Practical Assessment, Research, and Evaluation

Volume 18 *Volume 18, 2013* Article 14

2013

Test Administration Models

Kirk A. Becker

Betty A. Bergstrom

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 14, December 2013 ISSN 1531-7714

Test Administration Models

Kirk A. Becker & Betty A. Bergstrom, *Pearson*

The need for increased exam security, improved test formats, more flexible scheduling, better measurement, and more efficient administrative processes has caused testing agencies to consider converting the administration of their exams from paper-and-pencil to computer-based testing (CBT). Many decisions must be made in order to provide an optimal examination program from the perspectives of validity, customer service, and cost. Methods to administer computer-based testing include Computer Adaptive, Linear-On-The-Fly, Multistage testing, and multiple fixed forms. Each of these methods has pros and cons that must be considered in relation to the purpose and characteristics of the exam. Issues of security, access, psychometrics, and cheating will also be addressed.

Computer-based testing offers many options for test administration that are not possible with paper-based tests. These options have implications for test reliability and length, security, cost and upkeep, and other program needs. Based on an organization's needs and goals, it is possible to work through the characteristics of different testing models in order to evaluate how appropriate they are for a given program.

This article outlines the types of computer-based test (CBT) administration models available to testing organizations. When choosing a test design, organizations will want to evaluate the usefulness and feasibility of each model in order to choose the optimal model for their program. The following test administration models, while by no means exhaustive, represent the vast majority of testing models currently in use in computer-based testing.

Model Definitions

Fixed-Form (Linear)

While computer-based testing makes numerous testing models possible, linear test forms are still used by many testing organizations. These tests are similar to paper test forms in that the same set of test items is administered to all test takers who receive a given test form. CBT, however, typically administers only 1 item at a time, and frequently the order of test items is randomized. In both paper and CBT administration of linear forms, a limited number of parallel forms containing non-overlapping or partially overlapping item sets are typically constructed.

Linear-on-the-Fly Testing (LOFT)

The first nationwide computer-based testing program, the National Association of Securities Dealers (NASD) exam, began administering computer-based tests (CBT) in 1979 using this model. LOFT (Gibson & Weiner, 1998; Stocking, Smith, & Swanson, 2000) is a test administration model designed to address item-security concerns with linear forms. LOFT increases security by limiting the exposure of all items. This model makes use of a large, calibrated item pool to individually construct a test form for each test taker. A fixed-length test is assembled for each test taker, either at the beginning of the testing session or as the test is administered. Items are selected to satisfy a set of specified content and statistical constraints irrespective of the test taker's ability or responses to previous items. Item Response Theory methods are then typically used to score the test based on item statistics and test taker performance.

Pallet Assembly Model

A new addition to the family of CBT administration models, "Pallet Assembly" has been proposed by Boyle, Jones and Matthews-Lopez (2012). This design, an adaptation of LOFT, posits "pallets" of multiple test forms that are pre-assembled using automated test assembly techniques. Pre-equated test forms are built to specific psychometric and content targets. When

exams contain testlets or sets of items, Pallet Assembly can provide consistent score precision at the cut, positive item bank utilization, precise exposure control, decreased pairwise form overlap, precise adherence to test specifications, and simplified scoring. Although the authors note a slight compromise in score precision relative to adaptive models, the offset can be positive gains in exposure control. The Boyle et al research demonstrates that, when the purpose of the exam is credentialing, Pallet Assembly (which centers test information on the cut score rather than on test taker ability) provides good measurement precision, lower production costs, and increased security of test forms.

Computer Adaptive Testing (CAT)

Computerized adaptive testing has been a popular computer-based administration model since the 1990s because it enables shorter tests, greater reliability, and good test security. CAT operates on the principle that items that are too easy or too difficult for a test taker contribute little information about that test taker's ability (Bergstrom, Lunz, 1999; Green, Bock, Humphreys, Linn & Reckase, 1984). As a test taker takes a CAT, the estimate of his/her ability is continually estimated based on all items presented to that point in the test. The computer algorithm selects the next "best" item available given all test specifications and the current estimate of examinee ability. In this way, items that are too hard or too easy for the test taker are not administered, and the examination given is individualized. Competence is continually assessed, and the difficulty of the test is "targeted" or "tailored" to the estimated ability of the test taker. Efficiency is gained because each test taker answers items appropriate to his or her ability, so test length can be shortened without sacrificing reliability. With nearly all implementations of CAT, test takers are not able to review previously answered items.

