Organizational Research Methods
14(4) 608-630
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1094428110370715
http://orm.sagepub.com
SAGE

Unproctored Internet Test Verification: Using Adaptive Confirmation Testing

Guido Makransky[1,2] and Cees. A. W. Glas[3]

Abstract

Unproctored Internet testing (UIT) is commonly used in employment test administration. When the test is high stakes, the International Guidelines on Computer-Based and Internet-Delivered Testing recommend to follow up the results with a confirmation test in a controlled setting. This article proposes and compares two methods for detecting whether a test taker's original UIT responses are consistent with the responses from a follow-up confirmation test. The first method is a fixed length adaptive confirmation test using the likelihood ratio (LR) test to evaluate cheating and the second method is a variable length adaptive confirmation test using an extension of the stochastic curtailed truncated sequential probability ratio test (SCTSPRT) to evaluate cheating. Simulation studies indicated that the adaptive confirmation test using the SCTSPRT was almost four times shorter while maintaining the same detection power. The study also demonstrated that cheating can have a detrimental effect on the validity of a selection procedure and illustrated that the use of a confirmation test can remedy the negative effect of cheating on validity.

Keywords

unproctored Internet testing, computer adaptive testing, item response theory

Introduction

The increase in computer availability and the widespread use of the Internet have led to an increased application of unproctored tests that can be administered at any place and time via the World Wide Web. Unproctored Internet testing (UIT) is Internet-based testing of a candidate without a traditional human proctor. Although definitive usage data are lacking, it is probable that UIT accounts for the majority of individual employment test administrations that currently take place in the U.S. private sector (Pearlman, 2009). Fallaw, Solomonson, and McClelland (2009) found that more than two thirds of the employers who conduct testing for selection are already engaging in UIT.

[1] Master Management International A/S, Allerød, Denmark
[2] University of Twente, Enschede, The Netherlands
[3] Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioral Sciences, University of Twente, Enschede, The Netherlands

Corresponding Author:
Guido Makransky, Master Management International A/S, Gydevang 39-41, Allerød 3450, Denmark
Email: guidomakransky@gmail.com

The flexibility of Web-based test administration is attractive because it limits the resources necessary for administering tests, meaning that test proctors need not be hired, trained, or sent to testing locations; testing equipment does not have to be purchased, distributed, or maintained; and job candidates do not have to travel to testing locations. UIT also allows continuous access to assessments and limits the time it takes to process candidates.

The primary disadvantage of UIT stems from the many forms of cheating that are possible such as assistance from others who have knowledge of the items before the test, assistance from others during the test, or the substitution of test takers (Tippins, 2009). There is evidence that cheating is widespread in unproctored as well as proctored settings. Automatic Data Processing Inc. (2008) reports data indicating that 45% of job applicants falsify work histories. These applicants would presumably be willing to cheat if given the opportunity. Cizek (1999) and Whitley (1998) found that approximately half of all college students report cheating on an exam at least once during their college education. There is also abundant literature motivated by the concern that applicants may fake their responses in occupational testing in an attempt to obtain employment (e.g., Anderson, Warner, & Spector, 1984; LaHuis & Copeland, 2009; Ones, Viswesvaran, & Reiss, 1996). It is likely that cheating in UIT is particularly widespread because proctoring is the primary means by which cheating is curtailed (Arthur, Glaze, Villado, & Taylor, 2009). In fact, Chapman and Webster (2003) describe cheating as one of the most common reasons that human resources professionals are hesitant to implement UIT.

Some possible solutions for limiting the detrimental effects of cheating in a high stakes unproctored Internet ability test are to use a speeded test (e.g., Arthur et al., 2009; Nye, Do, Drasgow, & Fine, 2008) or to have a two-step testing process where a confirmation test is used to verify the results of the first test (e.g., Beaty, Dawson, Fallaw, & Kantrowitz, 2009; Burke, van Someren, & Tatham, 2006).

Nye et al. (2008) did not detect any differences due to cheating when comparing proctored and unproctored test results for a perceptual speed test in a selection setting. Similarly, Arthur et al. (2009) concluded that a speeded test format appears to reduce the prevalence of cheating. However, more research is needed before a conclusion about the validity of using a speeded test format without a verification test for UIT can be made.

Currently, the International Guidelines on Computer-Based and Internet-Delivered Testing (2005), developed by the International Testing Commission, explicitly stipulate the need for verification testing. Guideline 45.3 states:

> For moderate and high stakes assessment (e.g., job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e., at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores. Procedures should be used to check whether the test-taker's original responses are consistent with the responses from the confirmation test. Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (e.g., not seek assistance, not collude with others, etc.). This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept.

The most commonly used method for conducting a confirmation test is the test–retest approach. With this approach, an unproctored test is used as a prescreen, followed by a proctored administration of the same or a parallel version of the test in a proctored setting. Variations are to use randomized item selection (e.g., Burke et al., 2006) or computerized adaptive testing (CAT).

There are many statistical methods available for detecting forms of cheating such as collusion in proctored settings, including Angoff's β index (Angoff, 1974), error similarity analysis (Belleza &

Belleza, 1989), and the Z index (van der Linden & Sotaridona, 2002). There is also research investigating how to identify respondents who distort their responses in personality scales when they are used for personnel selection (e.g., Drasgow, Levine, & Zickar, 1996; LaHuis & Copeland, 2009). However, studies describing the statistical methods used in UIT confirmation testing or literature comparing the efficiency of different methods are still lacking. This is surprising, given the widespread use of UIT and the well-documented criterion validity evidence of ability tests used for personnel selection (e.g., Schmidt & Hunter, 1998).

The problem with current confirmation test approaches is that they largely undermine the purpose and attractiveness of UIT in the first place, which is to have a quick, seamless, automated, and candidate-friendly application-selection process that will reduce cost by limiting on-site or proctored testing. This is the case because most confirmation tests are unnecessarily long due to suboptimal designs. The lack of short confirmation tests means that many possible UIT benefactors discard or partially discard using UIT (Kaminski & Hemingway, 2009). Therefore, research is needed to investigate whether efficient follow-up testing is possible and to find effective methods of detecting suspicious test responses.

This study is designed to fill the gap in the literature by identifying a psychometric method for developing a confirmation test designed to verify the veracity of unproctored ability test results as efficiently (in terms of the number of items administered) as possible. Thus, a confirmation test, as it is investigated in this article, is used exclusively to confirm or reject the result from the unproctored test and is not intended to be used as a supplement to the unproctored test result.

An effective framework for developing such a confirmation test is to use CAT combined with item response theory (IRT). CAT has the advantage over traditional tests that the test can be tailored to the individual respondent, ensuring a precise result as quickly as possible. The use of CAT has been found to increase the efficiency of classification decisions in classification testing (Eggen & Straetmans, 2000; Rudner, 2002) by decreasing the number of items necessary for making an accurate decision without compromising precision. In classification testing, the goal of the examination is to classify respondents into a limited number of categories based on cutoff points selected on an ability scale. Confirmation testing can be considered a special form of classification testing, where the goal of the test is to classify each respondent by either accepting or rejecting the hypothesis that they have the same ability level in the controlled proctored condition as they did in the unproctored setting. The fundamental difference between confirmation and classification testing is that a unique cutoff point is used for each respondent in a confirmation test, whereas a classification test uses a general cutoff point for the whole sample. This difference is essential and implies that the methods used in classification testing cannot be applied directly to confirmation testing and that extensions of these methods are required.

Consequently, the design of the algorithm for an adaptive confirmation test could take the form of a classification CAT where the result from each respondent's unproctored test is used as a unique classification cutoff point in the confirmation test. This presents a challenge because the unproctored test result is an ability estimate and not a fixed point. In the following sections, we explore and test the consequences of using adaptive confirmation test methods for verifying unproctored test results, and propose a means of generalizing the methods used in classification testing to the present context, with the intention of identifying a way of increasing the efficiency of confirmation tests compared to those currently being used.

The article is organized as follows: The IRT model is presented in the next section. This is followed by an introduction of a statistical test that can be used in a fixed length confirmation test. The subsequent section covers two sequential procedures for conducting a classification test: the truncated sequential probability ratio test (TSPRT) and the stochastically curtailed TSPRT (SCTSPRT) and presents a method for generalizing these procedures to the current context. This is followed by an outline of the research questions in the study, and a description of the simulation studies designed to answer these questions. The final section provides an overview of the results including benefits and limitations of using the proposed methods for confirmation testing, as well as future research possibilities.

