilized, PPT exams can be scored with a number-correct approach. CAT and CCT, on the other hand, more readily provide opportunities to use scoring algorithms based on the mathematical models of IRT. Performance testing introduces its own set of often greater issues when developing scoring systems. Because a classical test theory approach works as well as IRT for PPT, this is typically the approach used by testing programs with insufficient sample sizes for IRT.

Test development

Form construction represents another way that PPT is simpler than other methods. PPT and CFT administer exams in a fixed-form method, or the traditional approach of giving every candidate the same set of items. If examinee volume is large and security is a concern, then multiple forms can be used, but they are still constructed with the same number of items to match the program's blueprint. Moreover, the number of additional forms is limited in number; because it is usually infeasible to print booklets and answer sheets for 50 or 100 different forms, the number of forms used per testing window rarely exceeds single digits. Fixed forms must also be constructed well before live administration.

The simplicity is also beneficial in terms of item development. PPT does not require complex items, as do performance tests and computer simulations. Items are typically administered using a multiple choice format, but may include short answer, extended response, and/or essay questions. The extent to which raters are needed to score items balances the costs of development versus scoring.

Stakeholders

Because the majority of the population is familiar with multiple choice exams administered on paper, the level of disapproval by stakeholders may be quite low. Examinees do not require substantive orientation/education to understand or participate in this particular delivery method. As with any delivery process, examinees should know what content to expect and what their results mean.

Performance Testing

Practical considerations

While performance testing in an informal sense may appear less sophisticated than PPT, standardized performance testing that is used for high stakes purposes is more involved. A programmatic approach must be taken, first identifying and defining the tasks, then establishing a systematic scoring system that can be universally applied to all examinees, and then addressing the sometimes substantial logistical needs. For example, a performance test for a certification in ophthalmic assisting requires all the necessary clinical equipment, patients that meet specific guidelines regarding their medical condition, the recruitment and training of raters who are expert enough to serve as scorers for the test, and evidence that demonstrate that these raters can then consistently and accurately apply the scoring criteria across candidates. This often leads to high costs, which are justified by the claim that the given tasks cannot be assessed with acceptable fidelity in any other way.

Psychometrics

A justification for live performance testing on the psychometric level is that we are possibly sacrificing some classical reliability evidence for greater fidelity and face validity, and if the tradeoff is positive, the use of live performance testing is warranted. However, this is not necessarily the case. Reliability can be reduced by the introduction of the human element in the form of raters. However, a primary goal of the scoring system is that it be designed to eliminate as much subjectivity as possible. This is often done by the application of objective checklists rather than subjective rating scales. With

proper scoring systems and training of expert raters, combined with a selection of tasks appropriate for assessment, this effect can be ameliorated.

Adequate rater training needs to be stressed (*Standard 2.23*). If there are any stakes associated with the test, its administration must be standardized as much as possible to ensure that each examinee had a similar experience and the same chance to pass the exam. It is therefore vitally important that raters are trained in the protocol of the exam and in correct, systematic use of the scoring system.

Test development

Performance testing leads to a test development process that is much more involved than the typical multiple choice item. To create an objective test delivery process, the tasks may need to be broken down into checklists or step lists which might be ordered, non-ordered, or semi-ordered. Related to the delivery, a scoring system must be designed to translate the results of the checklists into examinee scores or classifications. For example, different numbers of points might be assigned to each step, and examinees must attain a certain total score to pass.

Not only must these items be developed, but also be applied and revised if necessary. The amount of effort needed to pretest and revise such items is also obviously more than that required for items with simpler structure. However, because these items are more complex and tend to be longer, the bank will be much smaller and more manageable. Similarly, form construction is simpler because five or ten items might be sufficient.

Stakeholders

This issue often finds its way into stakeholder relations. High exam costs are never popular, but the high fidelity of appropriately representing the content domain is a major advantage that can sometimes not be escaped. It is impossible to argue about "tricky questions" when examinees are asked perform an actual task that is known ahead of time, and those in a supervisory role know that if examinees have passed the test, they are capable of doing the task.

Computerized Fixed-form Testing (CFT)

Practical considerations

CFT is the same as PPT with the exception that the test items are presented to examinees on a computer rather than on paper. It is used because the application of computers alleviates some of the drawbacks of PPT, such as the fact that tests are able to be scored at the time of administration as long as the test is developed in such a way that no psychometric steps are required after administration.

If the test is delivered via a vendor that has multiple sites, CFT also has the advantage of being able to be administered at as many sites as possible on a continuous basis, rather than a few sites during limited time windows.

