Computer Adaptive Testing (CAT) in an Employment Context

Mike Fetzer, Amanda Dainis, Sara Lambert, SHLPreVisor
Adam Meade, North Carolina State University
Out of all testing methods available today, computer adaptive testing (CAT) provides the maximal balance of accuracy and efficiency. Not to be confused with computer-based testing (a term that refers to any type of test administered using a computer), CAT is a method of testing that "adapts" to each individual test taker. In other words, CAT provides a tailored testing experience based on the test taker's level of knowledge, skill, ability, or other characteristic being evaluated by the test. As a result, a CAT requires fewer items and produces a more accurate score than traditional “static” or randomly-generated tests.

Over the past few decades, CAT has been used extensively in the areas of education, certification, and licensure. There are an estimated 30 operational CAT programs all over the world that evaluate 4 to 6 million men, women, and children each year. Before now, the only large-scale CAT program used for hiring purposes was the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB evaluates vocational abilities relevant to a wide variety of jobs in the military, and the CAT version has been in use since 1982.

SHLPreVisor has made CAT technology available for use in hiring and developing employees in organizations of all sizes and industries. Currently, SHLPreVisor offers nearly 300 different tests in CAT format, with more currently under development. Our online platform enables these tests to be delivered to anyone with a standard Internet connection, anywhere in the world, and at any time. More importantly, CAT provides a much greater level of test security and is thus a better option for remote, unsupervised (unproctored) testing situations.

How does CAT work and why is it better than traditional testing?

The basic premise of CAT is quite simple – the computer selects test items to administer based on the ability level of the test taker. If the test taker gets an item correct, the computer will select a more difficult item to administer next. If the test taker gets an item wrong, the computer will select a subsequently easier item. This process continues until either the computer has enough information to produce a reliable test score or the test taker has reached the maximum number of items to be administered, whichever comes first. As simple as this may sound, the science behind CAT is anything but basic. Only through the use of a complex series of analyses, sophisticated algorithms, and large item pools is a test able to be administered in CAT format.

In short, CAT represents the most sophisticated, advanced approach to testing available with current technology and psychometric (testing) theory. Widespread CAT testing has emerged fairly recently due to the computational requirements involved with CAT scoring and item selection. With increasing use of remote (unproctored) Internet testing for pre-employment purposes, CAT allows for maximal test security while simultaneously providing accurate estimates of examinee ability with maximal efficiency. No other method of assessment allows for the instantaneous creation of a test geared toward the ability of a single test taker. Moreover, through the use of item response theory (IRT) and item selection algorithms, a CAT requires fewer items to obtain more precise estimates of examinee ability than static or randomly-generated tests.

Some noted benefits of CATs include:

- Tests can be administered without the presence of a trained test administrator.
- Test scores are available immediately.
- Test security may be increased because there are no hard copy tests to be compromised, and varying item exposure leads to a reduction in examinee discussion about certain items.
- Under most conditions, less time is needed to administer CATs than fixed-item tests since fewer items are needed to achieve equal precision.
- Shorter testing times can reduce fatigue, thus removing a source of measurement error in the testing process.
- CATs can provide equally accurate scores over a wide range of abilities, while traditional tests are usually most accurate for average examinees.
- CATs can be administered in an unproctored setting more easily than fixed length tests as few items will be common to multiple test takers.
1. HISTORY AND CURRENT USE OF CAT

The first large-scale adaptive test can be traced back to the beginning of the 20th century. Alfred Binet’s intelligence testing processes were designed to align with the examinee’s IQ through the presentation of items that varied in level of difficulty, based on the examinee’s responses to previous questions. Of course, in 1905, this process was accomplished manually by the test administrator, who utilized booklets of test items divided into various difficulty levels.