The IRT Model

The proposed methods are defined in the framework of IRT. The fundamental concept of IRT is that each test item is characterized by one or more parameters and each test taker is characterized by a single ability parameter. The probability that a given test taker answers a given item correctly is given by a function of the test taker's ability level θ and the item parameters. Conditional on those parameters, the response on one item is assumed independent of the responses to other items. The IRT model used is the 2-Parameter Logistic, or 2-PL model, where the probability of a correct response is given by

$$P_i(\theta) = \frac{1}{1 + \exp(-a_i(\theta - b_i))} \tag{1}$$

(Birnbaum, 1968). Here, $P_i(\theta)$ is the probability of a correct response for item i, θ is the test taker's ability, and a_i and b_i are item parameters. Furthermore, a_i is called the discrimination and b_i the difficulty parameter. It is common to scale ability levels to have a mean of 0 and a standard deviation of 1. The item difficulty parameter represents the point on the ability scale at which the probability of answering the item correctly equals .5. Higher discrimination parameters indicate that the probability of answering an item correctly rises dramatically with small changes in ability in the region of the item difficulty.

Alternative IRT models that might be considered are the 1- and 3-Parameter Logistic models (1-PL and 3-PL models). In the first model, also known as the Rasch model (Rasch, 1960), the a parameters of all items are considered equal. This model was not pursued here, because it is often too restrictive to obtain acceptable model fit. In the 3-PL model (Birnbaum, 1968), a guessing parameter is added as a third item parameter to model guessing. This model was not pursued in this article because many authors report identification problems with the model (see, for instance, Luecht, 2006 or Partchev, 2009). Luecht (2006, p. 578) remarks that attempting to estimate stable item parameters in small samples can be extremely challenging and that the estimation problems include nonconvergence of the numerical estimation solutions from ill-conditioned likelihood functions and empirical underidentification of the model parameters leading to large error covariances of the parameter estimates. In addition, we focus on adaptive testing where the items are tailored to the ability level of the respondents, θ is usually close to b_i, and the probability of a correct response converges to .5. Therefore, there is little information in the data to support the estimation of a guessing parameter on one hand and, on the other hand, a guessing parameter adds little to the precision of the estimate of θ.

What differentiates CAT from a classical linear test is that an item selection function is used to optimally select the next item in the test. A selection function that is often used is Fisher's information function. For dichotomously scored items, the information function has the following form:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{2}$$

where $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ, and $Q_i(\theta) = 1 - P_i(\theta)$. The next item is selected that has the maximum information in the item pool at $\theta = \theta^*$, where θ^* is the current estimate of the ability of the test taker. CATs can also be used to classify examinees into a limited number of categories, by evaluating a hypothesis that θ is above or below a cutoff point on the ability scale. These algorithms usually select the next item with maximal information at the cutoff point, rather than at the current ability estimate.

The efficiency (in terms of the number of items administered) of CAT in classification testing can be improved by adjusting the decision method and the termination criteria, as well as the item selection method that is used. Different item selection methods are discussed later in this article. In the sequential procedures used in classification testing, the response pattern is analyzed after the administration of each item to determine whether there is sufficient information to make a classification decision without administering unnecessary extra items. However, the test length requirements for a confirmation test are more stringent than those for a classification test. Therefore, it is not certain that a sequential procedure will increase the efficiency of the confirmation test without severely decreasing the decision accuracy. The following sections describe both fixed length and sequential procedures for developing a confirmation test used for validating UIT.

Fixed Length Confirmation Test

A hypothetical job candidate will be used to illustrate the confirmation testing procedures used in this article; we will call this candidate C. Upon applying for a job, C is typically administered an initial unproctored online test that he can take at his convenience. If his result meets the criteria for selection, he is then offered a follow-up confirmation test under supervised conditions. A fixed length confirmation test consists of a fixed number of items. After the administration of all of the items in the test, a statistical test is used to evaluate the hypothesis that C's ability level is significantly lower than his unproctored test score. If C's confirmation test score is not significantly lower than his unproctored test score, then his unproctored test result is accepted and this score is used in making a selection decision. Note that the score from the confirmation test is not used to supplement C's unproctored score; it is simply used to confirm or reject the unproctored score. The example of C will be used again later in this article. We will now explain the fixed length confirmation test in more technical detail.

The LR test can be used to identify aberrant response patterns in a fixed length test. The LR test uses the ratio of the maximum likelihood under two different hypotheses. In the current context, the null hypothesis is that one ability parameter is sufficient to describe the response pattern in the two tests. This hypothesis represents the assumption that the respondent has the same level of ability in both tests. The alternative hypothesis is that two ability parameters are necessary to describe the response pattern: one for the unproctored test and one for the confirmation test. This represents the situation where the respondent's score has changed from the unproctored to the confirmation test.

Investigating the log-likelihood of each model compares the two models under the two hypotheses. The log-likelihood function of the ability parameter θ, given a response pattern $\boldsymbol{x}$ on a test of N items is given by

$$\operatorname{Log} L(\theta;x) = \log\left[\prod_{i=1}^{N} P_i(\theta)^{x_i}(1 - P_i(\theta))^{1-x_1}\right], \tag{3}$$

where $P_i(\theta)$ is the item response function from the 2-PL model given in equation 1, $x_i = 1$ for a correct response and $x_i = 0$ for an incorrect response to item i. The LR statistic for the test of the null hypothesis that the unproctored test is valid is given by

$$\mathrm{LR} = \frac{L(\theta;x_1,x_2)}{L(\theta_1;x_1)\,L(\theta_2;x_2)}, \tag{4}$$

where $\boldsymbol{x_1}$ and θ_1 are the response pattern and estimated ability from the unproctored test and $\boldsymbol{x_2}$ and θ_2 are the response pattern and estimated ability from the confirmation test. $L(\theta_1; \boldsymbol{x_1})$ and $L(\theta_2; \boldsymbol{x_2})$ are the likelihood of the unproctored and confirmation test, respectively, and $L(\theta; \boldsymbol{x_1}, \boldsymbol{x_2})$ is the likelihood of one single θ computed using the responses of the unproctored and confirmation test concurrently. The LR statistic has a chi-square distribution with degrees of freedom equal to the number of parameters in the unrestricted model minus the number of parameters in the restricted model. This difference is one in the current context.

In confirmation testing, the one-sided test is used because the only matter of interest is if the estimate for the unproctored test is significantly higher than the estimate for the confirmation test, because this would suggest that there is a possibility that the respondent has cheated.

The one-sided version of the null hypothesis of the LR test is

$$H_0 : \theta_1 \leq \theta_2. \tag{5}$$

The decision to reject (Decision = r) the null hypothesis is made if the LR test is significant and the unproctored test score (θ_1) is higher than the confirmation test score (θ_2). Otherwise, the null hypothesis that the result from the unproctored test is valid, is accepted (Decision = a).

Sequential Confirmation Tests

An organization that is interested in hiring C (the hypothetical job candidate introduced earlier) may want to reduce the resources necessary for test administration by reducing assessment time. In this situation, a sequential confirmation test could be used. In a sequential confirmation, test items are administered to C one at a time. After his response to each item, the decision algorithm evaluates if there is enough information available to accept or reject the null hypothesis that C's unproctored test result is valid. If a decision can be made to accept or reject C's unproctored test score at this point, then the test is terminated. If there is not enough information available to make a decision, then an additional item is administered. Additional items are administered until a decision is made or until the maximum number of items is reached, in which case a statistical test similar to the fixed length confirmation test is used to determine the decision. Two methods for conducting a sequential confirmation test are explored in this article: the TSPRT and the SCTSPRT.

The TSPRT is a method that is commonly used for classifying respondents in computerized classification testing (Eggen, 1999; Eggen & Straetmans, 2000; Reckase, 1983). The TSPRT offers substantially shorter tests than a conventional full-length fixed-form test, while maintaining a similar level of classification accuracy (Eggen & Straetmans, 2000; Rudner, 2002). The SCTSPRT is an extension of the TSPRT that can be used to shorten testing even further, without substantially compromising error rates (Finkelman, 2003, 2008).

The TSPRT and the SCTSPRT cannot be applied directly to the context of confirmation testing, because the hypothesis tested in this context is based on the change between two testing sessions, rather than the comparison of an estimate to one or more fixed cutoff points, which is the basis for the hypotheses tested in classification testing. Therefore, confirmation testing is fundamentally different from other applications where the TSPRT and the SCTSPRT are used. The following sections outline the TSPRT and the SCTSPRT as they are used in computerized classification testing. This is followed by an extension of these tests to confirmation testing.

TSPRT

The sequential probability ratio test (SPRT) is a sequential procedure used to make classification decisions in computerized classification testing. In applied conditions, a maximum for the number of items to be administered is usually set (e.g., Eggen & Straetmans, 2000; Jiao, Wang, & Lau, 2004). This is a truncation of the SPRT labeled TSPRT.

