Using Cumulative Sum Statistics to Detect Inconsistencies in Unproctored Internet Testing

Tendeiro, Jorge N.; Meijer, Rob R.; Schakel, Lolle; Maij-de Meij, Annette M.

Published in:
Educational and Psychological Measurement

DOI:
10.1177/0013164412444787

Educational and Psychological Measurement
73(1) 143–161

DOI: 10.1177/0013164412444787
http://epm.sagepub.com
SAGE

Using Cumulative Sum Statistics to Detect Inconsistencies in Unproctored Internet Testing

Jorge N. Tendeiro[1], Rob R. Meijer[1], Lolle Schakel[2], and Annette M. Maij-de Meij[2]

Abstract

Unproctored Internet Testing (UIT) is becoming more popular in personnel recruitment and selection. A drawback of UIT is that cheating is easy and, therefore, a proctored test is often administered after an UIT procedure. To detect inconsistent test scores from UIT, a cumulative sum procedure (CUSUM) is proposed. This procedure is applied to empirical data from an adaptive computer-based test in a real personnel selection context. The usefulness of the CUSUM is illustrated, and the unique contribution of the CUSUM to existing procedures is discussed.

Keywords

Unproctored Internet Testing, inconsistent response behavior, computer-based testing, cumulative sum statistic

The use of Unproctored Internet Testing (UIT) is rapidly becoming popular in personnel selection and recruitment (Naglieri et al., 2004; Tippins, 2009). There are many advantages associated with UIT: It can save time and money for both test providers and test takers, a broad range of candidates can be reached, it is available around the

[1]University of Groningen, Groningen, Netherlands
[2]PiCompany, Utrecht, Netherlands

Corresponding Author:
Jorge Tendeiro, Department of Psychometrics and Statistics, Faculty of Psychology, University of Groningen, Grote Kruisstraat 2/1, Groningen, 9712 TS, Netherlands
Email: j.n.tendeiro@rug.nl

clock at the best convenience for each person, it facilitates test scoring, and it speeds up feedback (Naglieri et al., 2004). Also, the continuing technological advances that we have witnessed in the past decade solved problems in UIT related to speed of computer systems and reliability of the Internet. Although UIT is still not commonly accepted by psychologists in general (Tippins, 2009; Tippins et al., 2006), many private organizations in the United States do seem to have embraced UIT for personnel recruitment and selection (Pearlman, 2009).

This is not to say that there are no challenges in UIT. Often mentioned problems are nonstandardized testing conditions, test security, candidate identification, and cheating (Buchanan & Smith, 1999; Naglieri et al., 2004; Tippins et al., 2006). The term *cheating* may refer to any kind of behavior from the test taker that intends to deceive the test provider by enhancing the ability or trait that the test tries to measure. Different types of cheating can be distinguished. For example, a candidate may take the test for others, or candidates may use nonallowed sources such as books, websites, or mobile phones. Although these types of cheating may also occur in proctored testing, the lack of a human proctor in UIT may increase cheating behavior. It is possible to use technology to try to prevent fraudulent behaviors (e.g., webcams), but as argued in Tippins et al. (2006, p. 205), that would mean that ''the setting is no longer unproctored, it is just proctored remotely.'' Naglieri et al. (2004) discussed that it is important (specially for high-stakes testing) to implement security measures on the Web servers involved in UIT in order to improve test security. Burke (2009) argued that using *Web patrols* to search the Internet for webpages that give nonauthorized access to testing contents can also be implemented. Technical methods to prevent fraud such as using webcams or Web patrols are of a *preventive* nature, that is, intended to discourage candidates from engaging in illegal practices, and are not further discussed.

The most popular method to validate scores from UIT consists of inviting candidates for a second, proctored test administration, often referred to as a *confirmation* or *verification* test. The confirmation test is taken in a secured, supervised environment provided by the organization responsible for the test administration. Of course, UIT is only efficient when the resources invested in the confirmation test are reduced to a minimum. This is typically achieved by only inviting for the confirmation test those candidates with a score on the unproctored test higher than some cutoff score. Also, the confirmation test should be constructed in such a way that only a minimum number of items is required to confirm or dismiss the results obtained in the first test. In the present article, we focus specifically on *statistical* methods designed to detect inconsistencies between the scores from the unproctored and the confirmation tests.