However, the increased availability comes with increased costs. The vendor must support testing centers at sometimes hundreds of sites worldwide, and these costs are transferred to the testing programs. This approach can be cheaper than PPT administrations by the testing program themselves at a large number of sites. The

comparison must be done on a case-by-case basis. An additional potential cost is the increased security and item exposure risk that comes with greater accessibility.

Psychometrics

The application of immediate score reporting means that item analysis, equating, and scaling must all be completed before the form is launched. This in turn often means that items must be pretested, or administered in an unscored fashion in previous forms. With PPT, because there is time to analyze and revise or delete items between the administration and scoring time, new items that have not been pretested can potentially be used as scored, live items.

Similarly, equating and scaling can take place after administration with PPT, but not with CFT with immediate score reporting. The equating and scaling can be done with either IRT or CTT.

Test development

Test development processes for CFT differ little from PPT. Items are typically multiple-choice format items, but can also be short answer or extended response items that are written to the test blueprints. They are assembled into fixed forms and then administered. The difference is, as previously mentioned, that items must be pretested and reviewed before being included as live items, which requires that new items must be included as unscored items to obtain data for analysis. This is necessary if the program wants to provide score reports to the examinees immediately following testing.

Stakeholders

Because it is computerized, CFT might receive some resistance from stakeholders that are accustomed to PPT (Folk, March, & Hurst, 2006). As with all CBT, diplomatic efforts must be made to communicate the benefits of computerized testing to stakeholders (Jones, 2000), such as the availability of immediate score reporting. However, because the tests are still of the conventional fixed-form multiple-choice format that most people are familiar with, the resistance should be less than the more sophisticated methods that remain to be discussed.

Linear-On-The-Fly Testing (LOFT)

Practical considerations

Also referred to as Automated Test Assembly (ATA), LOFT is similar to CFT in that each examinee is administered a fixed number of items via computer (Stocking, Smith, & Swanson, 2000). The difference is that CFT utilizes a single form or several forms that are constructed before the release of the test into the field. With LOFT, the test is released to the field as a pool of items, and the actual set to be administered to an examinee is not selected until they take the test. The selection is determined by algorithms that take into account relevant variables specified by the testing program, such as content constraints and item statistics.

The purpose of this approach is to enhance security of large testing programs that have continuous administration. If 1,000 examinees take a fixed-form test every week, it is likely that many items will become compromised by being dispersed candidate-to-candidate or via in Internet, even if there are several forms. LOFT addresses this by

giving every candidate a test that is equivalent in terms of content and statistical properties, but the actual set and order of the items are allowed to vary. When item pools are of sufficient size, very few, if any, candidates will see the same form, reducing one avenue of item compromise.

A drawback to LOFT is that it makes use of administration algorithms and IRT, but does not reduce test length and, by extension, the time is takes to estimate an examinee's abilities (i.e., seat time). CCT and CAT also utilize larger item pools that have sufficient density and comparatively complex algorithms during the actual administration. However, the efficiency of the administration algorithm may reduce test length for an examinee by 50%.

Because LOFT is usually based on IRT and a primary goal is to enhance security in high-volume situations, LOFT is most appropriate for very large testing programs where continuous testing is desirable because examinees do not arrive in cohorts (e.g., graduating in December and June), and it is necessary to have all examinees receive the same number of items. If the latter is not necessary, CCT and CAT present a more efficient alternative.

Psychometrics

Because the forms are algorithmically constructed to meet content and statistical specifications, all items in the available pool must be psychometrically analyzed beforehand, requiring the pretesting of items. While it is possible to build the algorithms and scoring to utilize CTT, IRT enables all the forms to be built more equivalently by targeting the test information function. This then requires sample sizes that are necessitated by the choice of IRT model. However, sample size is rarely an issue because the LOFT approach is only used for high-volume exams where the large number of possible forms addresses a test security need.

Test development

Like fixed-form tests, LOFT tests commonly employ multiple choice items. While it might seem that test development might be more extensive because of the need for a sufficiently sized pool, the number of items needed for the pool is not necessarily more than the number required for multiple forms. Suppose a desired exam length is 150 items, and is administered in four forms designed with 40% overlap. The number of unique items required in this situation is 150 + 90 + 90 + 90 = 330 items. A bank of 330 items that meet the inclusion criteria (e.g., content representation, item characteristics) can be sufficient for LOFT administration of 150-item tests. Additional considerations of security, item exposure, item drift, and retake policies may also influence the size of the item pool.

The remaining aspects of test development are similar to fixed-form approaches. The development of test blueprints, long-term development of an item bank, and the pretesting of items are not necessarily different. The primary difference is that rather than develop fixed forms for release, a pool of items and a psychometric algorithm are developed.