Barely Adaptive Testing (BAT)

Way (2010) used the term "Barely Adaptive Test" to refer to a partially targeted test used as an interim step when transitioning to CAT. Test items are selected based on the relative ability of the test taker (e.g., high vs. low ability), but they are not precisely targeted to the exact ability estimate. Using the randomesque item selection procedures discussed by Kingsbury & Zara (1989) and Bergstrom, Lunz, & Gershon (1992), it is possible to widen the item selection criteria to include relatively large numbers of items. As Muckle et al (2008) discussed, this process can dramatically improve test security while minimally impacting test precision.

Computer Adaptive Multistage Testing (MST)

Multi-Stage Testing is similar to CAT in that test taker performance on previous items determines which items are seen next. Unlike CAT, MST administers sets of items in modules or testlets. Test takers therefore receive a tailored test, allowing for increased reliability and decreased test length, while also allowing for greater control of the individual modules by test developers. MST also allows test takers to review items within a module while most CAT implementations do not allow review once an item is submitted.

Evaluation Continua

This paper presents six continua that are useful in evaluating the appropriateness of different testing models to the needs of a given program. Like the list of models, these continua are not exhaustive. Our goal is to demonstrate how the process of lining up potential testing models along meaningful evaluation continua can facilitate the discussion of which models are appropriate for the current needs of a given testing program.

1. Test developer control: The items administered on a test may be individually selected and/or reviewed by test development staff, or they may be drawn from a pool of items by a computer algorithm. While form-based versus pool-based testing models define the ends of this continua, the middle is defined by models involving a combination of both. An algorithm can select from pre-packaged item sets (as in MST), or fixed forms may be largely built by automated test assembly algorithms but with some human involvement. Linear test forms may also incorporate pretest items randomly selected from a pool. The extent to which item selection is formalized, the degree to which item coding and meta-data allow for automated item selection, and the comfort level of stakeholders with automatically generated tests will influence model choices on test developer control.
2. Time of assembly: Is the test assembled prior to, immediately preceding, or during administration? The time of test assembly has implications for the technology and infrastructure required to make use of some models, as well as the security of the exam.

DOI: https://doi.org/10.7275/pntr-yz21

3. Shape of test information: Non-adaptive testing models present a test with the same test information curve for all test takers. Adaptive testing models produce individually peaked levels of information for each test taker. Given the same number of items drawn from the same item bank, an adaptive testing model will always have precision greater than or equal to that of a non-adaptive model. For testing programs involving classification (e.g., pass/fail decision), high levels of precision for clearly passing or clearly failing test takers may not be necessary. Conversely, when ranking test takers (e.g., admissions testing) is the goal of a program, it may be desirable to maximize information for each test taker.
4. Level of pool utilization: Both the testing model and the item characteristics influence item selection and item pool utilization. The required test characteristics, including precision, content balancing, and length, will also influence pool use. Items that contribute minimally to the needs of a testing program will rarely be used under any model. Still, certain test models result in a greater variability in item use than others. Procedures used to improve item use (e.g., item exposure algorithms) are implemented at the expense of goals such as test information. Pool usage can be considered relative to the items currently being administered (on a set of test forms or in a pool) or relative to the set of items available (the bank).
5. Pool size required: Pool size requirements encompass both the number of items required to administer a desired test and also the number of items required to meet a test program's security and availability goals. For the purposes of this paper, we are assuming that continuous testing (testing on demand) is a goal.
6. Vulnerability to memorization: Computer-based testing provides multiple safeguards against cheating; however, no testing program is completely immune to all forms of cheating. In addition to the capability that the rich data collected during CBT administration offers for data forensics, certain test models make it more difficult to take advantage of advance knowledge of content or other forms of cheating.

