Assessment Systems Corporation
www.assess.com

Adaptive Testing: Is it Right for Me?

White Paper

Nathan A. Thompson, Ph.D.
Vice President, ASC
Adjunct Faculty, University of Cincinnati

Contact Information

Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, Minnesota 55114
Voice: (651) 647-9220
Fax: (651) 647-0412
E-Mail: solutions@assess.com
www.assess.com

Computer-based testing is here to stay because of it many benefits, including real-time feedback and scoring, flexibility in item formats, enhanced security, and data management capabilities. An additional benefit is the use of sophisticated delivery methodologies that capitalize on the computing power available. One example of this is linear-on-the-fly testing (also known as automated test assembly), where a form is assembled for each examinee at the beginning of the test, based on certain psychometric specifications. An even more sophisticated approach is computerized adaptive testing (CAT), where the test can update an examinee's score and determine the items they see *after every item*. Unfortunately, while the psychometric benefits are well-documented, CAT is not a viable option for many testing programs. However, this lack of viability is often overstated; CATs can be administered with a bank of as few as 200 items with a calibration sample as small as 100 examinees.

CAT assessments are interactive in their difficulty, meaning that an algorithm dynamically selects items that are neither too difficult nor too easy for each examinee. This interaction is an effort to significantly reduce the number of items administered as compared to traditional fixed-form testing, where every examinee receives the same set of items in the same order. The complex algorithm is based on the mathematical models of item response theory (IRT; Embretson & Reise, 2000), which model the probability of correctly responding to a test item as a function of a trait, ability, or knowledge. Among the many benefits of IRT is the fact that it can place items and persons on the same scale, θ, similar to a standard score scale. This is of paramount important to adaptive testing because it provides a defensible mathematical method of matching items to persons.

From a practical perspective, CAT requires five operational components:

1. Item pool – a set of items calibrated with a psychometric model (e.g., IRT);
2. Starting point – the location on the scale where the algorithm should begin;
3. Item selection method – the process of matching items to examinee θ;
4. θ estimation method – the mathematical approach to determining θ based on responses to items that have been administered to an examinee;
5. Termination criterion – the mathematical and/or practical constraint that must be satisfied for an examinee's test to end.

The item pool is a given that is utilized throughout the test. The actual CAT algorithm operates by beginning at the starting point and selecting an item. After the examinee responds, the item is scored, an updated estimate of ability is obtained, and then the algorithm checks to determine if the termination criterion has been satisfied. If it has not, the algorithm cycles back to step 3, selects another item, and repeats the process until the termination criterion is satisfied.

The reduction in number of items needed per test is due to the two intelligent, interactive components: adaptive item selection and the variable termination criterion. Adaptive item selection selects only items that are useful for a given examinee. If an item is likely too easy or too difficult to provide any information regarding a given examinee, there is little chance it will be selected. Simply put, it is a waste of time from a psychometric perspective to administer an easy item to a top student, and vice-versa. Yet traditional fixed-form tests administer many such items because they do not adaptively select items.

The variable-length termination criterion also plays a role in substantially reducing test length. The test is completed as soon as the criterion is satisfied. For example, a person can be administered as few as ten items (Eggen, 1999; Rudner, 2002), and if all are answered correctly, the algorithm might classify the person as a "pass" without requiring any more items to be administered. The converse is also true; if a person gets most or all of the items incorrect, a CAT will likely fail them after a small number of items. This virtually eliminates the possibility of persons taking the test just to memorize questions and then post them on the Internet; the examinee has to answer many items – but not too many items – correctly to continue.

CAT offers several important benefits. First, by only administering items that are of appropriate difficulty, CATs typically require only half as many (or fewer) items as a conventional fixed-form test while maintaining an equivalent level of precision (Weiss & Kingsbury, 1984). This saves substantial amounts of testing time. Second, it greatly enhances test security by not only presenting different sets of items in different orders to each examinee, but the fact that half as many items are required substantially reduces the usage of the item bank. Furthermore, CAT does not require the construction of parallel forms or form equating, and facilitates conversion of scores onto a scale for score reporting.