While the premise for adaptive testing was born in the early 1900s, use and interest in the procedure did not become widespread until later that century. The 1960s and 1970s saw a great deal of growth in adaptive testing applications. This escalation was due to newly developed computer technology as well as increased research in the area of item response theory (IRT). Computers allowed researchers to implement the premise of administration that Binet had worked with decades earlier, but with much less tedious, manual effort. The U.S. Department of Defense became interested in the extensive research that had been conducted over the years on CAT and tried to implement its use within the Army, Navy and Air Force. Unfortunately, the sizeable computers of the time period yielded insufficient power and incurred too much cost to be successful.

Despite the difficulties with technology that were experienced in the 1960s and 1970s, IRT theorists pursued the identification of psychometric models that could be used when computerized adaptive testing became more accessible. Improvements in technology led to the availability of personal computers in the early 1980s, and these new machines had enough power to run test administrations adaptively. Additionally, they were less expensive to purchase and operate than the larger computers of the 1970s. Thus, in the 1980s, many researchers implemented various types of computer adaptive testing programs. These CAT programs differed from each other in their item-selection methodology. Some of programs were pre-structured, meaning that the path followed for item selection allowed only a limited amount of individualization (as was the case in Binet’s administration of the Intelligence test). Examples of pre-structured procedures include pyramidal, flexilevel and stradaptive item selection procedures (Kingsbury & Zara, 1989).

More advanced techniques for CAT administration also became available in this time period. In variable step-size item selection models, the branching involved is interactive to the examinees’ overall ability estimate and allows for an indefinite number of item-selection paths. Today’s CAT engines most commonly make use of one of the variable step-size models such as the Maximum Information model or models which implement Bayesian techniques.

Notable CAT Programs

As noted above, CAT has been used extensively for education, certification, and licensure programs as well as within the military for selection purposes. The number of CAT programs is increasing each year (see Appendix), and we’ve selected five of these major programs to describe in further detail below.

**Armed Services Vocational Aptitude Battery (ASVAB)**

The Armed Services Vocational Aptitude Battery (ASVAB) was initially developed as a paper-and-pencil test in the 1960’s by the U.S. Department of Defense. The instrument is still widely used, testing approximately 700,000 students in 12,000 high schools annually (Lee, 2006). It is a multidimensional instrument, assessing ten vocational abilities (such as Mathematics Knowledge, Electronics Information, and Mechanical Comprehension) that help to determine potential placement for examinees in military occupations. There are currently three forms of the ASVAB, one of which is computer-adaptive (the CAT-ASVAB) that was developed in 1982. Research has shown the CAT-ASVAB requires a shorter amount of testing time than the other forms (Powers, 2007).

**Graduate Management Admission Test (GMAT-CAT)**

ETS also administers the Graduate Management Admission Test (GMAT-CAT) to those wishing to gain admission to graduate masters (typically MBA) programs. Approximately two-thirds of graduate business schools throughout the world require a GMAT score from applicants (www.west.net, 2007). The abilities that are purported to be measured by the test cover three areas: Analytical Writing, Quantitative Ability, and Verbal Ability. Currently, the computer-adaptive version of the GMAT, which was implemented in 1997, is the only form of the test given in North America.
Microsoft® Certified Professional Exams
To provide certification to information technology professionals, Microsoft® develops and administers both adaptive and fixed-format tests that assess certain skill areas in the technology industry. There are approximately 2 million Microsoft® Certified individuals throughout the world (www.microsoft.com, 2007).

American Institute of Certified Public Accountants Exam (AICPA)
Each year, the Uniform Certified Public Accountants Examination is administered to 110,000 people who are attempting to become Certified Public Accountants (www.cpa-exam.org, 2007). The test is composed of multiple-choice testlets and simulations that cover four areas: Auditing and Attestation, Financial Accounting and Reporting, Regulation, and Business Environment and Concepts. The multiple-choice testlets in the first three areas were converted to an adaptive format in 2004, but all the simulated case studies are static. The fourth area, Business Environment and Concepts, is currently considered to be undergoing a practice analysis and is composed of three, multiple-choice, static testlets.

