Computerized Adaptive Testing and eLearning

Just as electronically delivered instruction represents an important advance in the technology of learning, electronically delivered assessment is a correspondingly important advance. E-learning can provide an efficient, controlled method of instruction at a distance, and computer-administered testing provides an opportunity for a similarly efficient and controlled method of assessment. However, just as all methods of e-learning are not created equal, not all computerized assessments are created equal. Methods can differ in the time required on the part of the examinee, the time on the part of the test developers, fidelity, and sophistication – all of which are interrelated.

One of the most sophisticated technology-enabled assessment technologies is computerized adaptive testing (CAT). 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 the standard normal 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:

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 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 items needed per test is due to the two intelligent, interactive components: adaptive item selection and the variable termination criterion. Adaptive item selection only selects items that are useful for a given examinee. If an item is likely too easy or too difficult to provide any information regarding an examinee, there is little chance it will be selected. From a psychometric perspective, these items would be wasted if administered, but 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 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.

An important drawback to CAT that potentially limits the testing programs it can be applied to 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 1000 examinees. Nevertheless, it is possible to have an initial sample size of 100 or smaller, especially if classical test theory is applied (Frick, 1992).

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 requirements of the testing program, 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.

Additionally, CATs are flexible in their transmission 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. 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 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.

In conclusion, CAT is an advance in assessment technology whose benefits and goals strongly align with those of e-learning. Instructional 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. Besides the benefits to the organization, the reduction in exam time can be utilized for more detailed feedback to examinees or additional instruction. Therefore, CAT represents an opportunity to augment instruction by more efficient use of student time.

To learn more, visit assess.com/adaptive-testing

NATHAN THOMPSON, PHD
CEO
ASSESSMENT SYSTEMS CORPORATION
solutions@assess.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.