Several person-fit methods have been proposed in the literature to detect inconsistent item score patterns; see Meijer and Sijtsma (2001) and Karabatsos (2003) for reviews. An interesting practical example is given in Jacob and Levitt (2003), who investigated suspicious test score fluctuations and misfitting item score patterns to identify teacher cheating. Methods that directly compare test–retest scores are also available. For example, Guo and Drasgow (2010) introduced the z statistic, which

compares ability estimates from both tests. In the present article, we present a new methodology that consists of studying the compatibility between the ability estimates from the unproctored test and the item score patterns from the proctored test. Our method is based on CUmulative SUM control charts (CUSUMs); see Page (1954), Van Krimpen-Stoop and Meijer (2000, 2001), Armstrong and Shi (2009), and Tendeiro and Meijer (2012). One advantage of the CUSUM methodology is that it uses item-level information to detect inconsistencies, unlike alternative methods which only use ability estimates or sum scores. CUSUMs are specially useful in identifying aberrancies which are sequential in nature. Also, the CUSUM methodology allows displaying results in the form of charts, which have the advantage of requiring minimal knowledge for interpretation of results.

Our main goal is to identify which candidates display a decrease in performance from the first (unproctored) test to the second (proctored) test. We apply the CUSUM methodology to empirical computerized adaptive test (CAT) data containing scores on both unproctored and proctored tests in a personnel recruitment procedure, and we compare the performance of these statistics with alternative methods in the literature.

Control Chart Techniques

Statistical process control (SPC) covers a range of statistical procedures that allow to control and monitor different types of production processes. One of the most important statistical tools in SPC are the CUSUMs discussed by Page (1954). A CUSUM is a chart that allows following a production process in real time by using accumulated information. The procedure detects shifts in the measurements at an early stage in case some kind of anomaly disturbs the production process. A CUSUM is characterized by lower and upper control limits. Once a shift in measurements is big enough and the chart line crosses a control limit then an alarm signal is given, and measures should be taken in order to identify and eliminate the problem and to restore normal production conditions. At that point the CUSUM is reset and production is resumed.

Although SPC was traditionally focused on controlling industrial production, several studies proposed to use these methods to detect unusual response behavior in educational and psychological testing using item response theory (IRT) modeling (Embretson & Reise, 2000). One of the first studies that applied CUSUMs was discussed in Bradlow, Weiss, and Cho (1998), who used CUSUMs to identify candidates with inconsistent response patterns in a CAT. Applications of CUSUM procedures to CATs take into account that CATs are sequential and adaptive procedures, which are used to measure some latent ability θ of a candidate via the administration of a variable number of items. Van Krimpen-Stoop and Meijer (2000, 2001; see also Meijer & van Krimpen-Stoop, 2010) further developed and applied CUSUM procedures. They introduced statistics to determine upper and lower CUSUM control functions for CATs using dichotomous items scores. The CUSUM statistics are evaluated after administration of each item i ($i = 1, \ldots, n$, where n is the number of items in the test). The formula is

$$C_i^+ = \max\{0, T_i + C_{i-0}^+\}, \tag{1}$$

$$C_i^- = \min\{0, T_i + C_{i-0}^-\}, \tag{2}$$

for $i = 1$, and $C_1^+ = C_1^- = 0$. Statistic T_i (van Krimpen-Stoop & Meijer, 2000) can be defined by $\frac{X_i - p_i}{n}$, that is, the difference between the observed and expected scores on item i, X_i, and $p_i = P(\mathrm{X_i} = 1|\theta)$, respectively, corrected for test length. Expected probabilities p_i are computed using any suitable IRT dichotomous model, for example, the one-, two- or three-parameter logistic models (1PL, 2PL, or 3PL models, respectively); see, for example, Embretson and Reise (2000). T_i can be evaluated at the updated ability estimate $\hat{\theta}_{i-1}$ or at the final ability estimate $\hat{\theta}_n$. Upper and lower control limits $U_i(\alpha)$ and $L_i(\alpha)$, respectively, must be estimated (Hawkins & Olwell, 1998). A response string is to be considered inconsistent if, at any step i, $C_i^- \leq L_i(\alpha)$ or $C_i^+ \geq U_i(\alpha)$. The control limits are estimated such that false positives (i.e., falsely detecting inconsistent behavior) are limited by a preselected level α.