When implementing the TSPRT, first a cutoff point (θ_o) on the θ scale is selected, followed by a small region on each side of the point. The combination of the two regions of size $\delta > 0$ is referred to as the indifference zone. The indifference zone serves the purpose of identifying two points on the ability scale that can subsequently be used to calculate the LR. The statistical hypotheses are formulated as

$$\mathrm{H}_0 : \theta_2 \geq \theta_o + \delta = \theta_+ \,\text{against}\, \mathrm{H}_\mathrm{A} : \theta_2 \leq \theta_o - \delta = \theta_-. \tag{6}$$

After the administration of the kth item ($1 \leq k < N$, where N is the maximum number of items in the test), the following statistic is used to make a decision to accept or reject H_0:

$$\mathrm{LR}_k = L_k(\theta_+; x)/L_k(\theta_-; x) = \log L_k(\theta_+; x) - \log L_k(\theta_-; x), \tag{7}$$

where $L_k(\theta_+;\boldsymbol{x})$ and $L_k(\theta_-;\boldsymbol{x})$ are the likelihoods, θ_+ is the model under the null hypothesis and θ_- is the model under alternative hypothesis, both given the response pattern observed up to the kth item. The log-likelihood ratio of the values θ_+ and θ_- after k observations represents the relative strength of evidence for selecting θ_+ and θ_- as the correct decision. High values of the ratio indicate that the examinee's ability is above the cutoff point and low values indicate that their ability is below the cutoff point. Note that no estimation of θ is involved. Decision error rates are specified as

$$P(\text{accept } \mathrm{H}_0|\mathrm{H}_0 \text{ is true}) \geq 1 - \alpha, \text{and } P(\text{accept } \mathrm{H}_\mathrm{A}|\mathrm{H}_\mathrm{A} \text{ is true}) \leq \beta, \tag{8}$$

where α and β are small constants that can range from .01 to .4. Furthermore, we define $\mathrm{A} = \alpha/(1 - \beta)$, and $\mathrm{B} = (1 - \alpha)/\beta$. The TSPRT uses the following procedure after each item has been administered:

$$\text{Continue sampling if} : \log \mathrm{A} < \mathrm{LR}_k < \log\mathrm{B}, \text{ and } k < N, \tag{9}$$

$$\text{AcceptH}_0 \,|\, \text{Decision} = \text{a, if} : \mathrm{LR}_k \geq \log \mathrm{B}, \tag{10}$$

$$\text{RejectH}_0 \,|\, \text{Decision} = \text{r, if} : \mathrm{LR}_k \leq \log \mathrm{A}. \tag{11}$$

If the maximum number of items in the test is reached (that is, if $k = N$), an additional decision rule is implemented:

$$\text{Accept H}_0 \,|\, \text{Decision} = \text{a, if} : \mathrm{LR}_k \geq \log \ \mathrm{C}, \tag{12}$$

$$\text{RejectH}_0 \,|\, \text{Decision} = \text{r, if} : \mathrm{LR}_k < \log \ \mathrm{C}. \tag{13}$$

Here, C is a constant satisfying $\mathrm{A} \leq \mathrm{C} \leq \mathrm{B}$; its selection is an important factor in the error rates of the TSPRT. Spray and Reckase (1996) proposed using the relation $\log \mathrm{C} = (\log \mathrm{A} + \log \mathrm{B})/2$, which would result in $\log \mathrm{C} = 0$, whenever $\mathrm{A} = 1/\mathrm{B}$.

The SCTSPRT

An alternative to the TSPRT used to shorten testing without substantially compromising error rates is the SCTSPRT (Finkelman, 2003, 2008). Often, the TSPRT is inefficient because an extra item is presented even though the classification decision for the respondent can no longer be changed. The SCTSPRT stops testing once the decision cannot be altered on the basis of the remaining items in the test. This is called curtailment. The SCTSPRT also stops testing in cases in which the probability of changing the classification decision is smaller than a predefined value (y).

To evaluate the probability that the decision will change, it is first necessary to define what the current decision is, based on the k responses so far. D_k* represents the tentative decision at stage k; where $k < N$. Finkelman (2008) suggests using the following definition: $\mathrm{D}_k^* = \mathrm{a}$, if $\mathrm{LR}_k \geq \log \mathrm{C}$ and $\mathrm{D}_k^* = \mathrm{r}$ otherwise. The probability of changing decisions depends on the probabilities of the response pattern of future items starting at stage $k + 1$, which in turn depends on θ. Therefore, a value for θ must be selected for evaluating the probability that the decision will change between stage k and N, that is, evaluating whether $D_k{}^* \neq D_N$, where D_N is the decision at stage $k = N$. A conservative option is to use the conditional power approach by Jennison and Turbull (2000). The conditional power approach implies setting $\theta = \theta_-$ when $D_k{}^* = \mathrm{a}$, and $\theta = \theta_+$ when $D_k{}^* = \mathrm{r}$ (Finkelman, 2008).

The stochastic curtailed TSPRT is defined as the following sequential rule.

If $k < N$, set $k = N$ and D = a if

$$\{\log \mathrm{LR}_k \geq \log \mathrm{B}\} \text{ or } \{\log\ \mathrm{LR}_k > \log \mathrm{C} \text{ and } P_{\theta-}(\mathrm{D}_N = \mathrm{a}|\mathrm{LR}_k) \geq y\}; \tag{14}$$

and set $k = N$ and D = r if

$$\{\log\ \mathrm{LR}_k \leq \log \mathrm{A}\} \text{ or } \{\log\ \mathrm{LR}_k < \log \mathrm{C} \text{ and } P_{\theta+}(\mathrm{D}_N = \mathrm{r}|\mathrm{LR}_k) \geq y'\}, \tag{15}$$

$$\text{If } k = N, \text{set D} = \text{a if and only if } \log \mathrm{LR}_k \geq \mathrm{C}, \tag{16}$$

where y and y' are thresholds along the probability scale, which take values between .5 and 1. In particular, they determine how high the probability of retaining the current decision must be for the test to be stopped early. That is, the statement $P_{\theta-}(\mathrm{D}_N = \mathrm{a}|\mathrm{LR}_k) \geq y$ in equation 14 specifies that the probability of event $\mathrm{D_T} = \mathrm{a}$, given log LR_k, must be greater than or equal to y under θ_- for the criterion to be satisfied. Similarly, the statement $P_{\theta+}(\mathrm{D}_N = \mathrm{r}|\mathrm{LR}_k) \geq y'$ in equation 15 specifies that the probability of the event $\mathrm{D_T} = \mathrm{r}$, given log LR_k, must be greater than or equal to y' under θ_+ for the criterion to be satisfied.

Setting $y = y' = 1$ results in a test that has the same error rates as the TSPRT with the possibility of reducing the number of items. The stochastic curtailed SPRT is straightforward when all items that will be administered are known and k is close to N. Finkelman (2008) uses an approximation of the probabilities $P_{\theta-}(\mathrm{D_T} = \mathrm{a}|\mathrm{LR}_k)$, and $P_{\theta+}(\mathrm{D}_N = \mathrm{r}|\mathrm{LR}_k)$ in equations 14 and 15 when the remaining items in the test are unknown. Applying Finkelman's approximation in the current context is not straightforward, so we approximate these probabilities by simulation as a computational method. A simulation of the remaining item responses in the test at point k can be applied to calculate the probability of retaining the current decision at point $k = N$. The simulation uses the conditional power approach described above for selecting θ at point k. The remaining item responses from point k to $k = N$ can then be generated by replicating a number of CATs given θ_+ or θ_- 100 times. The probability of retaining the decision is computed by adding up the number of times the decision remains unchanged in the 100 replications.

Extending the SPRT Framework to Computerized Confirmation Testing

The SPRT framework cannot be applied directly to confirmation testing because the cutoff point is not a fixed point but varies over respondents based on their unproctored test result. Therefore, an extension of the SPRT framework is needed where the cutoff point is defined at the level of the individual based on the unproctored test result. Defining the cutoff point based on the unproctored test result presents a further challenge because the unproctored test result is an estimate, which contains error, and is not a fixed variable. A solution is to develop a test taking into account the sampling distribution of the estimate. Using a cutoff point from the sampling distribution, the probability of exceedance or significance probability can be determined.

When using the SPRT framework for a confirmation test, the selection of the cutoff point determines the ratio of Type I and Type II error rates in the test. Decreasing the Type I error rate automatically means an increase in Type II error; therefore, a prioritization of the two is important. Incorrectly rejecting a respondent who has completed the unproctored test without cheating (a Type I error) can have serious ethical and possible legal consequences. Burke et al. (2006) provide a good overview of considerations that are important in handling such respondents, which can be demanding in terms of required resources for those administering the test. Alternatively, the cost of committing a Type II error includes hiring the incorrect applicant, which can result in negative organizational consequences and negative consequences for the applicants who were not selected. In the current context, raising questions about someone's integrity usually outweighs the negative consequences of a Type II error. Therefore, the development of a confirmation test method that controls for the Type I error rate is prioritized in this article.