2. IRT & CAT BASICS

In order to fully understand how a CAT operates, it is important to have a basic knowledge of the science that drives it. The first area to become familiar with is item response theory (IRT). IRT has entered the mainstream of all aspects of the testing field: selection, education, military, and cognitive ability areas. Many of the large-scale, standardized tests are now developed, scored, and revised under the IRT theoretical framework. This shift from classical test theory to item response theory can be attributed to the “more theoretically justifiable measurement principals and the greater potential to solve practical measurement problems” (Embreton and Reise, 2000). It is a model-based measurement framework, which works to estimate the examinee’s ability on a certain construct by gleaning information from their responses and from certain properties of the items themselves.

Basic IRT Concepts

To begin a discussion about IRT, it is useful to differentiate between ability scores and test scores. An ability score is the true ability level of a person for the construct being tested. Regardless of how difficult or easy the test, a person’s true ability score is innate and remains the same, and is therefore test independent. A test score, on the other hand, is test dependent, in that the difficulty or ease of the test will affect the resulting value. The difference between these two types of scores manifests itself in an important way: two people with the same ability level may end up with different test scores if one person receives a much more difficult test than the other.

Item Response Theory is a mathematical model which allows a description (score) of an examinee to be formed that is independent of the particular test and the particular items administered. The score received at the end of a test administration is an estimate of the examinee’s ability level, not a test score as described above. This ability level is often referred to as theta (θ) and has a scale that typically ranges from -3 to +3. On the theta scale, 0 indicates an average level of ability while negative numbers indicate less than average ability and positive numbers indicate greater than average ability.

Assumptions

The capability of IRT methodology to estimate an examinee’s ability level rather than merely produce a test-dependent score stems from assumptions concerning the data. The following two assumptions must be met in order to justify using an IRT model:

1) Unidimensionality. For an IRT model to estimate an examinee’s ability level properly, it is necessary that the set of items administered only measure one single construct or latent trait (such as HTML programming ability). When there is only one ability being measured, it can be assumed that examinee responses to the test items are due to that ability and that ability alone.

2) Local Independence. The concept of local independence refers to the idea that the single ability being measured by the test is the only factor that influences how a person responds to the item. A math test including an item which held some clue to the answer in the question text would not meet the assumption of local independence, as the clue is a second factor above math ability which influences a person’s response to the item. (Hambleton, Swaminathan, Rogers, 1991)
IRT Models

After the assumptions have been met, the model to use for CAT administration is determined. Over the years, many models have been formed under the IRT framework, however this section will discuss the three most common models.

The one parameter logistic model (1-PL), also known as the Rasch model, estimates a person’s ability \( \theta \) by making use of just one parameter. This parameter is difficulty, often denoted as \( b \). For each item administered, the model uses a logistic function containing both the item’s difficulty level and the examinee’s estimated theta to obtain the probability that the examinee will answer the item correctly. The actual value of \( b \) indicates the ability level at which the examinee would have a 50% chance of answering the item correctly.

This relationship is seen graphically below, where item difficulty (\( b \)) and ability (\( \theta \)) are measured on the same dimension and the probability of a correct response is measured on the other. The plot of the logistic function below, or the Item Characteristic Curve (ICC) as it is also called, indicates that this particular item has a difficulty of 0, or in other words, a person of average ability has a 50% chance of answering this item correctly.

As the ICC shifts to the right (remaining parallel to the existing ICC), the function represents a more difficult item. As it moves to the right, one can see that it would take a person of above average ability (1, 2, 3) to have a 50% chance of answering the item correctly. Conversely, as the ICC shifts to the left, an easier item is depicted because a person with below average ability (-1, -2, -3) could still have a 50% chance of answering the item correctly.

It is important to note that within the 1-parameter model, the ICC’s will always be parallel to one another, regardless of difficulty level. A change in difficulty level of an item does not alter the shape of the ICC, it simply shifts the ICC to the left and to the right. While the 1-parameter model offers the benefit of being the simplest model, it often times will not fit the data due to the restriction of parallel ICCs. When items do not share the same slope, also called item discrimination, they will not meet the restriction of parallel ICCs. If this occurs, the 1-PL model is not sufficient, and the 2 parameter logistic model (2-PL) should be applied to the data.