Armstrong and Shi (2009) defined an alternative statistic T_i to compute upper and lower CUSUMs. Besides modeling the conditional probability of a correct response to an item, p_i, Armstrong and Shi (2009) also modeled the probability of correctly answering an item in case of underperformance (p_i^L) or overperformance (p_i^U). The latter probabilities should accurately model the specific types of inconsistent behaviors of interest to the researcher. Armstrong and Shi (2009) presented a method that allows to model both p_i^L and p_i^U as quadratic functions of p_i. Tendeiro and Meijer (2012) suggested some improvements for this method. Once models for both p_i^L and p_i^U have been chosen it is possible to define the CUSUM statistics,

$$\begin{aligned} \gamma_i^U &= \ln \frac{(p_i^U)^{x_i}(1-p_i^U)^{1-x_i}}{p_i^{x_i}(1-p_i)^{1-x_i}}, \qquad \gamma_i^L = \ln \frac{p_i^{x_i}(1-p_i)^{1-x_i}}{(p_i^L)^{x_i}(1-p_i^L)^{1-x_i}}, \\ C_i^U &= \max\{0, \gamma_i^U + C_{i-1}^U\}, \qquad C_i^L = \min\{0, \gamma_i^L + C_{i-1}^L\}. \end{aligned} \tag{3}$$

Statistics γ_i^U and γ_i^L are the logarithms of likelihood ratios (Neyman & Pearson, 1933). Upper and lower control limits are estimated such that the rate of false positives is not superior to a preselected level α.

Two-sided CUSUM statistics which combined both C_i^U and C_i^L given in Equation (3) were also considered by Armstrong and Shi (2009),

$$\begin{aligned} &C_{\max}^U = \max\{C_i^U\} \text{ and } C_{\min}^L = \min\{C_i^L\}, \quad i = 1, \ldots, n, \\ &C^{LR} = C_{\max}^U - C_{\min}^L. \end{aligned} \tag{4}$$

A score pattern is out of control whenever C^{LR} is larger than a CUSUM control limit. Tendeiro and Meijer (2012) considered a similar two-sided statistic using the CUSUM statistics defined in Equations (1) and (2), denoted by C_{VM}^{LR}.

The use of lower or upper CUSUM strategies is related to the type of inconsistency that a researcher is mostly interested in. In theory, lower CUSUMs are more sensitive in detecting inconsistent behaviors in which candidates perform under their

real ability level, for example, because of tiredness, subexpertise on the contents being tested, or problems in understanding the language in which the test is written. Upper CUSUMs, on the other hand, are more sensitive in detecting overperformance, for example, because of dishonest behaviors such as cheating. A two-sided CUSUM statistic can be used if a researcher would like to detect both under- and overperformances.

CUSUM charts can be depicted for each candidate and for each CUSUM statistic, if desired. In the ''Results'' section we show examples of CUSUM charts and we show how these charts can be interpreted. CUSUM charts are one of the strengths of the CUSUM methodology because charts facilitate and enrich the interpretation of results.