Discussion of Continua Relative to Models

Test Developer Control

Figure 1 provides a visual representation of how much control test developers have over the forms assembly process. Linear forms allow for hand-crafted item selection and review of intact forms by subject matter experts (SMEs). Even when item selection methods such as automated test assembly are used to create linear forms, content developers or SMEs can still review and revise the test forms. We present Pallets to the right of linear forms (less developer control) because even though Pallets make use of linear forms, the large number of forms that need to be created and reviewed may limit the extent to which content developers and SMEs can review and refine test forms. MST is also presented as a model with less test developer control, in this case because of the algorithmic selection of item modules. While one of the advantages of MST is the ability to adapt to test taker ability while also allowing test developer review of modules, the number of potential paths through modules makes the review process more complicated than for linear forms. Finally, LOFT, BAT, and CAT are pool-based test administration models with no capability for review above the item level prior to the administration of a test to an individual test taker.

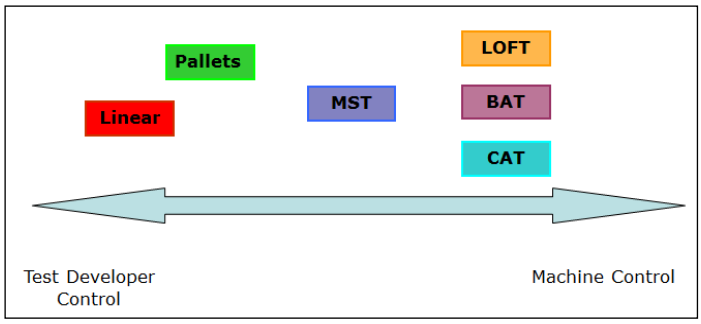

Figure 1. Level of Test Developer Control Over Forms Assembly

Time of Assembly

Figure 2 shows the distribution of test models relative to time of form assembly, with the far left showing a priori test assembly and the right side showing assembly during the test. The specific forms a test taker sees in linear and pallet models must be built and published prior to the testing event. With LOFT, the specific items that a test taker will see are typically selected prior to the administration of the first item on the exam, which could be at the time of registration, upon arrival at the testing location, or upon sitting

down at the computer to take the test. It is possible for a LOFT exam to select items (non-adaptively) during test administration; however, most or all implementations of LOFT with which the authors are familiar select items prior to administration. MST modules are assembled prior to test administration; however, the modules a test taker sees are selected during test administration. Finally, CAT and BAT necessarily select items during administration.

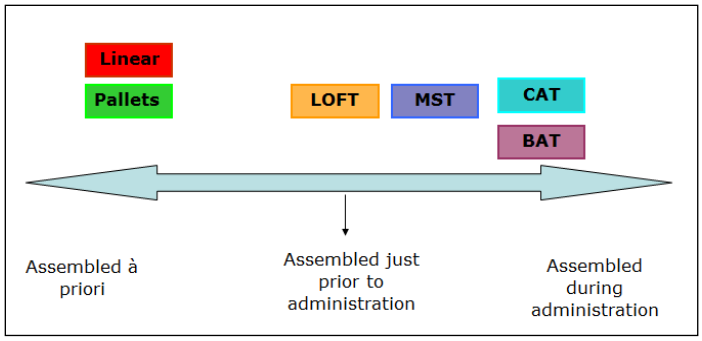

Figure 2. Time at Which Individual Test Forms are Assembled

Shape of Test Information

The relationship between testing model and the shape of test information, which determines the precision of test taker measurement, is presented in Figure 3. Adaptive testing models are designed to produce measures with less error/greater reliability than a non-adaptive test of the same length. Because it is adaptive at the item level, CAT provides the most information. MST adapts at a more molar level (modules), resulting in slightly lower test information. BAT adapts at the item level but intentionally increases the range of information in the items selected, resulting in slightly lower information as well (how much lower will depend on the characteristics of the bank and the test parameters). Generally speaking, all three of these models will produce more uniform precision across the range of test taker ability. LOFT, linear, and pallets may have comparable precision to CAT at one point on the ability scale (e.g., at the passing standard for a peaked test), but information will decrease for test takers with higher or lower ability.

