



White Paper white paper

PREVISOR®

1805 Old Alabama Road, Suite 150, Roswell, GA 30076
800-367-2509 • www.previsor.com

PreVisor ConVerge: The Best Practice for Unproctored/Unsupervised Internet Testing

Michael Fetzer, PhD

Global Director of Advanced Assessment Technologies

Darrin Grelle, PhD

Senior Research Scientist

January 1, 2010

Introduction – Unproctored/Unsupervised Internet Testing

Unproctored/unsupervised Internet testing, or the administration of Internet-based tests to individuals outside a traditional proctored test setting (like a testing center), is an increasingly popular practice in the area of personnel selection. An unproctored Internet test (UIT) can be completed by applicants in their homes or in libraries – literally anywhere they can access the Internet. Recent evidence indicates that more than two thirds of employers who utilize testing for hiring purposes engage in some form of UIT (Fallaw, Solomonson, & McClelland, 2009), which represents a significant increase from 31% in 2005 (Fallaw, Muñoz, & Dawson, 2005).

UIT provides a number of advantages in the selection process. Specifically, UIT can result in decreased costs and improved efficiency while increasing the size of the applicant pool. HR leaders, line managers and executives consistently note that the costs associated with components of traditional proctored testing programs, including test proctors, office space for testing, and computers and IT maintenance make this mode of testing prohibitively expensive compared to UIT. In fact, our research indicates that the costs associated with proctored testing programs are most salient for high volume positions (Lahti & DeKoekkoek, 2006).

In addition to substantial reduction in costs associated with UIT, Internet recruitment is becoming one of the standard methods by which companies must vie for applicants. The Internet has revolutionized how employers source and process applicants. Rather than dealing with a manageable number of applicants that HR personnel can sort through, Internet-based recruitment may mean that literally hundreds or thousands of applications are now received for relatively few job openings. UIT facilitates the management of applicants throughout the application process. Rather than serving as a decision making step relatively late in the hiring process, the advantages associated with UIT mean that it can be used as one of the first screening tools to effectively deal with the influx of applicants (Shepherd, Drasgow, & Beaty, 2004). Validated testing at the recruitment stage has the potential to pay enormous dividends over notoriously less valid screening procedures (e.g., résumé review).

Another benefit of UIT is the accessibility of the hiring process through the Internet. Since many UITs can be accessed from any location that has a computer with Internet access, applicants can take significant steps in the hiring process. This feature has the potential to allow companies to attract high quality “passive” job seekers. While these people who may be well-qualified for a job but are not actively looking for a new position, job postings or ads may be encountered while on the Internet may lead to passive job seeking. Job seekers can easily respond to a job posting or ad, quickly be qualified by the test, and fast-tracked through the hiring process.

The primary criticism of UIT is its susceptibility to cheating, which can have far-reaching consequences for both applicants and organizations using UIT. For instance, in an unproctored setting, an applicant could enlist the aid of a friend or relative to assist in completing the test, or simply have that person take the test for him/her. This vulnerability is especially true for unproctored cognitive ability or knowledge-based tests, as there is typically only one right answer to each question. This vulnerability also exists for unproctored non-cognitive assessments (e.g., personality, biodata, or situational judgment tests) as applicants can work together to determine what patterns of responses yield a “passing” score. Cheating on a test could easily result in an inappropriate hiring decision, as the test score does not represent the true qualifications of the applicant.

Perhaps more damaging to the long-term viability of UIT is the greater degree of risk associated with maintaining the security of the test items used in unproctored testing environments. In other words, there are more opportunities to electronically “copy” items in an unproctored environment as compared to a proctored test center, if the individual has the motivation and the skills. These items could then be sold to others and/or posted to a website, thereby compromising the security (and likely the validity) of the test.

In order to address both of these potential drawbacks to UIT, PreVisor has developed and implemented a process known as PreVisor ConVerge, a new method that utilizes sophisticated computer adaptive testing technology and innovative confirmatory scoring techniques. In order to better understand the value of ConVerge and how it can help organizations take advantage of the benefits of UIT without assuming the risks, a discussion of computer adaptive testing technology will be presented first, followed by an explanation of how this drives the power of PreVisor ConVerge and why it is the best practice.

Computer Adaptive Testing (CAT) – An Overview

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 over 30 major CAT programs in operation all over the world that evaluate four to six million people each year. Until recently, 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. With PreView, PreVisor’s proprietary computer adaptive testing engine, CAT has been made available for use in hiring and developing employees in organizations of all sizes and industries.