Stakeholders

LOFT exams are administered via computer, so immediate score reporting is possible. Therefore, the resistance of examinees to testing in computerized testing centers can be an issue. However, the primary examinee issue is the fact that most examinees will receive a different set of test items. Because psychometrics is perceived as esoteric and is rarely understood by examinees, there might be claims of test unfairness even though the forms are equivalent. Therefore, it is important to educate examinees about this approach and its justifications. For instance, a converse argument is that the use of only a few fixed forms would invite test security problems, which is a much larger test fairness inequality.

Computerized Classification Testing (CCT) and Computerized Adaptive Testing (CAT)

CCT and CAT are extremely similar, differing only on a psychometric level due to different scoring, and to a moderate extent at that. Therefore, they can be considered simultaneously for the purposes of this paper.

Practical considerations

Like LOFT, CCT and CAT are often seen as sophisticated methodologies that are too expensive to be utilized except by the most massive testing programs. However, as demonstrated in the LOFT section, the number of items required for an acceptable item pool is not necessarily more than the number of unique items required for traditional fixed-form administration with multiple forms. The cost of CCT/CAT may not be prohibitive, especially if a testing program already utilizes IRT.

Psychometrics

CCT is designed only to classify examinees into broad categories, while CAT is designed to obtain a precise estimate of examinee ability. They are often grouped together due to their similarity on the surface, but there are distinct differences in the psychometric algorithms. However, the fact that CCT and CAT utilize algorithms, like LOFT, places them apart from fixed-form testing as *variable-form* or *algorithmic* tests.

CCT and CAT differ substantially from the other multiple choice methods in that they utilize interactive algorithms, which enable them to be *variable-length*. A variable-length test is one that terminates when certain statistical criteria have been satisfied; some examinees will encounter this after 20 items, while some might require 200 items. This greatly enhances test security by potentially reducing test length and examinee seat time by up to 50% on average (Weiss & Kingsbury, 1984; Stocking, Smith, & Swanson, 2000), which substantially reduces item exposure. Fixed-length tests can be designed with CCT and CAT, but such tests reduce the advantages of efficiency.

The parameters of the criteria are specified by the testing program, and include practical constraints such as test length. Because some examinees might object to a test of only 20 items – especially if they fail – a minimum of 50 or 100 items can easily be instituted. In an extreme case, it is even possible to fix the test length to be the same for all examinees. This example of modifying algorithms based on stakeholder relations eliminates the substantial reductions in item exposure and examinee seat time simply to reduce objections by examinees.

Computerized mastery testing (CMT; Sheehan & Lewis, 1992) is a special case of CCT where the number of classification categories is limited to two, and applied to a test

where the goal is to ascertain "mastery" of a knowledge domain. Mastery is generically defined and represents the defined performance level of the minimally qualified candidate. Common examples of this include certification and licensure tests.

Test development

The test development process for CCT and CAT is highly comparable to that of LOFT. Items are typically multiple-choice, calibrated with IRT, and released as a pool to be assembled into tests for each examinee. The difference is that LOFT will select all items before the examinee begins the test, while CCT and CAT interactively select items throughout the test.

Stakeholders

Not only does CAT/CCT face the issue of each examinee receiving different sets or sequences of items, they face the possibly huge issue of variable test length. Examinees will need to understand how some might fail after only 30 items while others might receive 100 items and pass; the former could try to argue that they would have passed given more items, not understanding how the item selection and scoring algorithm works. One approach that has been used to address this problem is to fix the test length for all examinees, which unfortunately reduces one of the major benefits of CAT/CCT: the potential 50% reduction in items needed. An adaptation of this approach that might be considered is to allow examinees that pass to receive fewer items, but administer a certain minimum number of items for examinees that fail.

Computerized Simulations

Practical considerations

Computerized simulations (Pucel & Anderson, 2003) represent a significant technological advance in test administration. But like computerized multiple-choice testing, the computerization does not necessarily entail greater costs, and in many cases might actually be less expensive. Computerized simulations require a much greater initial investment just to be launched, but can be more economic in the long run because the logistical problems discussed previously are reduced. There is no need to recruit patients and expert raters and fly them to a test location (or fly examinees to a location), tests can be administered year-round on demand, and the actual equipment does not need to be obtained every time a test is administered. It is important to note that the initial investment might be significantly reduced if scoring systems already exist for the live performance test that can be adapted to the simulations.

A practical consideration that is especially important with computerized simulations, given the expense of their development, is the length of time they are relevant. For a testing program whose content changes often because of new technology, the development of new simulations to keep up with this technology can be too expensive. If, however, the simulation will likely remain applicable for several years, it is more likely that the investment is worthwhile.