Besides the benefits to the testing organization, the reduction in exam time provides benefits to the examinees and other stakeholders. A reduced exam time can increase examinee motivation. Motivation is also increased by the fact that lower-ability examinees are not discouraged by difficult items, and high-ability examinees are not bothered with very easy items. But perhaps the most important point is that the time saved can be utilized for other purposes, such as additional instruction.

Additionally, CATs are flexible in their delivery to a candidate. They can be administered over the Internet, on a local area network, on a standalone computer, or via CD or flash drive. They are not limited to major corporate testing centers. However, the same cautions regarding administration that apply to paper-and-pencil testing still apply to CAT. For example, if a test has high enough stakes that there is incentive for cheating, it is obviously not prudent for the test to be self-administered by the examinee. A CAT over the Internet would be just as ineffective as a paper-and-pencil test mailed to an examinee. Such practical issues, always important to assessments, should be evaluated on a case-by-case basis.

An important disadvantage to CAT that potentially limits the testing programs to which it can be applied is that it requires larger sample sizes than are needed to launch traditional fixed-form tests. Depending on the specific IRT model utilized, initial sample size requirements can be as high as 1,000 examinees. Nevertheless, it is possible to have an initial sample size of 100 or even smaller, especially if classical test theory or decision theory is applied (Frick, 1992; Rudner, 2002). Therefore the sample size requirement is not the roadblock that it is often perceived to be.

Because CAT pools typically require several hundred items, item pool size requirements are also often perceived as a drawback, but this is not necessarily true. Many fixed-form testing programs utilize several forms; if three forms of a 100-item test are utilized per year, with 20% overlap, a total of approximately 260 items are needed. Depending on the characteristics of the testing program, such as number of content areas, accuracy needed, and item quality, a pool of 260 items might be sufficient to launch a CAT exam. So the investment required in item development is not necessarily greater than fixed-form testing.

A substantial hurdle faced in the development of a CAT is the extensive psychometric expertise that is required. It requires specialized knowledge to properly calibrate an item bank and to set up an effective CAT. Software must be purchased or developed for three important components: IRT calibration, CAT simulations, and a CAT testing engine.

Most testing organizations have neither the knowledge nor the software. Fortunately, the necessary software is available commercially (and is free in some cases), and psychometric consultants are available to provide the necessary expertise. Of course, none of this is an issue if you are purchasing CAT tests that have already been developed, only if you are attempting to develop a CAT from an existing item bank.

However, the most important issue regarding the application of CAT to a testing program is whether there is a business case. Are the foreseen benefits worth the investment required by the program? Obviously, this is also on a case by case basis. If you are purchasing software for use by psychometricians at your organization, you would need the first two programs presented in Table 1, and one of the FastTEST delivery platforms.

Table 1: Software necessary for CAT development

Software	Purpose	Academic Price
Xcalibre 4	IRT calibration and item analysis report	$249
CATSim	CAT simulations to determine CAT design	$399
FastTEST Pro	PC item banker and testing engine	$1399 plus testing station licenses
FastTEST Web	Web-based item banker and testing engine	Customized

Such prices are not out of reach for most organizations that have the examinee volume to be interested in CAT. The time spent by psychometricians and test developers also needs to be estimated; but again, that is not necessarily more than required to develop fixed-form tests in your organization.

In conclusion, CAT is an assessment technology that can greatly benefit many testing programs. Programs that require efficient, accurate assessment could be better served by a CAT approach as opposed to traditional fixed-form testing, whether delivered via paper-and-pencil or computer-based. Tests are even shorter if the purpose is a simple pass/fail classification rather than precise scores. But on top of the psychometric and security benefits, the accurate scores and testing time savings represent opportunities to further the mission of the testing organization.

Websites for software above:

Software	URL
Xcalibre 4	http://www.assess.com/xcart/product.php?productid=569
CATSim	http://www.assess.com/xcart/product.php?productid=555
FastTEST Pro	http://www.assess.com/xcart/product.php?productid=570
FastTEST Web	www.fasttestweb.com

References

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

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-261.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8,* 187-213.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

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