The 2 parameter logistic model (2-PL) is similar to the 1-PL, except that a second parameter is taken into account when estimating examinee ability. This parameter is item discrimination, or slope, also denoted as \( a \). The larger the value of the slope, the more discriminating and useful an item will be. In the 2-PL, regardless of differing or similar difficulty levels, the slopes of the items are free to differ. This can be seen graphically below.
The comparison of Item 1 (blue) and Item 2 (red) is an example of how ICCs look for two items that have different slopes and different difficulty levels. The properties of Item 1 are a slope of 1.7 and a difficulty of -0.5 where the properties of Item 2 are a slope of .6 and a difficulty of 1.0.

Introducing this second parameter increases the chances that an IRT model will fit the data and will be applicable for a particular item pool because the restriction of parallel ICCs does not have to be met. Still, there is a third model which increases the chances of fitting the data more accurately, known as the 3-Parameter logistic model (3-PL). The 3-PL model is the most used in large scale testing programs (Wainer, 2000).

This model incorporates a third parameter above difficulty and item discrimination into the model; this parameter is often referred to as the guessing parameter, denoted as $c$. Particularly relevant to multiple choice tests, the guessing parameter attempts to take into account the idea that a person with very low ability has the opportunity to obtain a correct response simply based on chance. The 3-PL takes guessing into account by implementing a floor on the probability of getting the item correct (Wainer, 2000). The 3 parameters affect the relationship between ability and the probability of a correct response on the item, as represented graphically below.
The lower limit for the probability of getting an item correct is now set at .25 rather than at 0, as it was in both the 1-PL and the 2-PL models above. This particular ICC represents an item of average difficulty ($b=0$) and a fairly discriminating slope ($a=1$).

Once an IRT model is chosen and the parameters for each item are estimated, a test can be administered using CAT technology. As with IRT, there are some basic CAT concepts and processes that should be reviewed in order to fully appreciate the underlying technology.

**Basic CAT Concepts**

The central concept of CAT is to best match an item to the test taker’s ability level ($\theta$). By optimally matching $\theta$ and item difficulty, the maximum amount of information is learned about the test taker’s ability level. In order for CATs to function properly, the following two criteria must be satisfied: (a) a large database of items exists for which the item properties are already known, and (b) all items are unidimensional; they measure the same general construct or trait.

Imagine a test administered to a group of examinees which is far too easy for the group. Such a test would afford very little discrimination among examinee ability levels as nearly everyone would obtain a perfect score. A higher quality (and more useful) test would allow for maximum variance in examinee test scores, enabling distinctions among test takers to be made. Such a test would incorporate many items which approximately half of the examinees would get correct and half would get incorrect.

The same concepts apply for a single respondent. If a respondent has previously answered several difficult items correctly (and thus is very likely to be of high ability), very little is learned about this examinee by administering a very easy item. Conversely, little is learned by administering very difficult items to very low ability examinees. However, this is exactly what happens in traditional tests in which a single version is designed and administered to all examinees. Because most test scores must typically discriminate among respondents at all levels of ability, items vary widely in difficulty and cover much of the item difficulty continuum. As a result of the diversity of item difficulty, the test is not optimally constructed for any examinee.

In a CAT, the goal is to obtain as much information about the test taker’s ability using as few items as possible (with consideration also given to test security, item exposure rates, etc.). In order to do so, a basic cycle is established in which the following primary tasks are accomplished:

- Test taker’s ability is estimated.
- Most appropriate item is chosen based on that ability estimate.
- Test taker responds to item.

This basic cycle is repeated until one of several possible stopping rules is reached (these rules will be discussed in the next section).

The flowchart below serves as an illustration of the item selection, scoring, and ability estimation process implemented by the CAT engine.
There are several potential rules for selecting the most appropriate item, just as there are several ways to estimate test taker ability. In general, though, the item chosen will maximize the information available about the test taker's ability. As stated previously, the concept of information is directly related to the slope of the ICC at the level of ability in question. Thus, by choosing items with high ICC slopes (i.e., parameters) and difficulty values closely aligned with the current theta estimate, discrimination is maximized and much information is gained about the examinee.