Psychometrics

Simulations of performance tests require that complex tasks be broken down into simple, objective components for scoring, just as live performance tests do. However,

the application of computerized scoring algorithms eliminates the need for human raters to do the scoring, which can provide two benefits. First, it eliminates the need to train human raters, which means one less thing for the testing organization to do as the assessment window nears. Second, this facilitates the potential for more reliable scores by removing human judgment as a random error factor in scoring.

Test development

While live performance tests only entail the use of the systematic breakdown for scoring; computerization requires the complete reconstruction of the complex tasks in a computer interface with maximum fidelity. This makes item development efforts greater with this approach than with any other. Just the development of storyboards for complex tasks can be an extensive production, not even considering the amount of software development that is needed to transform those storyboards into the actual test. Moreover, it can be difficult to develop alternative forms with this approach, though it might be as simple as setting up a different patient with different numbers (e.g., a medical task). However, as previously mentioned, if this amount of expense is less than the expense required to hold live performance tests, then computerized simulation can be an attractive method of assessment.

Stakeholders

An additional hurdle to the initial cost is implementation of simulations with the stakeholders. However, while the simulation of complex tasks might be found objectionable by many stakeholders, cost savings can offer a compelling argument. Computerized administration greatly enhances the standardization of the test administration providing a greater likelihood of the same experience for all examinees. For example, with live testing, certain examinees might feel that their raters are harsher than other raters who could have potentially evaluated their work. Computerized simulation allows the computer to systematically apply the same rating criteria to all examinees, thus making the scores and subsequent decisions more defensible.

Conclusions

The choice of test administration method is an important one that faces all testing programs. In some cases, the decision might be made by outside influences, but in most cases the directors of the program evaluate methods in terms of practical considerations, applicability, psychometric characteristics, and effects on stakeholder relations. The choice can be quite difficult given the number of methods available and the similarities among them.

The size of a testing program is of paramount importance when evaluating test administration methods. More sophistication is available with greater volume, as much for psychometric reasons as for the fact that larger testing organizations have more resources. However, the need for large volume to effectively utilize greater sophistication is often overestimated. If a testing program has sufficient volume and resources to produce four fixed-forms every year, they likely have the capability to adopt a LOFT or CCT/CAT approach and realize the accompanying benefits if the disadvantages of these administration approaches do not outweigh the advantages.

Likewise, the transition from live performance tests to computerized simulations may offer benefits for organizations willing to make the initial investment when weighed against the costs.

Nevertheless, the traditional fixed-form PPT is still the most viable method of administration for many testing programs. In low-stakes or low-volume situations, it is usually not economically feasible to publish computerized test and reserve computerized testing centers for proctored administration.

Adequate knowledge of the details of the methods discussed in this paper is essential for the choice of the most appropriate method. Testing program directors should take the time to further understand the methods and be able to make an objective comparison, including details such as projected costs and volumes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Downing, S.M. (2006). Twelve steps for effective test development. In Downing, S.M. & Haldyna, T.M. (Eds.) *Handbook of Test Development* (pp. 3-26). Mahwah, NJ: Erlbaum.

Embretson, S.E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fan**,** X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58* (3), 357-381.

Feng, Y. (1994). *From the Imperial Examination to the National College Entrance Examination: the Dynamics of Political Centralism in China's Educational Enterprise*. Paper presented at the Annual Meeting of Association for the Study of Higher Education, Tuscon, AZ.

Folk, L.C., March, J.Z., Hurst, R.D. (2006) A Comparison of Linear, Fixed-Form Computer-Based Testing versus Traditional Paper-and-Pencil-Format Testing in Veterinary Medical Education. *Journal of Veterinary Medical Education, 33*(3), 455-464.

Jones, J.P. (2000). Promoting stakeholder acceptance of CBT. *Journal of Applied Testing and Technology*, Volume 2. Accessed 5/22/08 from http://www.testpublishers.org/jattart.htm.

Pucel, D.J. and Anderson, L.D. (2003). Developing computer simulation performance tests: Challenges and criteria. *Computers and Advanced Technology in Education* (pp. 170-174). ISBN: 0-88986-361-X. International Association of Science and Technology Development: Calgary, AB.

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

Stage, C. (2003). Classical test theory or item response theory: the Swedish experience. *Educational Measurement* No 42. Umeå, Sweden: University of Umeå, Department of Educational Measurement.

Stocking, M.L., Smith, R., & Swanson, L. (2000). *An investigation of approaches to computerizing the GRE subject tests*. Research Report 00-04. Princton, NJ: Educational Testing Service.

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

Yoes, M.E. (1995). *An updated comparison of microcomputer-based item parameter estimation procedures used with the 3-parameter IRT model* (Technical Report 95-1). Saint Paul, MN: Assessment Systems Corporation.