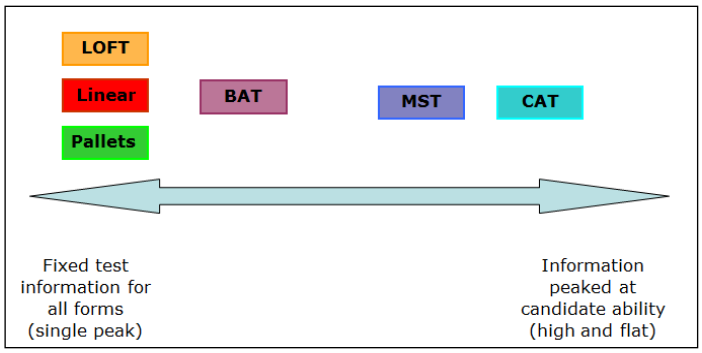

Figure 3. Shape of Test Information

Level of Pool Utilization

Frequently, the converse of test information is pool usage, and Figure 4 shows how different testing models make use of the item pool. Pool use is also highly affected by program features other than test model, so the place of models in Figure 4 could easily be shifted. Without item exposure controls, CAT pool usage will mirror the test taker population, with items that have the highest information at the population mean being used the most, and items with low information away from the population mean being used the least. The main limitation on linear forms is the number of forms built. Linear forms could be built using every available item in the item bank, or a small number of forms representing a fraction of the bank could be in use at any given time. MST can be constructed such that each module receives approximately equal exposure; however, the authors believe that most programs using MST focus on measurement precision rather than uniform exposure. A pallet design should make use of most of the available items in a bank. If some forms in the pallet design are retired for some reason, or if forms are reserved, pool use may suffer. The choice of randomesque item selection used, as well as the measurement model (e.g., Rasch or 3 parameter IRT model) will affect the way the BAT uses the available items in the pool. Finally, the test constraints and the content and statistical characteristics of the items in a LOFT pool may result in unused items.

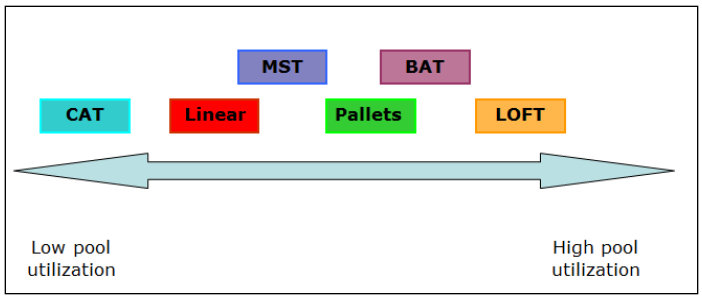

Figure 4. Level of Pool Utilization

Pool Size Required

The pool size required to operationalize each model is presented in Figure 5. Assuming a medium- to high-stakes testing program that administers exams every day, CAT will require more items than other testing models. LOFT, BAT, Pallet Assembly, and MST may be effectively implemented with fewer items than a CAT. Finally linear forms typically require fewer items overall, although the number of live, non-overlapping forms desired by a program could increase that number.