Stopping Rules

All CATs must incorporate some type of stopping rule to terminate the item selection/ability estimation cycle. These stopping rules can include one or more of the following:

- A minimum number of items must be administered.
  - This rule may take priority over other stopping rules such that no exam can terminate before the minimum number of items is reached.
- A maximum number of items have been administered.
- A maximum time limit is reached, at either the item or test level
- The minimum level of precision (or accuracy) for the ability level estimate ($\theta$) has been reached.
  - The precision of $\theta$ is represented by the standard error, similar to the concept of reliability in static tests

3. DEVELOPMENT OF A CAT

The process undertaken to produce a CAT is much more involved in comparison to the development process of a static, classical test theory-based test. However, the benefits greatly outweigh the costs related to time, resource requirements, and enhanced rigor. While most steps of the process are similar to those in a static-test development process, the extent and scale of the steps are quite different.

Item development for a CAT is the first and perhaps lengthiest step in the development process. The final pool of items should consist of approximately five to ten times the number of items that an examinee will eventually see when taking the test. Thus, for a 30-item test, an item bank of 150-300 quality items is highly recommended. Item writing, in of itself, is a tedious and rigorous process. Developing the initial set of items that will eventually be reduced through the analysis process is a major undertaking, as upwards of 400 items may be needed in order to get to a final pool of 150-300 items.

Once the initial item pool is established, data is collected on each item. IRT analyses typically require at least 300 data points for each item, with 500 being preferred. Since it is not advisable to attempt to get 300 people to complete all items in the initial item pool, often the items have to be split into sub-pools small enough to collect accurate data.

With a sufficient sample size of examinees, the item parameters (i.e., discrimination ($a$), difficulty ($b$), and guessing ($c$)) can be estimated. These parameters are used to determine which items will be retained in the final item pool, and which items will be revised or discarded. The final pool is then entered into the CAT system, which then creates optimal item selection paths for test takers.

From a continuous improvement perspective, experimental items (those that are administered to test takers but not scored) may be added to the item bank. These items are presented along with other, scored items during administration. Parameter estimates can then be yielded so that decisions about the retention or exclusion of the item from the item pool can be made, thus providing a continuous stream of new items.

Content Balancing

Many SHLPreVisor CATs will be validated using a content validation strategy in which crucial knowledge and skills (of the job for which the test is being used to select examinees) are mapped to the test content. Most SHLPreVisor CATs contain items that, while directly related to the general topic of the test (e.g., HTML), can also be grouped into various sub-topics (e.g., links, images, frames). The content validation approach will utilize the sub-topics within the CAT to establish the extent to which the content of the CAT is related to the job or role in which it is to be used for selection or development. Thus, it is important to administer a proportionate number of items from each sub-topic to ensure that each CAT
4. ADMINISTRATION OF A CAT

The administration of a SHLPreVisor CAT can occur wherever a PC and Internet connection are available. There is no special software needed or downloads required, and SHLPreVisor CATs are suitable for use in unproctored testing environments (i.e., from the applicant’s home or local library) as the adaptive nature of the test allows for enhanced security of the test content.

One of the administration aspects of a CAT is similar to other computer-based testing scenarios. The ability to deliver unproctored assessments is preferred by many hiring organizations as it reduces costs associated with test administrations/facilities and it also reduces the time it takes to complete the hiring process. Since CAT is necessarily computer-based, and can be administered in an online format, there reduces the need for a proctor. However, while static tests can be administered without a proctor, CAT provides features that are better suited to unproctored testing as not every examinee receives the same test, which makes the test content more difficult to obtain and share. With CAT there is no single answer sheet that can be compromised as each test administration consists of different items. The items seen by a specific applicant is dependent on his/her own ability level.

