

COMPUTER ADAPTIVE TESTING: THE PROS AND THE CONS

BY JEFF WEEKLEY, PH.D., RESEARCH ANALYST, KENEXA[®]

In a typical ability test, groups of items are administered, and the number of items an examinee answers correctly is used to estimate his/her ability. The more items an individual answers correctly, the greater his/her ability is assumed to be. However, because everyone responds to every item, most examinees are administered items that are either too easy or too difficult. Adding these items to the test is similar to adding constants to a score; they provide relatively little information about the examinee's ability level.

In Computer Adaptive Testing (CAT), the estimated ability level of the examinee is used to predict the probability of getting an item correct. With no knowledge about an examinee in the beginning, it is assumed he/she is of average ability. CAT begins by administering an item of average difficulty. An examinee who correctly answers the first item is then given a more difficult item; if that item is answered correctly, the computer administers an even more difficult item. Conversely, an examinee who gets the first item wrong is administered an easier question. In short, the computer interactively adjusts the difficulty of the items administered based on the success or failure of the test taker.

CAT consistently administers items appropriate to the examinee, which maximizes the information gained about the examinee's level of ability. CAT stops administering items when certain criteria are met, such as when the standard error of the ability estimate falls below a set threshold, indicating a reliable assessment (CAT is commonly based upon item response theory, which enables the test developer to calculate the reliability of a test taker's ability). Other stopping criteria include time and the number of items administered.

One basic requirement of CAT is that the content domain be one-dimensional. In other words, CAT can only be used to measure a single ability or skill. Where multiple skills/abilities need to be assessed, it is necessary to develop a separate CAT for each domain.

Assuming there is only one skill/ability to be assessed, the challenge that remains is development of a high-quality item pool. CAT developers must ensure that the test measures the examinee's true ability level. Because examinees (i.e., applicants) may be of high or low ability levels, the CAT must be able to assess across the entire range of ability represented in the applicant population. This is accomplished by the development of many items for low-ability examinees, average-ability examinees and high-ability examinees (as well as points in between). Some have argued that an effective CAT can be developed with only 100 high quality items distributed evenly across ability levels (more items are always preferred). For very "high stakes" exams, or those covering a very broad domain, many items may be necessary.

NON-TRADITIONAL DOMAINS

Recently, there have been developments aimed at using CAT in non-traditional domains. For example, some preliminary research has hinted at the potential for using CAT to measure personality traits. In addition to reducing assessment time, this approach has the potential

advantage of reducing faking on such measures. The same computer adaptive logic has been applied to performance measurement, with raters being presented new items based on ratings of previous items. One can easily imagine a computer adaptive multi-rater feedback process wherein a group of raters converges on a more accurate competency measure in less time.

Despite the development challenges, the many benefits of CAT to both examinees and administrators alike ensure that this technology will see increasingly greater use in the future.

See Figure 1 for advantages and limitations of CAT. ■

FIGURE 1: ADVANTAGES AND DISADVANTAGES OF CAT OVER TRADITIONAL ABILITY TESTING

ADVANTAGES	DISADVANTAGES
<p>Increased Accuracy Each examinee takes a unique test that is tailored to his or her ability level. Questions that have low information value about the test taker’s proficiency are avoided. The result of this approach is higher precision across a wider range of ability levels. CAT provides accurate scores over a wide range of abilities while traditional tests are usually most accurate for average examinees.</p>	<p>CAT is not applicable for all subjects and skills, especially for those where the item response theory cannot be applied.</p>
<p>Challenge Test takers are challenged by test items at an appropriate level. They are not discouraged or annoyed by items that are far above or below their ability level.</p>	<p>Traditional ability tests are constructed to assess a specific ability; item constraints may result in an overly narrow selection of questions presented to test takers.</p>
<p>Improved Test Security Because each test is unique to the examinee, it is more difficult to capture the entire pool of items. Doing so would require the careful collaboration of many examinees of varying ability levels.</p>	<p>The constraints imposed in selecting the next question can, in practice, result in test takers completing sets of items that are broadly the same—losing the advantage over traditional tests.</p>
<p>Time Savings Less time is needed to administer CAT than fixed-item tests because fewer items are needed to achieve acceptable accuracy. CAT reduces testing time by more than 50%, while maintaining a comparable level of reliability.</p>	<p>CAT requires careful item calibration. This, in turn, requires that extensive data be collected on a large item pool. The development of a sufficiently large item pool is one of the biggest constraints to the widespread use of CAT.</p>
	<p>CAT requires computers for test administration and the examinees must be minimally computer literate. While the lack of computers is becoming less of a limitation, many facilities still do not have the necessary hardware available.</p>
	<p>With each examinee receiving a different set of questions, there can be perceived inequities when examinees get together to “compare notes.”</p>

About the Authors

Jeff Weekley, Ph.D.

Jeff Weekley, Ph.D., serves as a Kenexa® executive consultant. Before joining Kenexa, Dr. Weekley held senior human resource management positions with Zale Corporation, Southland Corporation and Greyhound Lines. Weekley has designed, validated and implemented numerous large-scale employee selection systems for retail store managers and associates, customer service employees, health care providers, hospitality guest-contact employees, drivers and mechanics. He has also created many organizational development programs including succession planning processes, performance management systems, leadership training and internal customer satisfaction surveys.

Dr. Weekley has authored numerous articles for the Journal of Applied Psychology, Personnel Psychology, Academy of Management Journal, Human Performance and Journal of Management. He is a member of the American Psychological Association, the Society for Industrial and Organizational Psychology and the Academy of Management. Weekley holds a Doctorate degree in organizational behavior from the University of Texas at Dallas, a Master of Science degree in industrial and organizational psychology and Bachelor of Science degree in psychology from Texas A&M University.

www.kenexa.com
contactus@kenexa.com