The cutoff point for the confirmation test can be established using the lower limit of the confidence interval for the ability estimate in the unproctored test. This confidence level would hold for the Type I error rate in the confirmation test, if the confirmation test had perfect precision. In practice, a more conservative value is needed because the error is compounded due to a combination of the error in the two tests. One method for establishing the Type I error rate for the confirmation testing procedure is to take the error of both tests into account with the following equation:

$$\theta_0 = \theta_1 - (SE(\theta_1) + g), \tag{17}$$

where θ_1 and $SE(\theta_1)$ are the ability estimate and standard error obtained from each respondent's unproctored test, respectively, and g is a positive constant representing a combination of the anticipated error in the confirmation test and the desired Type I error rate of the test.

The SPRT requires a predefined cutoff point before the confirmation test begins; therefore, the exact value of g that corresponds to a specific Type I error rate cannot be obtained because it will depend on the items that are administered in the confirmation test. However, a value of g can be chosen in such a way that the Type I error rate aggregated over respondents is close to a predefined overall nominal Type I error rate. The choice of g only takes the characteristics of the confirmation test into consideration; the error in the unproctored test is already included in equation 17. Therefore, one value of g would result in a stable Type I error rate, regardless of the precision of the unproctored test. The value of g that corresponds to a Type I error rate can be obtained by conducting an a priori simulation of the confirmation test.

In summary, the purpose of the proposed method is to set a cutoff point at some position below the θ estimate in the unproctored test, to take into account the error in the unproctored test and anticipated error in the confirmation test with the intention of controlling the Type I error rate in the confirmation test.

Research Questions

There are four main research questions in this study: Which item selection method performs best in fixed length and sequential confirmation testing procedures? What are the benefits of using a sequential confirmation test, rather than using a fixed length confirmation test based on efficiency in terms of the number of items administered and the power to identify cheaters? What are the benefits of using the stochastic curtailed version of the TSPRT compared to using the TSPRT? Does a confirmation test decrease the effect of cheating on the validity of a selection procedure?

Simulation Studies

The research questions outlined above were investigated by means of simulation studies. The first simulation study was designed to compare different methods for conducting a fixed length confirmation test. The second study was used to define the cutoff points that correspond to nominal Type I error rates for a sequential confirmation test and to test whether the empirical Type I error rates matched the nominal Type I error rates. The third study explored different options for conducting a TSPRT for sequential adaptive computerized confirmation testing. This study also compared the results for the fixed length adaptive confirmation test with those of the TSPRT to investigate if the TSPRT improved the efficiency of the fixed length adaptive confirmation test. The fourth study investigated the possibility of shortening the confirmation test even further using the SCTSPRT procedure. Finally, a selection procedure was simulated to test the utility of the computerized adaptive confirmation test for a personnel selection procedure.

The simulation studies were programmed in Digital Visual FORTAN 6.0 for Windows. The ability estimation methods used in the fixed length confirmation test can also be calculated using standard IRT software packages such as Parscale (Muraki & Bock, 1996) or ConQuest (Wu, Adams, Wilson, & Haldane, 2007). It was assumed that during the unproctored test, a respondent who had cheated had another θ value than during the confirmation test. This was simulated by drawing θ from a standard normal distribution for the confirmation test. The θ value in the unproctored test was set to $\theta = \theta + \Delta$, where Δ represented the effect of cheating. Three groups of respondents were simulated. The first was non-cheaters: Here, the effect of cheating was 0 ($\Delta = 0$). The next group represented a moderate cheating effect of one standard deviation unit ($\Delta = 1$). This means that all of the respondents in this group had a true ability level that was one standard deviation unit higher in the unproctored test compared to the confirmation test. The final group represented a large cheating effect of two standard deviation units ($\Delta = 2$). In this group, all of the respondents had a true ability level that was two standard deviation units higher in the unproctored test compared to the confirmation test. The values of Δ were selected based on previous literature. Burke et al. (2006) report values of $\Delta = 2$ in testing the efficiency of a confirmation test used for high stakes occupational ability testing. van Krimpen-Stoop and Meijer (2000) use values of $\Delta = 1$ and $\Delta = 2$ in investigating non-fitting response behavior in the second half of a CAT. These values are large; however, in this context, they are appropriate because the consequences of rejecting an unproctored test score are serious and would only be considered when the effect of cheating is large. In addition, the type of cheating expected in an unproctored test, such as the substitution of test takers, would likely lead to greater effect size differences than the cheating that is widespread in proctored settings.

An item bank was simulated by drawing item difficulty parameters from a standard normal distribution and item discrimination parameters from a lognormal distribution with an expectation of 1. The item responses were generated according to the 2-PL model in equation 1. This was repeated for each respondent, both for the unproctored test and for the confirmation test. The unproctored test had a fixed length where the first 3 items were selected with maximum information for $\theta = 0$ and the remaining items were administered by selecting the item with maximal information at the current θ estimate. The estimation of θ was done with either maximum likelihood estimation (MLE) or weighted likelihood estimation (WLE; Warm, 1989). Each condition was repeated for 10,000 respondents.

Items in the confirmation test were administered according to one of 3 item selection procedures: random (RAN), maximum information at the cutoff point (MIC), and maximum information at the current θ estimate (MIE). The first procedure randomly selects the next item from the available item bank, excluding items used before. This selection method is similar to a nonadaptive confirmation test (e.g., Burke et al., 2006). The other two procedures select the next item maximizing the item information in equation 2. In the MIC item selection algorithm, the next item is selected to maximize the information for the cutoff point. In MIE, the next item selected is the item for which the information at the current ability estimate is maximal. The MIE item selection algorithm was initiated after 3 items had been administered in the confirmation test to ensure that a fairly stable ability estimate had been established. The item with maximum information at the cutoff point was administered up to this point. MIC has been found to obtain comparable power with fewer items than MIE in computerized classification testing (Eggen, 1999; Spray & Reckase, 1994); however, MIE item selection is favored when the purpose of the test is to estimate the ability of an examinee (Thissen & Mislevy, 2000).

Table 1. Percentage of Aberrant Response Classifications for a Fixed Length Confirmation Test Using the LR Test With a Significance Level of 5%

		MLE Length of unproctored test								
		10 items			30 items			50 items		
Test length	Δ	RAN	MIC	MIE	RAN	MIC	MIE	RAN	MIC	MIE
5	0	2.28	0.88	1.09	1.60	0.81	0.50	1.37	0.32	0.70
	1	9.88	1.04	4.57	11.07	8.65	7.47	11.09	4 .14	8.23
	2	19.21	1.32	9.36	33.46	6.08	13.19	35.83	23.71	20.20
10	0	6.67	10.45	9.71	3.28	5.02	5.45	2.93	4.23	4.09
	1	22.55	32.03	29.19	26.96	38.84	37.97	27.07	37.36	40.73
	2	35.72	32.83	36.52	62.05	58.00	60.07	65.22	65.58	67.13
20	0	13.06	16.49	15.72	5.48	7.86	7.75	4.30	5.73	5.42
	1	39.98	56.60	54.02	45.80	64.03	64.76	47.45	65.05	67.30
	2	59.88	65.07	65.55	81.57	88.09	87.63	88.57	92.34	93.44

		WLE Length of unproctored test								
		10 items			30 items			50 items		
Test length	Δ	RAN	MIC	MIE	RAN	MIC	MIE	RAN	MIC	MIE
5	0	3.99	6.43	6.63	3.18	4.80	3.84	3.04	4.12	4.06
	1	18.46	27.70	25.04	16.91	24.55	28.14	17.01	28.03	27.05
	2	43.42	57.70	61.20	45.11	66.59	64.17	45.79	57.48	66.33
10	0	6.30	10.13	8.94	4.15	4.79	5.57	3.97	3.84	3.92
	1	29.89	44.97	43.73	27.99	42.39	45.21	28.38	43.07	45.43
	2	66.04	81.37	81.28	68.30	87.25	89.02	70.82	88.65	89.84
20	0	9.73	13.72	15.13	5.17	6.77	6.81	4.25	4.95	5.89
	1	44.84	59.23	60.50	47.33	64.79	64.69	45.82	64.89	65.26
	2	82.32	92.30	94.14	87.47	96.52	97.79	89.51	98.02	98.90

Note: Δ = cheating effect in standard deviation units; LR= likelihood ratio; MIC = optimal item selection at the cutoff point; MIE = optimal item selection at the current ability estimate; MLE = maximum likelihood estimation; RAN = random item selection; WLE = weighted likelihood estimation.