CUSUMs in Test–Retest Context

As far as we know, there are no CUSUM applications in a test–retest context. The present article tries to fill in this gap. We propose to check, using CUSUMs, whether item scores from the confirmation test are compatible with ability estimates obtained from the unproctored test, $\hat{\theta}_{Un}$. In case of normal response behavior on the unproctored test, and assuming that cheating on the confirmation test is not likely, item scores from the confirmation test and ability estimates obtained from the unproctored test should be compatible. However, when a candidate overperforms during the first test, the item scores on the confirmation test may be unlikely given the ability estimate $\hat{\theta}_{Un}$. Specifically, some form of underperformance might be expected in the confirmation test if $\hat{\theta}_{Un}$ is taken as an acceptable estimate of the candidate's true latent ability. CUSUMs can be used in this context to assess whether $\hat{\theta}_{Un}$ and the response scores from the confirmation test are compatible. A researcher may use lower or two-sided CUSUMs computed using the ability estimate from the unproctored test and using the scores and item parameters from the confirmation test. Control limits can be estimated so that a decision rule can be created without too many false positives. We discuss this further in the ''Method'' section.

In this study, we focus on the lower CUSUM statistics C^{-} and C^{L}, since these are the most sensible statistics to detect deterioration of performance from the first test to the second test, which is our primary goal. We also used the two-sided C^{LR} statistic, because Tendeiro and Meijer (2012) found that this statistic performed well in different test situations.

Method

Data

The Connector Ability (CA; Maij-de Meij, Schakel, Smid, Verstappen, & Jaganjac, 2008) is a CAT for the measurement of a person's cognitive ability and is used in selection settings. The test consists of three subtests: series of figures, series of matrices, and series of numbers (in this order). A G-factor θ estimate is computed as

a weighted combination of the θ values estimated from the three subtests. There are two steps in the selection process using the CA. The first step consists of administrating the test through the Internet. This test is unproctored, and candidates are free to choose the location where they want to do the test. The second step consists of administrating the test in a supervised environment provided by the organization. The aim of this proctored test is to confirm the results of the unproctored test.

The CA is used by organizations for selecting candidates for positions requiring a master's (MA), bachelor's (BA), or upper vocational educational level. Item calibration was based on administration of items in pilot tests in unproctored settings. For each item, about 300 responses were obtained to estimate the item parameters using the 2PL model (Birnbaum, 1968). The CA was administered under proctored conditions in real selection settings to compute the norms for the three educational levels MA, BA, or upper vocational level. Normalized values for ability estimates were determined for each educational level on the basis of these norms. The candidates invited for the proctored test were the ones who scored at or above the normalized cutoff score. The testing organization could still invite candidates for the second test who did not pass the cutoff score in the unproctored test, if desired.

Data were obtained from candidates of 14 different organizations, between January 2010 and June 2011. The total sample consisted of 850 candidates (67% male, 28% female, 5% unknown) applying for a job requiring an MA educational level (82%), a BA educational level (14%), and an upper vocational educational level (4%). The majority (52%) were between 18 and 29 years old. Forty-eight percent of the candidates obtained an MA, 25% a BA, and 6% had a midlevel educational background. The sample contained 69% autochton candidates, 6% Western minorities, 11% non-Western minorities, and 14% had an unknown ethnic background. All candidates in the data set were invited for a confirmation test.

The number of items administered in the unproctored test was between 30 and 45 (mean = 37.0, SD = 5.1). The confirmation test was typically shorter than the unproctored test. Candidates who were administered the CA in 2010 (half of the sample) were given a confirmation test consisting of 15 items. An update of the procedure led to an increase from 15 to 21 items in the confirmation test for all candidates in 2011. Thus, candidates who were administered the CA in 2011 (exactly half of the candidates in our data set) were given a longer confirmation test than those who went through the procedure in 2010.

Abilities were estimated using maximum likelihood estimation for the 2PL model. Two separate item pools were used for the unproctored and proctored tests. The item pool for the unproctored test was larger than that for the proctored test. Also, only highly discriminating items were used for the confirmation tests. Table 1 summarizes some characteristics of the item pools. Note that no specific methodology was implemented to control item exposure rates (ratio of the number of times an item is administered to the number of candidates). In practice, only 21% and 67% of the item pools for the unproctored and the confirmation tests were administered, respectively. The low percentage of items used from the unproctored item pool may be