Initially, there may be concern that the technical complexities of CAT would require sophisticated software or other upgrades to a regular computer-based testing environment. This concern is unfounded. A normal, user-friendly PC and an internet connection (it does not have to be high-speed) are all that are necessary for an examinee to complete a CAT. Further, the item selection process is not lengthy: the CAT engine selects the next optimal item in a matter of seconds. In fact, an accurate ability score will be yielded in less time, generally, than a test score based on a classical test theory-based test. As stated in previous sections, less items are needed to accurately pinpoint the examinee’s ability in a CAT environment because the items are selected in a manner that optimally measure their ability.

5. IMPLEMENTING AND USING A CAT IN SELECTION (USE OF SCORES)

Being that a reliable, valid test can provide critical information about applicants that would be difficult to gather any other way, test scores can be used to gauge an applicant’s level of knowledge and skill regarding important aspects of the job. From a hiring decision standpoint, a score produced from a computer adaptive test can be used like any other test score one obtains from a validated measure that is administered in a non-adaptive format. The common thread is that the test be validated for the particular job/position and that it produces scores in a reliable manner. Once those two conditions are met, regardless of the administration style, the test scores can be used in hiring decisions.

There are multiple strategies for using test scores in making decisions about whether a given applicant should be hired or advance to the next stage of a selection process. One strategy is to set an appropriate passing score on specific assessments or a combination of assessments. Of course the ‘appropriate passing score’ one chooses to use should be calculated carefully and the process used to determine this score should be well documented. Also, one can use a “top-down” selection approach by creating a rank-ordered list of applicant assessment scores and choosing a certain number or percentage of applicants ranked at the top. Test scores can also be used in the hiring decision process by implementing “hurdles” in the selection process, whereby obtaining a passing test score may represent one of these hurdles allowing or not allowing the applicant to move on in the selection process.

As with any other type of test score used in selection, a computer adaptive test score should be used as one piece of information gathered through a selection process that is consistent and well documented across an organization. Ultimately organizations are free to set their own hiring policy, but they should first analyze how test results should be used to make hiring decisions. Additionally, policy decisions should be reviewed and approved by an organization’s human resources and legal counsel.
Legal Defensibility of Computer Adaptive Testing

One important business goal for any organization should be to use fair and legally compliant hiring and promotion practices. Any decision made as part of a hiring or promotion process – including reviews of résumés and job interviews in addition to more formal assessments – may come under scrutiny as part of an internal grievance process, an Equal Employment Opportunity Commission (EEOC) complaint, or an OFCCP audit. A typical example of these challenges include claims that a specific selection procedure is unfair because it is not related to a person’s ability to perform well in a particular job or because candidates believe the process screens out a disproportionate number of people in a protected group.

When implemented appropriately, assessments add an increasingly objective component to the hiring process – a benefit to organizations who are concerned about potential litigious complaints about their process. While some organizations may be reluctant to use any formal selection assessments due to (often unfounded) legal concerns, the use of a relatively subjective hiring process can actually increase the risk of legal exposure. Inserting objectivity in the hiring process is a key factor in enhancing candidates’ feelings about the organization - that they are being treated fairly during the hiring process. In turn, this may decrease the likelihood that candidates file complaints regarding the process. Further, effective use of assessments can help organizations by introducing increasingly objective methods to compare candidates on competencies that are important for job performance rather than relying on initial impressions or other factors that may not be related to success in the job.

SHLPreVisor’s assessments and services, including job analysis and assessment validation, provide clients with specific data and documentation to support the job relatedness of their hiring process. Ultimately, clear documentation of data gathered and analyses conducted that conform with the Uniform Guidelines on Employee Selection Procedures (1978) and Principles for the Validation and Use of Personnel Selection Procedure (2003) will help clients defend against any challenges to their selection procedures.

While computer adaptive tests represent a more complex methodology in terms of the administration and scoring procedures, the use of these tests falls under the same requirements as any other test, interview, or other criteria used to make hiring decisions. Two major requirements include reliability and validity. Reliability is the extent to which a test accurately measures a given individual skill or attribute. Validity refers to the job-relatedness of the test. In other words, does the test measure an individual skill or attribute necessary for successful job performance. Two well-accepted methods for establishing validity include content and criterion-related validity. These validation strategies, as well as CAT reliability, is discussed in more detail in the following paragraphs.