Study 1: Fixed Length Confirmation Test Using the LR Test

The first simulation study was designed to answer two research questions: Which of the three item selection methods described above is the most efficient? Can ability estimation using WLE increase classification accuracy over the more commonly used MLE? The second question was motivated by research that has found that ability estimation accuracy can be increased using alternative estimation algorithms compared to MLE (e.g., Warm, 1989). The questions were investigated for adaptive unproctored tests consisting of 10, 30, and 50 items, and fixed length confirmation tests consisting of 5, 10, and 20 items for item bank sizes of 100, 200, and 400 items. The outcome variable was the proportion of correct classifications.

The results of the first simulation study did not vary by size of the item bank so only the results for an item bank consisting of 200 items are presented in Table 1. The table reports the percentage of respondents in each condition who had their unproctored test score rejected.

The number of accurate classifications was higher when using WLE compared to MLE for cheaters ($\Delta = 1$ and $\Delta = 2$), while both methods maintained similar Type I error rates for confirmation test lengths of 10 and 20 items.

The item selection algorithm also had an effect on the efficiency of the confirmation test. Although, there was not a great difference between MIC and MIE item selection procedures, both performed better than random item selection (RAN) across most conditions. In fact, MIC and MIE item selection methods performed similar to random item selection with half the number of items in the confirmation test (e.g., 10 instead of 20), when WLE was the estimation method. An unproctored test length of 10 items resulted in high Type I error rates in the confirmation test due to a lack of precision in the unproctored test.

The results of the first simulation study suggest that the most effective method for constructing a fixed length confirmation test is to use WLE for ability estimation and MIC or MIE for item selection. With these methods, a confirmation test length of 10 items was necessary before the confirmation test obtained acceptable power. Additionally, the confirmation test performed best when the unproctored test was at least 30 items.

Study 2: Defining the Cutoff Points That Correspond to Nominal Type I Error Rates for a Sequential Confirmation Test

The second simulation study was used to define the cutoff points that correspond to nominal Type I error rates for a sequential confirmation test and to test whether the empirical Type I error rates matched the nominal Type I error rates. The decision algorithm for the sequential confirmation test procedures requires a cutoff point from equation 17 before testing begins. The algorithm uses the cutoff point to test the probability that the respondent's ability based on the confirmation test response pattern is above or below that point. Equation 17 includes θ and the standard error obtained from the unproctored test, as well as g, which is a constant depending on a combination of the anticipated error in the confirmation test and the desired Type I error rate for the combined testing procedure. The value of g is obtained by simulating the confirmation test a number of times for a representative sample of respondents. Therefore, a simulation of the values of g that correspond to particular Type I error rates based on the characteristics of the confirmation test was necessary before a sequential confirmation test could be performed.

The choice of g only takes the characteristics of the confirmation test into consideration; therefore, one value of g should result in a stable Type I error rate, regardless of the precision of the unproctored test. Precision in the unproctored test was varied in the current example by varying the length of the unproctored test. Table 2 compares the empirical and nominal Type I error rates across several conditions to test the empirical consequences of using g in equation 17 to set a cutoff point for a sequential confirmation test. The optimal values of g that correspond to Type I error rates of .10, .05, and .01 are reported. To estimate the Type I error rates under optimal conditions, a conservative critical value of $\alpha = \beta = .01$ and an indifference zone of .1 were set for this simulation so all respondents were administered all ($N = 10$ or $N = 20$) items in the confirmation test.

The table provides optimal values of g, for setting the Type I error rate in the sequential confirmation test; these are used in the remainder of the article. The difference between empirical and nominal Type I error rates in Table 2 appear to be due to random error and do not provide evidence for systematic differences. Therefore, this method is acceptable for selecting a single value for *g* that maintains a consistent Type I error rate independent of the precision of the unproctored test.

Table 2. Theoretical and Empirical Type I Error Rates for an Adaptive Confirmation Test Using the Truncated Sequential Probability Ratio Test (TSPRT) for Different Values of the Constant g

			Confirmation test length = 10 Unproctored test Length				Confirmation test length = 20 Unproctored test Length		
Nominal Type I error	Item selection method	g	10	30	50	g	10	30	50
.10	RAN	0.80	0.088	0.102	0.102	0.55	0.950	1.010	0.980
	MIC	0.50	0.091	0.097	0.110	0.35	0.097	0.089	0.102
	MIE	0.50	0.093	0.098	0.104	0.35	0.099	0.100	0.103
.05	RAN	1.20	0.053	0.056	0.051	0.85	0.520	0.490	0.520
	MIC	0.75	0.052	0.050	0.045	0.55	0.054	0.049	0.048
	MIE	0.75	0.055	0.055	0.051	0.55	0.055	0.049	0.048
.01	RAN	3.00	0.023	0.022	0.017	1.50	0.008	0.007	0.009
	MIC	1.30	0.010	0.004	0.007	0.95	0.016	0.075	0.072
	MIE	1.30	0.013	0.008	0.010	0.95	0.015	0.082	0.080

Note: g = constant representing a combination of the anticipated error in the confirmation test and the desired Type I error rate for the combined testing procedure; MIC = optimal item selection at the cutoff point; MIE = optimal item selection at the current ability estimate; RAN = random item selection.

Study 3: Sequential Confirmation Test Using the TSPRT

There were two main objectives for conducting the third simulation study. The first was to establish an effective method for using a TSPRT for sequential confirmation testing. The second was to evaluate if a sequential procedure could reduce the number of items needed in the confirmation test while maintaining the same level of classification accuracy. A maximum number of items administered in the test was set at $N = 10$. In addition to the number of correct classifications, the average number of items needed to make a classification was used as a criterion to measure the increased efficiency of the procedure compared to the fixed length test. Varying the following variables assessed the most effective method of conducting a confirmation test using the TSPRT:

a) Item selection procedure: RAN, MIC, and MIE.
b) Unproctored test length: 10, 30, and 50.
c) Cheating effect in the unsupervised test: $\Delta = 0$, 1, and 2.
d) Critical value for the LR test: $\alpha = \beta = .01$, .05, .1, .2, and .4.
e) Indifference zone: $\delta = .1$ to .3 in increments of .05.

The results of the third simulation study are presented in Table 3 for an item bank size of 200 items; similar findings were obtained with different size item banks. To save space, only the results for an indifference zone of $\delta = .25$ are presented in Table 3. $\delta = .25$ is larger than the indifference zone typically used in classification testing; however, it produced the best results in the current context. This is a consequence of a short maximum test length because small values of δ resulted in nearly all tests progressing until the maximum number of items had been reached. Eggen (1999) reported a slight decrease in the percentage of correct decisions when increasing δ; however, the decrease in precision was not evident until $\delta = .3$ in this study. Therefore, a value of .25 will be used in the remainder of this article. Similarly, the results for critical values of $\alpha = \beta = .01$ and .05 were too conservative for the current context because the test was never terminated before the maximum number of items had been reached. Therefore, less conservative critical values of $\alpha = \beta = .1$, .2,

Table 3. Percentage of Aberrant Response Classifications and Average Number of Items for a Sequential Adaptive Confirmation Test Using the Truncated Sequential Probability Ratio Test (TSPRT)

		Unproctored test length = 10						Unproctored test length = 30						Unproctored test length = 50					
		RAN		MIC		MIE		RAN		MIC		MIE		RAN		MIC		MIE	
$\alpha = \beta$	Δ	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*
.1	0	5	9.7	5	9.2	5	9.3	5	9.8	5	9.5	5	9.7	6	9.9	4	9.4	5	9.6
	1	22	10.0	34	9.8	34	9.8	27	10.0	44	9.9	43	10.0	28	10.0	47	9.9	48	10.0
	2	53	10.0	76	9.6	74	9.8	67	10.0	90	9.6	89	9.8	72	10.0	93	9.6	93	10.0
.2	0	5	8.5	5	7.1	5	6.9	5	8.7	4	7.5	5	7.5	5	8.8	5	7.9	6	7.9
	1	22	9.4	34	8.6	34	8.6	26	9.7	44	9.1	45	9.3	29	9.8	48	9.2	48	9.3
	2	57	9.7	75	8.4	76	8.5	69	9.7	91	7.8	90	7.9	72	9.7	93	7.8	93	7.7
.4	0	6	3.9	9	2.5	9	2.4	8	4.2	11	2.6	10	2.8	7	4.2	10	2.5	12	3.1
	1	24	4.8	38	3.1	38	3.1	29	5.0	46	3.1	47	3.3	32	5.2	49	3.1	49	3.3
	2	55	5.0	72	2.9	73	2.8	67	5.1	84	3.0	83	3.1	70	5.0	87	3.0	87	2.9

Note: Type I error = 5%; $\alpha = \beta$ = values that determine the required magnitude that the likelihood ratio (LR) test must surpass to stop the confirmation test early; Δ = cheating effect in standard deviation units; *C* = percentage of aberrant response classifications; *K* = average test length; MIC = optimal item selection at the cutoff point; MIE = optimal item selection at the current ability estimate; RAN = random item selection.

and .4 are reported in Table 3. The table reports the percentage of respondents who had their unproctored test result rejected (*C*) and the average number of items used in the confirmation test (*K*) for each condition.