The basic premise of CAT is quite simple – the CAT engine selects test items to administer based on the ability level of the test taker. If the test taker answers an item correct, the CAT engine will select a more difficult item to administer next. If the test taker gets an item wrong, however, the CAT engine will then select an 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 technological requirements involved with CAT scoring and item selection. With increasing use of UIT for pre-employment purposes, CAT maximizes test security and efficiency while simultaneously provides accurate estimates of test taker ability. 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 a class of

statistics known as item response theory (IRT) and item selection algorithms, a CAT often requires fewer items to obtain more precise estimates of examinee ability than static or randomly generated tests.

Some noted benefits of CATs include:

- CATs require large item pools to ensure that sufficient a number of items are available across the range of item difficulty. As such, they can be more securely administered in an unproctored setting than static tests as few items will be presented to multiple test takers and the chances that any two test takers will receive the same items are very small.
- Tests can be administered without the presence of a trained test administrator.
- Test scores are available immediately.
- Test security is increased because there are no hard copy tests to be compromised, and varying item exposure leads to a reduction in examinee discussion about test 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 examinees of average ability.

Basic CAT Concepts

The central concept of CAT is to optimally match test items to the test taker's ability level by tailoring the test based on the difficulty of the items. This allows for gaining the maximum amount of information about the test taker's ability level in a minimal amount of time. 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) the test is unidimensional; that is, the items measure the same general construct or trait.

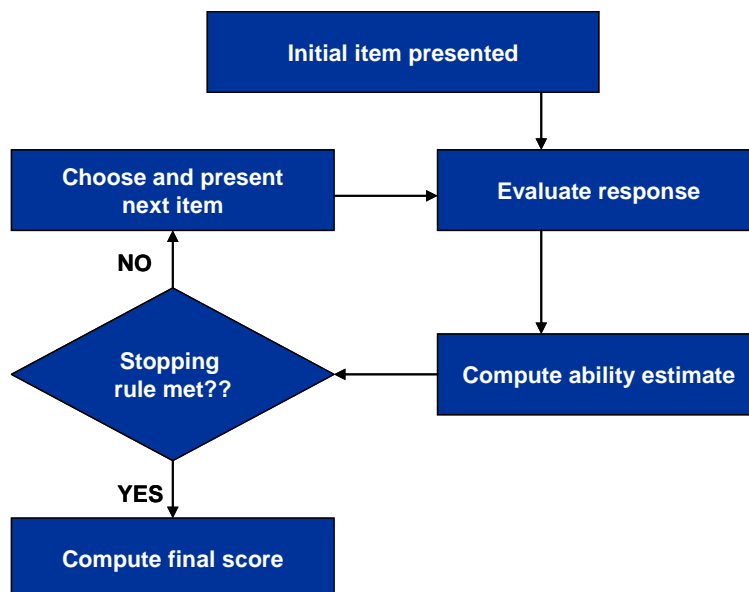
Imagine a test administered to a group of test takers which is far too easy for the group's ability level. Such a test would afford very little discrimination among examinee ability levels as nearly everyone would obtain a perfect score. A higher quality (and more useful) test would allow for maximum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

The same concepts apply for an individual test taker. If a test taker has previously answered several difficult items correctly (and thus is likely to be of high ability), very little is learned about this individual by administering a very easy item. Conversely, little is learned by administering very difficult items to very low ability individuals. However, this is exactly what happens in traditional tests in which a single version is designed and administered to all test takers. Because most test scores must typically discriminate among test takers 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 one individual.

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:

- The test taker's ability is estimated.
- The most appropriate item is chosen based on that ability estimate.
- The test taker responds to item.

The flowchart on the following page serves as an illustration of the item selection, scoring, and ability estimation process utilized by the CAT engine.



This basic cycle is repeated until one of several possible stopping rules is reached. 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 (θ) has been reached.
 - The precision of θ is represented by the standard error, similar to the concept of reliability in static tests.

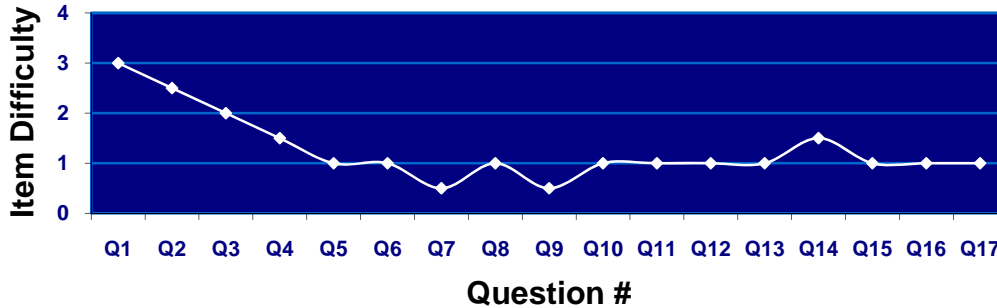
The sophisticated algorithms and rigorous data requirements behind computer adaptive tests are well suited to UIT as CAT helps mitigate many of the challenges associated with UIT. However, using a single, unproctored/unsupervised administration of cognitive ability testing does not remove all potential for cheating. Specifically, there is still the challenge of making sure that the test taker completed the test without help from others. Towards that end, PreVisor has developed a CAT-based method for administering a follow-up proctored/supervised test that leverages the information from the unproctored/unsupervised test administration without raising concerns about false accusations of cheating. This method is known as PreVisor ConVerge and it provides the hiring organization with information about the ability score of the test taker which can be utilized to make accurate and effective hiring decisions.