Content Validity

Many tests used in employee selection rely heavily on content validity evidence to justify their use as selection tools. A content-related validity strategy focuses on demonstrating that the content of the computer-adaptive test is relevant to the work requirements on the target job. For example, a word processing skills CAT can be validated for an administrative assistant job by showing that the operation of word processing software, as measured by the CAT, is an important work requirement of the job. The focus of this validation strategy is on demonstrating the correspondence between the tasks or competencies required by the CAT and the tasks or competencies performed on the job.

SHLPreVisor typically establishes the link between assessment content and job requirements based on carefully documented judgments made by subject matter experts (e.g., incumbent employees and immediate supervisors), both in the original design of the assessment and through evaluation by job experts when examining the relevance of an assessment for a particular job. Another important aspect of content validity evidence is the delineation of all areas of the content knowledge domain demanded by the job and the direct mapping of these aspects of the content domain to test items. This aspect also drives the content-balancing algorithm which, in turn, enhances the content validation process.
Criterion-Related Validity

Another indicator of validity is the degree to which CAT scores are related to an important outcome for the job or organization – typically some measure of employee performance. A criterion-related validity strategy investigates whether there is a significant statistical relationship between CAT scores and “criterion” measures such as job performance, training performance, and/or job tenure. If a significant relationship is observed, candidates with more favorable scores will be expected to demonstrate better performance on the job. Feasibility of this validation approach depends on a number of factors (e.g. sufficient numbers of job incumbents or candidates who can participate, availability of appropriate performance measures, etc.).

Reliability of CATs

The reliability of a test refers to the extent to which it is free from measurement error. In other words, differences in test scores among individuals are due to differences in the skill or trait being measured rather than a function of the test. Methods for establishing reliability for “static” (non-adaptive) tests include internal consistency (e.g., coefficient α, KR-20), alternate forms, test-retest reliability, and standard error of measurement (SEM). The first method (internal consistency) is not possible with CATs due to their adaptive nature. The other methods are appropriate, but discussion of SEM is perhaps the most compelling to highlight the greater reliability of CATs over traditional “static” tests.

For static tests, the SEM is typically denoted as a single value for any given test. That is, it represents an estimate of the reliability of the test across all test takers in the sample used to calculate the SEM. However, static tests tend to be more reliable for assessing test takers who are of average ability, and less reliable for those test takers who are of high or low ability. With CATs, the reliability of the test is used as a stopping rule in the sense that the test administration engine can be programmed to end the test once a desired score reliability estimate is reached.

Use of reliability as a stopping rule introduces the notion of “controlled reliability” and greatly improves the accuracy of CATs over traditional/static tests. This control results in more reliable test scores over a broader range of ability levels, and is especially useful in hiring situations where finer distinctions among top performers are critical to making the right hiring decision.

6. SUMMARY

The use of CAT technology provides a more accurate and efficient method for testing individuals in an employment context. Based on its widespread success in other fields, CAT will become the preferred method of test administration due to the many benefits it affords. When properly implemented, computer adaptive tests can be a valuable (and legally defensible) tool for hiring and developing the best employees.
7. REFERENCES