The results of the TSPRT indicate that the MIC and MIE item selection procedures performed similarly in terms of classification accuracy. The MIC procedure used an average of .1 fewer items to obtain a similar level of accuracy. Both MIC and MIE outperformed random item selection in terms of power and efficiency across all conditions. These item selection methods increased the percentage of cheaters identified by approximately 20% over random item selection, while maintaining a similar Type I error rate. Changing the critical value from $\alpha = \beta = .1$ to .2 led to a similar number of correct classifications with fewer items. The use of the critical value $\alpha = \beta = .4$ led to inflated Type I error rates and decreased the power in the test. Therefore, a critical value of $\alpha = \beta = .2$ obtained the best results and will be used in the remaining simulations.

The results from the sequential and fixed length tests can be compared for a confirmation test consisting of 10 items and an unproctored test of 30 items because the Type I error rate was approximately 5% for both methods. The sequential adaptive confirmation test using the TSPRT resulted in a greater number of correct classifications for the MIC and MIE item selection methods compared to the fixed length tests. The sequential method was also more efficient and saved an average of 1.9 items with the MIC item selection method, 1.8 with the MIE method, and 0.6 with random item selection method when averaging across the three cheating effects.

Using a single value for g in equation 17 has the disadvantage that the Type I error rate applies to the entire distribution and is not defined conditionally at each point on the ability scale. Therefore, an analysis of classification accuracy across ability levels was done to assess whether the probability of being classified as a possible cheater depended on the respondent's ability. The conditional precision was simulated in exactly the same way as in the previous simulation; however, 10,000 response patterns were simulated for each of the following θ values: -2, -1, 0, 1, and 2. Table 4 presents the conditional precision of the confirmation test using the TSPRT.

Table 4. Conditional Precision of the Sequential Adaptive Confirmation Test Using the Truncated Sequential Probability Ratio Test (TSPRT) With an Unproctored Test of 30 Items

	$\theta = -2$		$\theta = -1$		$\theta = 0$		$\theta = 1$		$\theta = 2$	
Δ	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*
	RAN									
0	19	9.8	8	9.3	3	8.5	2	8.1	3	8.9
1	36	9.9	30	9.8	25	9.6	23	9.7	23	9.9
2	72	10.0	73	9.7	72	9.6	64	9.8	52	10.0
	MIC									
0	6	7.8	5	7.7	5	7.4	4	7.6	5	8.0
1	46	9.3	46	9.1	46	9.1	41	9.1	36	9.1
2	93	7.9	92	8.0	91	8.1	87	8.5	78	8.9
	MIE									
0	7	8.7	4	7.5	5	7.8	5	7.7	5	7.9
1	45	9.6	47	9.4	45	9.2	42	9.3	40	9.2
2	92	7.8	92	8.2	91	7.9	88	8.4	80	8.9

Note: Type I error = 5%; Δ: cheating effect in standard deviation units; *C* = percentage of aberrant response classifications; *K* = average test length; MIC = optimal item selection at the cutoff point; MIE: optimal item selection at the current ability estimate; RAN = random item selection.

For random item selection, the empirical Type I error rate (displayed in the columns labeled *C*) ranged from 2% to 19% across ability levels. The Type I error rate ranged from 4% to 6% and from 4% to 7% for MIC and MIE item selection methods, respectively. The deviation from the nominal Type I error rate of 5% was greatest for the lowest ability level of $\theta = -2$. The power of the MIC and MIE methods at detecting cheaters was also relatively consistent for ability levels $\theta = -2, -1$, and 0 and only became lower at ability levels $\theta = 1$ and 2. Therefore, it is justifiable to use a single value of g in equation 17 with MIC and MIE item selection. For random item selection, the percentage of aberrant response classifications was far less stable across values of θ. For instance, for $\theta = -2$, the percentage was 19, while it was 3 for $\theta = 2$. So, this method leads to local bias and should not be used with random item selection.

Study 4: Sequential Confirmation Test Using the SCTSPRT

The main purpose of the fourth simulation study was to check if the SCTSPRT could be used to shorten the TSPRT in computerized adaptive confirmation testing. An SCTSPRT was developed based on the most effective method identified in Study 3. The SCTSPRT halts testing in cases in which the probability of a change of the classification decision is smaller than the predefined value y from equations 14 and 15. Therefore, y was varied from .75 to 1 in intervals of .05.

The results of the fourth simulation study are presented in Table 5. To save space, results for only a selection of the values of y are presented.

As expected, Table 5 indicates that the accuracy of the confirmation test decreased as the value of y decreased. The Type I error rate was maintained as long as $y \geq .95$; however, the empirical Type I error rate was above the nominal Type I error rate, for $y \leq .9$. Therefore, values of $y \geq .95$ would be suggested for designing a confirmation test with a predefined Type I error rate. The detection accuracy was only .02% lower for $y = .95$ compared to $y = 1$ when averaging across MIC and MIE for all conditions of Δ. However, there was a decrease in the number of items necessary for making a decision from 5.96 to 5.4 items for MIC and from 6.07 to 5.5 items for MIE. These results suggest that the SCTSPRT with $y = .95$ is the most efficient (in terms of the number of items used), yet accurate

Table 5. Percentage of Aberrant Response Classifications and Average Number of Items for a Sequential Adaptive Confirmation Test Using the Stochastic Curtailed Truncated Sequential Probability Ratio Test (SCTSPRT)

		Length of the unproctored test											
		10 items				30 items				50 items			
		MIC		MIE		MIC		MIE		MIC		MIE	
y	Δ	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*	*C*	*K*
1.00	0	4	4.0	5	5.0	5	5.9	4	5.3	4	5.3	5	5.6
	1	32	6.3	34	6.6	45	7.1	48	7.3	45	7.0	46	7.5
	2	77	6.3	75	6.2	91	5.6	91	5.7	93	6.1	93	5.4
.95	0	5	4.5	5	4.2	5	4.2	5	5.0	5	5.1	5	5.2
	1	34	6.1	34	6.2	43	6.4	41	6.2	48	6.3	48	6.7
	2	78	5.6	77	5.9	91	4.9	90	5.3	93	5.5	93	4.8
.90	0	6	4.1	7	4.3	6	4.4	6	4.7	6	4.8	6	4.6
	1	34	5.4	34	5.7	44	6.1	43	6.3	50	6.0	47	6.2
	2	73	5.2	75	5.1	88	4.3	88	4.7	92	5.2	92	5.2
.80	0	7	3.9	7	3.9	8	3.9	8	4.1	7	4.2	7	4.3
	1	36	4.9	36	5.0	43	4.9	45	4.8	47	5.4	48	5.1
	2	75	4.6	74	4.7	87	4.3	87	4.3	90	4.5	91	4.4

Note: Type I error = 5%; *y* = determines how high the probability of retaining the current decision must be for the test to be stopped early; Δ = cheating effect in standard deviation units; *C* = percentage of aberrant response classifications; *K* = average test length; MIC = optimal item selection at the cutoff point; MIE = optimal item selection at the current ability estimate.

(in terms of correct classifications), method for conducting a computerized confirmation test. The accuracy of the MIC and MIE item selection methods was virtually identical; however, the MIC used an average of .1 fewer items when averaging across all conditions for $y \geq .95$. The Type I error rate was consistent across unproctored test lengths; however, the power of the test increased as the length of the unproctored test increased.

A comparison of the two sequential confirmation test methods, the TSPRT and the SCTSPRT, showed that both tests retained similar precision; however, the SCTSPRT led to classifications with fewer items than the TSPRT. The SCTSPRT led to an average decrease of 2.2 and 2.1 items for the MIC and MIE item selection methods, respectively, when y was set to 1; and of 2.8 and 2.7 when y was set to .95. This represents a reduction of over 25% of the items in the confirmation test.

Study 5: A Simulated Personnel Selection Procedure

So far, we have discussed the accuracy of a confirmation test in terms of classification accuracy. However, we were also interested in the effect of cheating and the returns obtained from using the sequential adaptive confirmation test for a selection procedure where those candidates who obtain the highest score on a test are hired for a job. A sample of 10,000 job candidates was generated with a normal distribution with a mean of 0 and a standard deviation of 1 representing their true ability. Cheating was defined in the same way as in the previous simulations with the cheating effect (Δ) = 0, 1, and 2 standard deviation units in the confirmation test. The proportion of cheaters in the sample was varied as 0%, 10%, and 20%. Candidates were assigned randomly to a cheating or honest condition. Based on research findings presented in Cizek (1999), assignment to the cheating condition was also based on an assumed correlation between ability and the propensity to cheat of -0.3. The selection ratio for the example presented below was 50%, although other ratios were also explored.