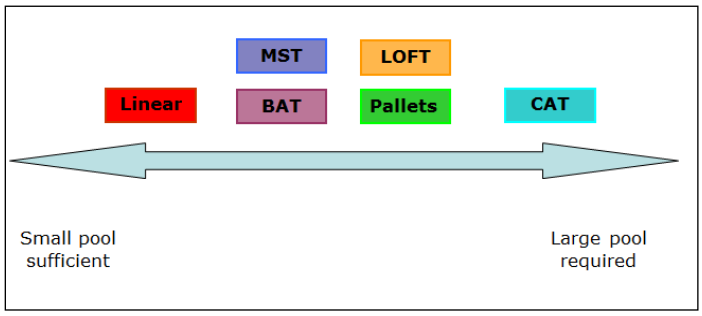

Figure 5. Pool Size Required

Vulnerability to Memorization

The final evaluation of vulnerability to memorization is provided in Figure 6. Our perspective here is relative to the amount of content a test taker would need to learn to be significantly advantaged. For example, a linear form can be compromised in its entirety, meaning that a dishonest test taker would only need to memorize items on a single form to gain a significant advantage (assuming that form was administered). Likewise with MST, because items are administered in discrete blocks, access to the content of specific blocks would provide a significant advantage. The advantage of Pallet Assembly is slightly lower than CAT, BAT, and LOFT because the items exist in discrete blocks; however, the large number of forms makes pallets very similar to pooled testing in practice. In the CAT, BAT, and LOFT models, there is a low probability of administering a particular item to a specific test taker, which means that a dishonest test taker would have to memorize an extremely large amount of content to receive a significant advantage.

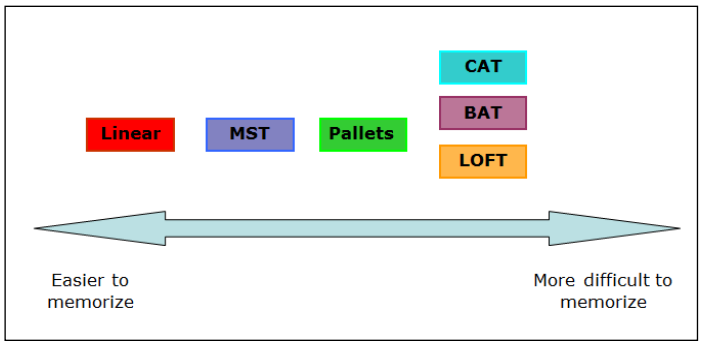

Figure 6. Vulnerability to Memorization

Key Questions to Ask When Selecting a Model

How does a testing organization decide which CBT administration models will give optimal reliability and which models will be feasible and provide an adequate level of security? Following are key questions to help in this decision:

What is the Purpose of the Test?

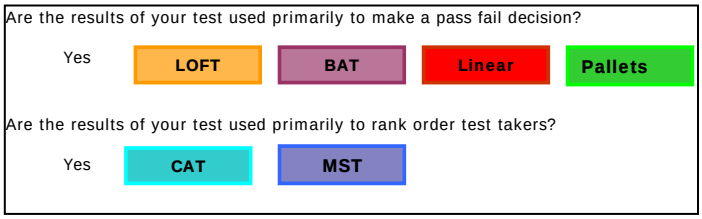

Figure 7. Test Purpose

When the results of the test are used to make a pass/fail decision, certain test designs like LOFT, BAT, or Pallet Assembly may bring higher reliability and better use of the item bank. Alternatively, if the purpose of the test is to give accurate scores across a range of ability estimates, a model like CAT may be optimal.

Does Your Organization Have the Resources to Create and Maintain a Large Item Bank?

More sophisticated models require large IRT-calibrated item banks. If your organization is just starting out, you may need to begin with linear tests and build an item bank to support one of the other CBT models.

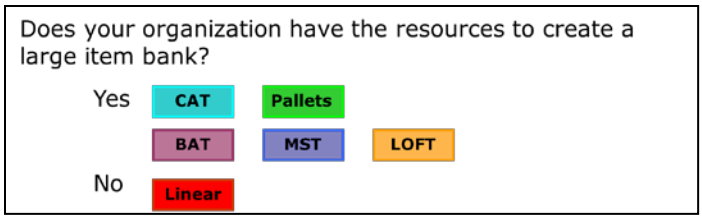

Figure 8. Program Resources

Is It Important to Your Organization to Review Intact Test Forms Prior to Administration?

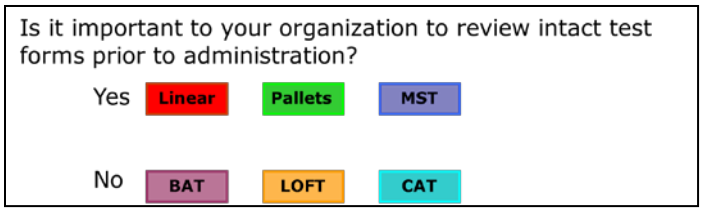

Figure 9. Need for Test Form Review

Some CBT models create the test on-the-fly or during test administration. For these models, the emphasis on quality shifts from subject matter expert review of the test form to review of the items in the bank. Organization comfort level with computer selection of items may influence your choice of CBT models.