### Computer Adaptive Testing (CAT) in an Employment Context

<table>
<thead>
<tr>
<th>CAT Program</th>
<th>For More Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive Matrices Test (AMT)</td>
<td><a href="http://www.schuhfried.at/eng/wts/amt.htm">www.schuhfried.at/eng/wts/amt.htm</a></td>
</tr>
<tr>
<td>American Institute of Certified Public Accountants Exam (AICPA)</td>
<td><a href="http://www.cpa-exam.org">www.cpa-exam.org</a></td>
</tr>
<tr>
<td>American Society of Clinical Pathologists Exam (ASCP)</td>
<td><a href="http://www.ascp.org/bor">www.ascp.org/bor</a></td>
</tr>
<tr>
<td>Armed Services Vocational Aptitude Battery (ASVAB)</td>
<td><a href="http://www.military.com/ASVAB">www.military.com/ASVAB</a></td>
</tr>
<tr>
<td>Business Language Testing Service Computer Tests (BULATS)</td>
<td><a href="http://www.bulats.org/about/about_computer_test.html">www.bulats.org/about/about_computer_test.html</a></td>
</tr>
<tr>
<td>CAT of Written English for Spanish Speakers</td>
<td></td>
</tr>
<tr>
<td>CAT-Pharmacology</td>
<td><a href="http://www.rasch.org/rmt/rmt64cat.htm">www.rasch.org/rmt/rmt64cat.htm</a></td>
</tr>
<tr>
<td>Cito Company Educational CATs</td>
<td>cito.byuh.edu</td>
</tr>
<tr>
<td>College Level Academic Skills Test (CLAST)</td>
<td><a href="http://www.firm.edu/doe/sas/clast/clstcat.htm">www.firm.edu/doe/sas/clast/clstcat.htm</a></td>
</tr>
<tr>
<td>Computer Adaptive Screening Test (CAST)</td>
<td></td>
</tr>
<tr>
<td>Computer Adaptive Test of English (CATE)</td>
<td><a href="http://www.lrs.mdx.ac.uk/lang/ells/cate.htm">www.lrs.mdx.ac.uk/lang/ells/cate.htm</a></td>
</tr>
<tr>
<td>Computer-Adaptive Test for Reading Chinese (CATRC)</td>
<td>eall.hawaii.edu/yao/catrc/default.htm</td>
</tr>
<tr>
<td>Graduate Management Admission Test (GMAT-CAT)</td>
<td><a href="http://www.west.net/~stewart/gmat/gmat101.htm">www.west.net/~stewart/gmat/gmat101.htm</a></td>
</tr>
<tr>
<td>Learning Potential CAT (LPCAT)</td>
<td><a href="http://www.lpcat.co.za">www.lpcat.co.za</a></td>
</tr>
<tr>
<td>Measures of Academic Progress (MAP)</td>
<td><a href="http://www.nwea.org/Products/MAPHTM">www.nwea.org/Products/MAPHTM</a></td>
</tr>
<tr>
<td>Microsoft Certified Professional Exams</td>
<td><a href="http://www.microsoft.com/traincert/mcepexams">www.microsoft.com/traincert/mcepexams</a></td>
</tr>
<tr>
<td>Middle Years Information System Tests (MidYIS)</td>
<td><a href="http://www.cemcentre.org">www.cemcentre.org</a></td>
</tr>
<tr>
<td>National Council Licensure Exam (NCLEX)</td>
<td><a href="http://www.ncsbn.org/testing/psychometrics.asp">www.ncsbn.org/testing/psychometrics.asp</a></td>
</tr>
<tr>
<td>National Registry of Emergency Medical Technicians Exam (NREMT)</td>
<td><a href="http://www.nremt.org">www.nremt.org</a></td>
</tr>
<tr>
<td>Navy Computer Adaptive Personality Scale (NCAPS)</td>
<td><a href="http://www.sm.nps.navy.mil/">www.sm.nps.navy.mil/</a></td>
</tr>
<tr>
<td>North American Pharmacist Licensure Exam (NAPLEX)</td>
<td><a href="http://www.nabp.net">www.nabp.net</a></td>
</tr>
<tr>
<td>Performance Series - Scantron Corporation</td>
<td><a href="http://www.edperformance.com">www.edperformance.com</a></td>
</tr>
<tr>
<td>Scholastic Reading Inventory (SRI)</td>
<td>teacher.scholastic.com</td>
</tr>
<tr>
<td>Standardized Test for Assessment of Reading (STAR)</td>
<td><a href="http://www.fwps.org/cur/assess/star/index.html">www.fwps.org/cur/assess/star/index.html</a></td>
</tr>
</tbody>
</table>