Table 6. Selection Accuracy for an Unproctored Adaptive Test of 30 Items: Selection of the Top 5,000 Candidates Based on the Unproctored Test Score in a Sample of 10,000 Applicants

		Candidate's true ability			
Cheating %	Δ	False selections	Correct selections	Benefit ratio	Correlation between true ability and unproctored test score
0%	0	493	4,507	9.14	0.95
10%	1	735	4,265	5.80	0.90
10%	2	738	4,262	5.78	0.80
20%	1	974	4,026	4.13	0.87
20%	2	1,053	3,947	3.75	0.72

Note: Δ = cheating effect in standard deviation units.

The benefits ratio of the test is measured as the ratio of correct selections to incorrect selections. Table 6 presents the benefits ratio based on the cross-classification of true and reported scores for the unproctored test, without a confirmation test. As a more traditional validity measure, Table 6 also presents Pearson's correlation coefficient between the true ability and the unproctored test score for all 10,000 job candidates.

The benefits ratio of the unproctored test alone, when there was no cheating, was 4,507:493 or 9.14 correct selections for each incorrect selection. The effect of cheating decreased the benefits ratio of the test to 5.8 with 10% cheaters and 4.13 with 20% cheaters, when the cheating effect was moderate. The decrease was even larger when the effect of cheating was large; 5.78 with 10% cheaters and 3.75 with 20% cheaters. The correlation between the true ability and the unproctored test score was 0.95, when there was no cheating in the sample of job candidates. The correlation was reduced to 0.90 and 0.80, respectively, for moderate and large cheating effects when 10% of the candidates in the sample were cheaters. The correlation became 0.87 and 0.72, respectively, for moderate and large cheating effects when 20% of the candidates in the sample were cheaters.

Table 7 presents a continuation of the selection procedure presented above where the candidates who were selected based on their unproctored test result were administered an adaptive confirmation test using the SCTSPRT as a follow-up. The table illustrates a cross-classification of the candidates' true ability and the decision based on the confirmation test. A benefits ratio of the combination of the unproctored test and the confirmation test is presented based on the candidates whose unproctored test result was accepted. The table also presents Pearson's correlation coefficient between the true ability and the unproctored test score, after eliminating the candidates who had their unproctored test score rejected in the confirmation test. The last two columns of Table 7 present a cross-classification of candidates' true condition (honest/cheater) and their confirmation test result.

The addition of the confirmation test in the selection procedure only slightly increased the benefits ratio of the selection procedure from 9.14 to 9.90 (4,274:432) when there was no cheating. The increase occurred because a larger percentage of falsely selected candidates (14%) had their unproctored test result rejected compared to the group of respondents that was correctly selected (5%). This occurred because the unproctored test score was inflated based on random error for the falsely selected candidates. This finding illustrates that the use of a confirmation test also helps detect candidates who are falsely selected even when the source of the inflated unproctored test result is not due to cheating. The benefit of using the confirmation test increased as the percentage of cheaters in the applicant pool increased. When the effect of cheating was moderate, the increase in benefits ratio went from 5.80 to 8.24 with 10% cheaters and from 4.13 to 6.28 with 20% cheaters. The benefits ratio increased from 5.78 to 14.73 with 10% cheaters and from 3.75 to 19.44 with 20% cheaters

Table 7. Selection Accuracy for an Unproctored Test of 30 Items With a Follow-Up Stochastic Curtailed Truncated Sequential Probability Ratio Test (SCTSPRT) Confirmation Test

			Candidates true ability				Candidates true condition	
Cheating %	Δ	Decision from confirmation test	False selections	Correct selections	Benefit ratio	Correlation between true ability and UIT score[a]	Cheater	Honest
0%	0	a	432	4,274	9.90	0.95		4498
		r	61	233				286
10%	1	a	458	3,773	8.24	0.93	476	3,755
		r	277	492			478	291
10%	2	a	245	3,608	14.73	0.93	75	3,778
		r	493	654			921	226
20%	1	a	535	3,360	6.28	0.91	980	2,915
		r	439	666			866	239
20%	2	a	147	2,896	19.44	0.93	150	2,893
		r	906	1,051			1801	156

Note: Δ = cheating effect in standard deviation units; a = accept UIT score; r = reject UIT score.
[a] Correlation between true ability and unproctored test score after eliminating the candidates who had their unproctored test score rejected in the confirmation test.

when the effect of cheating was high. This occurred because the proportion of respondents who had cheated became larger in the sample of selected candidates. Therefore, the benefit of using the confirmation test was larger because the percentage of cheaters who were identified by the confirmation test increased. The benefit of using a confirmation test was also larger when the ratio of the candidates who were selected became smaller because the proportion of cheaters in the group of selected candidates increased.

A comparison of the correlation between the true ability and the unproctored test score in Tables 6 and 7 gives an indication of the impact of the confirmation test in terms of a correlation coefficient. The correlation remained unchanged at .95 when there was no cheating. The inclusion of a confirmation test increased the correlation coefficient from .90 to .93 and from .80 to .93, for moderate and large cheating effects, respectively, when 10% of the candidates had cheated. The correlation between the true ability and the unproctored test score increased from .87 to .91 and from .72 to .93, for moderate and large cheating effects, respectively, when 20% of the candidates had cheated on the unproctored test.

Discussion

With the increased need for flexible selection procedures, we foresee an increased use of UIT in the future. This study showed that cheating can have a detrimental effect on the validity of a selection procedure and the possibility of using a confirmation test can decrease this effect. In addition, there is evidence that a significant number of test candidates would cheat in an unproctored high stakes test; therefore, confirmation testing is important for validating UIT results. The article investigated four research questions related to the problem of verifying unproctored Internet test results by means of a short confirmation test. The research questions are repeated below with conclusions and practical considerations based on the results of the study.

Which item selection method performs best in fixed length and sequential confirmation testing procedures? The results illustrated that adaptive selection methods were more powerful than random item selection in terms of accurately classifying respondents as honest or cheaters and were more efficient in terms of the number of items required to make a decision in the sequential procedures. A comparison of the two adaptive methods led to the conclusion that maximum information at the cutoff point (MIC) was slightly more efficient than maximum information at the current θ estimate (MIE) for the sequential procedures. This is consistent with previous findings from adaptive classification test literature (e.g., Eggen, 1999; Spray & Reckase, 1996). However, this difference was small and may not have practical significance in applied settings.

What are the benefits of using a sequential confirmation test, rather than using a fixed length confirmation test based on efficiency in terms of the number of items administered and the power to identify cheaters? Regarding this research question, we first discuss the benefits in terms of efficiency. The primary advantage of using the sequential methods was that they required fewer items to obtain a similar level of power. The SCTSPRT method required one half of the items compared to the fixed length procedure using the WLE for ability estimation; and approximately one fourth compared to using the MLE for ability estimation. One of the criticisms of UIT has been that it lacks a quick method for verifying unproctored test results. This improvement will make it more attractive to take advantage of the benefits of UIT because it provides a quick testing process that can reduce cost by limiting on-site or proctored testing. An additional advantage was that the sequential methods provide the option of controlling the Type I error rate in the test. This allows the test users to set the error rate according to their particular needs. Although the sequential procedures were more efficient than the fixed length confirmation test methods, there may be applications where a fixed number of items are desired. When a fixed length confirmation test was used, the results showed that ability estimation using WLE increased classification accuracy compared to the more commonly used MLE.

The power of the sequential confirmation test methods introduced in this article is directly related to the characteristics of the test and the degree of cheating. The adaptive SCTSPRT method resulted in over 90% accurate detection of cheaters when the unproctored test length was at least 30 items and the respondent's cheating effect was large (two standard deviation units). The power of the test was lower for short unproctored test lengths and for moderate cheating effects (one standard deviation unit). It is expected that an unproctored test length of 30 items is acceptable for most applied settings because the flexibility of UIT allows the respondent to take the test at their convenience. The practical consequences of the lower power for moderate cheating effects will vary based on the intended use of the test for the particular testing organization. The item bank in the confirmation test can be expanded, if the power of the test is not sufficient in certain situations. Other alternatives, such as a long supervised test that counts as the result of record could also be considered. The SCTSPRT method proposed in this study provides the option of increasing power by increasing the Type I error rate in the test; however, this results in an increase in the number of false positives, which can lead to a number of negative ethical and possible legal consequences.