Is It Important to Your Organization to Shorten the Length of Your Current Examination?

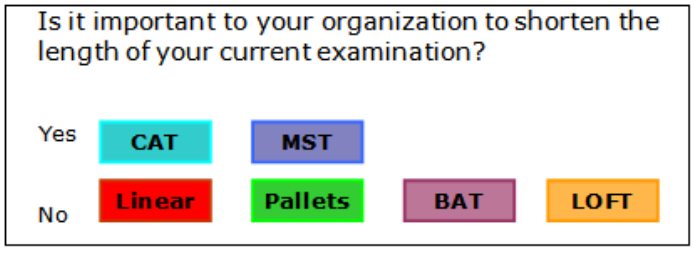

Figure 10. Test Length and Timing

Adaptive models that target the difficulty of the items to the ability of the test taker can result in greater reliability for shorter test lengths. If shortening your test is a consideration, you may want to consider one of the adaptive models.

Is Test Taker Memorization of Items (cheating) a Problem With Your Current Examination?

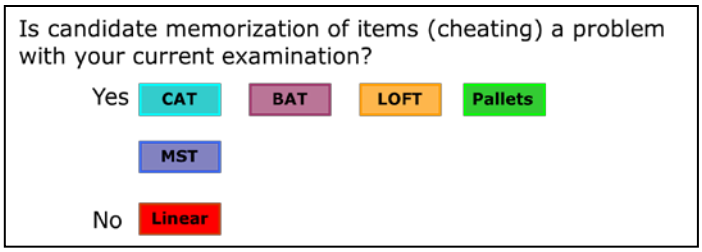

Figure 11. Memorization and Cheating

If you believe your current test is being compromised, you may want to consider a CBT model that maximizes the number of test items and/or test forms.

Discussion

Testing organizations have many choices with regard to good computer-based testing designs. Different models will be appropriate in different circumstances. First and foremost, you need to assess your organization's capability to create and maintain the administration model that you select. Large IRT-calibrated item banks give great flexibility, but they require dedicated teams of content experts and psychometricians to maintain.

Steps to help with the decision of an appropriate model include:

- Evaluate your current organizational capability for test and item bank development.
- Review the algorithms for CBT test administration models.
- Talk to organizations using the models that you are interested in exploring.
- Run simulations with your current item banks to evaluate which models can be supported.
- Create a long-term plan to take you from your current status to a more optimal model.

The benefits of using sophisticated CBT administration models are many. Don't be afraid to get started!

References

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow (Ed.), *Innovations in Computerized Testing.* Mahwah, NJ. Lawrence Erlbaum.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5(2),* 137-149.

Boyle, M., Jones, P. and Matthews-Lopez, J. (Feb, 2012). Replacing an Adaptive Decision-Making Test with a Pallet of Automatically Assembled Fixed Forms. Breakout Session at Association of Test Publishers Conference, Palm Springs, CA.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in computer-based environment. *Journal of Educational Measurement, 35(4),* 297-310.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

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

Muckle, T. J., Bergstrom, B. A., Becker, K., & Stahl, J. A. (2008). Impact of altering randomization intervals on precision of measurement and item exposure. *Journal of Applied Measurement*, 9(2), 160-167.

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

Way, D. (2010). Some perspectives on CAT for K-12 Assessments. Presented at the 2010 National Conference on Student Assessment, Detroit, MI.

Citation:

Becker, Kirk A. & Bergstrom, Betty A. (2013). Test administration models. *Practical Assessment, Research & Evaluation*, 18(14). Available online: http://pareonline.net/getvn.asp?v=18&n=14

Corresponding Author:

Kirk Becker
Pearson
1 North Dearborn Street, Suite 1050
Chicago, Illinois 60602

Email: Kirk.Becker [at] pearson.com