PreVisor ConVerge – How it Works

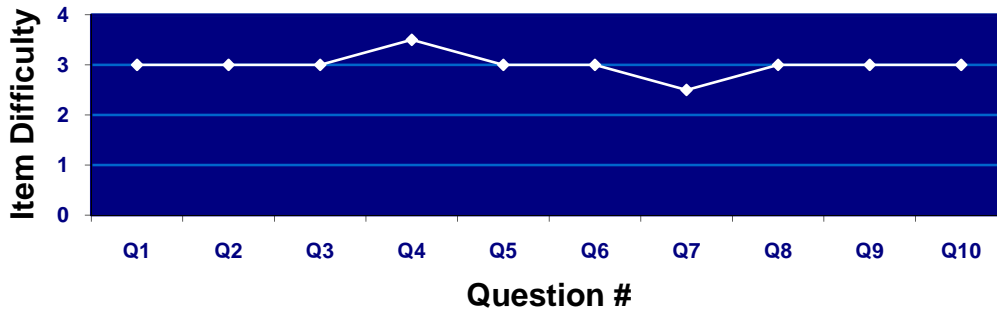
In a CAT, the first item that is administered to the test taker is randomly drawn from a group of items that are of average difficulty, since the majority of test takers are of average ability level. As the test progresses, the CAT engine estimates the ability level of the test taker and produces a final score when the stopping rules have been met. This score can then be used by the CAT engine to determine the first item to be administered to the same test taker when he/she comes on-site for the proctored/supervised follow-up test. This allows for a shorter on-site testing time since the test will start near the test taker's

estimated ability level and thus converge on his/her ability level more quickly, as long as his/her response pattern is consistent with that from the unproctored/unsupervised administration.

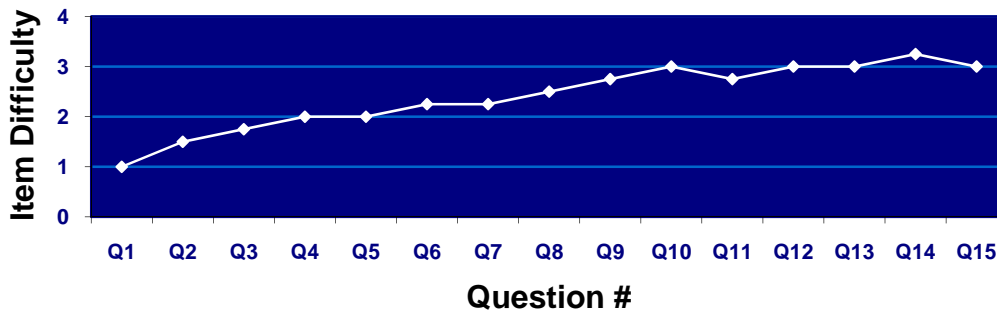
For example, say a test taker of low ability level cheats by having his/her smart friend help out on the unproctored/unsupervised version and ends up with a very high score. When this test taker comes in for the follow-up test and his/her identity can be verified, the CAT engine will choose a very difficult item to administer first. In all likelihood, the test taker will get this item and several subsequent items wrong, as the CAT engine will gradually reduce the difficulty level of the items until the test taker starts to answer items correctly. Since the test taker is of low ability level, s/he will end up with a low score and may not meet the criteria to move forward in the hiring process. This process is illustrated below.



On the other hand, if a test taker of high ability level goes through the process, the time s/he spends on the follow-up test will be much shorter because the CAT engine will converge on the true ability score more quickly. This process is illustrated below.



In a third scenario, a test taker's performance on the unproctored/unsupervised administration may not reflect his/her true ability due to noise or other distractions. PreVisor ConVerge is also able to address this situation by providing the test taker an opportunity to display his/her true ability in a proctored/supervised environment (minus the distractions). This scenario is illustrated below.