What are the benefits of using the stochastic curtailed version of the TSPRT compared to using the TSPRT? The stochastically curtailed version of the TSPRT led to a reduction in test length of over 25% or over 2 items compared to the TSPRT. Therefore, the effect of using the stochastically curtailed version of the TSPRT for confirmation testing is even larger than what has been reported for computerized classification testing (e.g., Finkelman, 2008).

Does a confirmation test decrease the effect of cheating on the validity of a selection procedure? The simulation study demonstrated that cheating could have a detrimental effect on the validity of a selection procedure and illustrated that the use of a confirmation test could hinder the negative effect of cheating on validity. The advantage of using a confirmation test increased when the effect of cheating increased and as the number of cheaters in the applicant pool increased.

Limitations

The confirmation test methods proposed in this article are limited to the context of knowledge and ability tests and are only applicable to situations in which the unproctored test score is used as the operational score. A further limitation is that the SCTSPRT is a CAT/IRT-based confirmation test approach that demands a large item bank, which requires considerable resources to develop. Many developments in CAT research have been focused on decreasing the resources required for CAT item bank development. Advances such as automatic calibration procedures (e.g., Kingsbury, 2009; Makransky & Glas, 2010) and automated item generation (e.g., Glas & van der Linden, 2003; Glas, van der Linden, & Geerlings, 2010) provide options for decreasing the resources required for developing an adaptive test; however, the cost is still higher than the price of developing a traditional test. A related issue is that an adaptive approach requires computer access at the supervised testing location and relies on fast and reliable Internet access to allow communication between the local computer and the decision algorithm. These technological limitations are no longer as prominent in most testing locations, but they may still be prominent in others.

An additional limitation to using the SCTSPRT for confirmation testing is that this method is designed to test the hypothesis that the result in the first test is valid. Therefore, it does not add to the reliability of the ability estimate. Another disadvantage to using the SCTSPRT is that it is an adaptive procedure, which means that respondents are administered a different number of items depending on their responses. This can result in perceptions of injustice, specifically for respondents who have their UIT score rejected after the administration of few items. However, in practice, CAT procedures prove easy to explain to respondents. Some test administrators may also want to know the number of items that will be administered so that they can plan the assessment in relation to other activities such as interviews. Therefore, it is conceivable that these practitioners may prefer a fixed length test even though it is less efficient.

Finally, the SCTSPRT method for confirmation testing proposed in this article takes into account the Type I error in the test but not the Type II error. This limits the generalizability of the method to contexts where the importance of Type I error outweighs the importance of Type II error. The generalization of the method to control the Type II error would be straightforward and would consist of setting a less conservative cutoff point with the consequence of increasing the number of false positives. A related issue is the modest power of this method for detecting moderate cheating effects. Although the SCTSPRT method represents an improvement compared to existing methods, the lack of power for detecting moderate cheating effects continues to be a challenge for confirmation testing.

A general challenge for practitioners using a confirmation test is the issue of how to deal with UIT scores that are rejected by the confirmation test. The International Guidelines on Computer-Based and Internet-Delivered Testing (2005) provide a good outline of the procedures that should be in place before conducting a confirmation test. Burke et al. (2006) also offer a detailed description of a procedure that can be followed when a UIT score is rejected. One suggestion is to provide the possibility of taking a full-length supervised test to respondents who have their unproctored score rejected.

Future Research

We anticipate several future research directions related to confirmation testing, the application of the SCTSPRT, and UIT in general. The first is to add to the results of the current study by incorporating practical constraints such as content balancing and item exposure control. In addition, methods for controlling the Type I and Type II error rates simultaneously might be investigated. The current article generalized research from classification testing to the context of confirmation testing. Future research could investigate other approaches for verifying UIT results to explore if there are alternative methods that can increase the power to detect cheaters while maintaining test efficiency. Another area of research is related to candidate's perceptions of confirmation tests and UIT in general. Different methods for processing UIT scores that are rejected could be investigated.

A further line of future research is related to methods of proctoring online tests. Biometric authentication systems are currently used together with online proctoring through the use of webcams or a confirmation test (e.g., Foster, 2009). Online proctoring can limit the flexibility of the test and can be expensive. Therefore, future research and technological advances could investigate ways to make these methods cheaper and more widely available.

Finally, the adaptive confirmation test using the SCTSPRT was applied specifically to the verification of UIT in this article. However, there are many assessment applications where the goal of the examination is to evaluate if there has been a change on the latent trait. Some examples could be to assess change: after a training program, after administering a drug in clinical trials, or after administering a test at two different time points. In these applications, the fundamental objective is not to reassess the position of the individual on the latent trait, but alternatively to measure if a change on the position of the latent trait has taken place. Future research could investigate the possibility of using the SCTSPRT in these contexts.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The corresponding author's research is funded by Master Management International A/S and a grant from the Danish government.

References

Anderson, C. D., Warner, J. L., & Spector, C. E. (1984). Inflation bias in self-assessment examination: Implications for valid employee selection. *Journal of Applied Psychology*, *69*, 574-580.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*, 44-49.

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored Internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 39-45.

Automatic Data Processing Inc. (2008). *2008 ADP screening index*. Retrieved December 21, 2009, from www.adp.com/media/press-releases/2008-news-releases/adp-annual-pre-employment-screening-index.aspx

Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist-practitioner model: How I-Os should respond to unproctored Internet testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 58-63.

Bellezza, F. S. and Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, *16*, 151-155.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Burke, E., van Someren, G., & Tatham, N. (2006). *Verify range of ability tests: Technical manual*. Thames Dittion, Surrey: SHL Group.

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment*, *11*, 113-120.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal detection of mismeasured individuals. *Applied Measurement in Education*, *9*, 47-64.

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

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*, 713-734.

Fallaw, S. S., Solomonson, A. L., & McClelland, L. (2009). *Current trends in assessment use: A multi-organizational survey*. Poster submitted for the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Finkelman, M. (2003). *An Adaptation of Stochastic Curtailment to Truncate Wald's SPRT in Computerized Adaptive Testing* (CSE Report 606). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442-463.

Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 31-34.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 249-263.

Glas, C. A. W., van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 303-331). New York, NY: Springer.

International Guidelines on Computer-Based and Internet Delivered Testing (2005). Retrieved January 5, 2010, from www.psychtesting.org.uk/downloadfile.cfm?file_uuid=9BD783C8-1143-DFD0-7E98-798CD61E4F00&ext=pdf

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials.* Boca Raton, FL: Chapman & Hall/CRC.

Jiao, H., Wang, S., & Lau, C. A. (2004). *An investigation of two combination procedures of SPRT for three-category classification decisions in computerized classification test*. Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

Kaminski, K. A., & Hemingway, M. A. (2009). To proctor or not to proctor? Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 24-26.

Kingsbury, G. G. (2009). *Adaptive item calibration: A simple process for estimating item parameters within a computerized adaptive test*. Paper presented at the 2009 GMAC conference on computerized adaptive testing, Minneapolis, Minnesota.

LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, *12*, 296-319.

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575-596). Mahwah, NJ: Lawrence Erlbaum.

Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, *11*.

Muraki, E., & Bock, R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3) [Computer software]. Chicago, IL: Scientific Software.

Nye, C. D., Do, B., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16*, 112-120.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660-679.

Partchev, I. (2009). 3PL: A useful model with a mild estimation problem. *Measurement*, *7*, 94-96.

Pearlman, K. (2009). Unproctored Internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 14-19.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York, NY: Academic Press.

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.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computerized adaptive test*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, April 5–7, 1994, New Orleans, LA.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational & Behavioral Statistics*, *21*, 405-414.

Thissen, D., & Mislevy, R. J. (2000). Test algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101-134). Hillsdale, NJ: Lawrence Erlbaum.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 2-10.

van der Linden, W. J., & Sotaridona, L. (2002). *A statistical test for detecting answer copying on multiple-choice tests*. University of Twente Research Report 02-04. Enschede, the Netherlands.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston, MA: Kluwer-Nijhoff.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, *39*, 235-274.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest: General item response modelling software (Version 2.0) [Computer software]. Camberwell, Australia: ACER.

Bios

Guido Makransky holds the position of senior psychometrician at Master Management International A/S, Denmark. He is currently working on his PhD at the University of Twente, the Netherlands. His research focuses on the development and application of item response theory and computer adaptive testing to practical problems in organizational settings.

Cees. A. W. Glas holds the position of head of the Department of Research Methodology, Measurement and Data Analysis, at the Faculty of Behavioral Science of the University of Twente Enschede, the Netherlands. The focus of his scientific work is application of item response theory models to educational and psychological testing. He conducted research projects for such institutes as the Dutch National Institute for Educational Measurement (Cito, the Netherlands), and the Law School Admission Council (USA). With his department, he participates in the 2009-cylce of the PIS A project for the development of background questionnaires and the analysis of the relation between background variables and educational outcomes. He published more than 60 book chapters and articles in scientific journals.