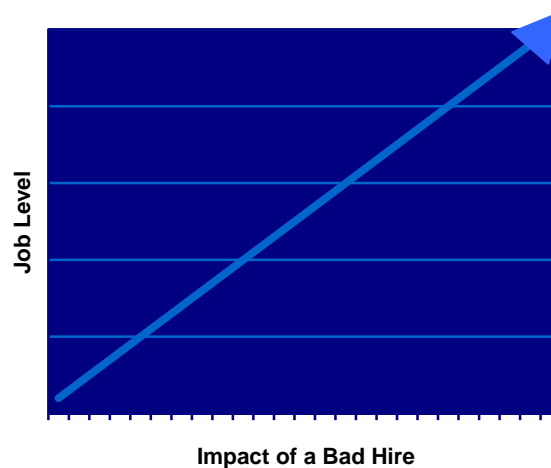
In all three scenarios, the only piece of information used to determine whether or not the test taker moves forward in the process is the score obtained on the follow-up test. There is no need to verify the unproctored/unsupervised score and draw any tentative conclusions (true or false) about the integrity of the test taker. Simply put, cheaters will be screened out and the more qualified (higher ability) test takers will move efficiently through the process.

Why is PreVisor ConVerge the “Best Practice”?

PreVisor ConVerge addresses all the issues traditionally associated with UIT. First, the adaptive nature of the tests greatly increases the security of the item pools since the probability of any two test takers receiving exactly the same set of items is almost zero. Consequently, PreVisor ConVerge greatly reduces the risk of cheaters creating and distributing/selling an answer key that would be useful. Second, computer adaptive tests are much more reliable and accurate for all levels of skill/ability than are static or randomly-generated tests. Third, the ability to utilize the information obtained in the UIT to drive the initial item selection on the follow-up test reduces the second stage testing time. Fourth, there is no need to “verify” the UIT score or administer a third test if a test taker’s score is not verified. Fifth, PreVisor ConVerge also allows for high ability candidates to “put their best foot forward” in the proctored/supervised administration by providing an opportunity to increase their score if their first attempt was not reflective of their true ability. Finally, by eliminating the need for a verification step, the challenges with potentially unfounded perceptions and/or assumptions of cheating are also eliminated.

Although PreVisor ConVerge is by far the best practice when it comes to UIT, there may be situations where follow-up proctored/supervised testing is not practical. This is especially true of high volume hiring situations where thousands of test takers need to be tested before progressing on to the next stage in the hiring process (e.g., interviews). These situations warrant additional considerations that can be partially addressed by the use of CAT. However, without leveraging the full power of PreVisor ConVerge, the main issue centers around the lack of confidence that the test score truly represents the ability level of the test taker.

In these situations, it is important to consider the impact of making a hiring decision based on inaccurate information. In other words, organizations must understand the consequences of hiring someone who may not possess the level of skill or ability that is represented by his/her UIT score. In addition, what other consequences may be associated with hiring someone who has cheated on their pre-employment test? To put this in perspective, one way to quantify the impact of a “bad hire” is by job level, which can be considered a proxy for job complexity, anticipated applicant flow/volume, and level of risk to the organization. For example, a bad hire for an entry-level fast food worker would have less impact on the organization than a bad hire for the regional manager. This perspective is graphically represented below, where the impact of a bad hire goes up as the job level increases.



Typically, the number of jobs available and the resulting size of the applicant pool are inversely related to the job level. For example, an organization may have hundreds of entry-level positions that need to be filled, but only a handful of managerial positions. The administrative cost of proctored/supervised testing would be far greater for the entry-level applicants than for those seeking the managerial positions. Given the potentially lower impact of making a bad hire for entry-level workers combined with the higher administrative costs, the decision to administer pre-employment tests completely unproctored/unsupervised would be relatively more appropriate for those jobs than for managerial positions. Again, the best practice recommendation for any UIT situation would be to utilize PreVisor ConVerge. However, we hope that the above discussion provides some insight for organizations who are debating the costs and benefits of the proctored/supervised stage of PreVisor ConVerge.

References

- Fallow, S. S., Muñoz, C. S., & Dawson, C. R. (2005, April). *Administering online testing: A benchmarking study*. Poster presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.
- Fallow, S. S., Solomonson, A. L., & McClelland, L. (2009, April). *Current trends in assessment use: A multi-organizational survey*. Poster presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Lahti, K. & DeKoekkoek, P. (2006). ROI for proctored versus unproctored assessment programs: Estimates from multiple utility models and identification of moderators. Presentation at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Shepherd, W.J., Drasgow, F., Beaty, J. (2004). New applications of computerized employment testing: Using human capital measurement for business decisions. *IHRIM Journal*, Sept/Oct, 40-46.