Table 1. Descriptives of the Item Pools

	Item pool unproctored test			Item pool proctored test		
	Figures	Matrices	Numbers	Figures	Matrices	Numbers
Number of items	306	248	247	54	49	50
Discrimination	1.26 (0.44)	1.26 (0.42)	1.01 (0.31)	2.31 (0.65)	2.10 (0.63)	1.88 (0.52)
Difficulty	−0.67 (1.09)	−0.53 (0.74)	−0.36 (0.84)	−0.48 (1.06)	−0.37 (0.99)	−0.47 (1.00)

Note. Values under "Discrimination" and "Difficulty" are the mean and standard deviation values for the discrimination and difficulty parameters of the items in each subtest (series of figures, matrices, and numbers).

Table 2. Proportions of Item Exposure Rates in the Item Pools of the Unproctored and Proctored Tests

Item exposure rates	Unproctored test	Proctored test
Never used	.79	.33
0% to 5%	.09	.32
5% to 10%	.02	.07
10% to 20%	.03	.13
>20%	.07	.15

problematic with respect to test security. However, there was no indication that items became known. The average item exposure rates were .16 and .13 for the unproctored and proctored tests, respectively, which seem to be acceptable rates (see Impara & Foster, 2006). Table 2 summarizes statistics regarding item exposure rates. In practice, the percentage of items administered from the item pool for the unproctored test was larger than 21% if we also consider the items used by candidates who were not selected for the confirmation test (38% of the item pool).

For the CA used in 2010, the ability parameter was set at zero at the beginning of each subtest, for both the unproctored and the confirmation tests. The updated 2011 version of the CA set the ability parameter at the beginning of each subtest at prespecified normed values according to the background of the candidate: $\theta = 0.5$, 0, and -0.5 for candidates with an MA, BA, or upper vocational educational level, respectively.

CUSUMs

We used C^{-}, C^{L}, and C^{LR} as CUSUM statistics. For each candidate a CUSUM sequence was determined using the item parameters and item scores from the confirmation test, and using the ability estimate from the first test. Probabilities p_i^L and p_i^U were estimated using the algorithm presented in Armstrong and Shi (2009) with some adjustments discussed in Tendeiro and Meijer (2012).

Next, control limits were estimated for each CUSUM statistic and for each candidate. We used bootstrapping (Efron & Tibshirani, 1993) to estimate the control limits. Given a candidate with ability estimate $\hat{\theta}_{\text{Un}}$, we selected all candidates (S) from the data set such that

$$|\hat{\theta}_{\text{Un}} - \hat{\theta}_{\text{Un}}^{(S)}| < k \cdot SE_{\text{Un}}. \quad (5)$$

Constant k was initially set equal to 0.5, and it was increased in steps of 0.25 until a sample of at least 10 candidates was collected. Only 24 candidates (2.8%) required a value larger than $k = 0.5$. The sample sizes for these cases ranged from 10 through

16. For all the other candidates (97.2%) the sample sizes were relatively larger (mean = 91.3, SD = 41.1).

For each sample we computed bootstrap distributions for the 1% and 5% control limits of each CUSUM statistic (number of resamples = 1,000). Our final estimates were computed as the medians of the corresponding bootstrap distributions. The median was used because we observed that the bootstrap distributions were often nonsymmetric and/or multimodal. The control limits were then used to classify the answering behavior of each candidate as inconsistent or normal, for each CUSUM statistic.

Alternative Methods

The l_z statistic (Drasgow, Levine, & Williams, 1985) is a standardization of the log likelihood function,

$$l_0 = \sum_{i=1}^{n} [X_i \ln p_i(\theta) + (1 - X_i) \ln (1 - p_i(\theta))], \tag{6}$$

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{V(l_0)}}. \tag{7}$$

We computed the l_z statistic for each candidate in the data set using the ability estimate from the first test, and the item parameters and item scores from the confirmation test. The distribution of l_z is standard normal only when true θ values are used (Molenaar & Hoijtink, 1990). We used maximum likelihood estimates of θ, which has the effect of changing the variance of l_z more than expected under the standard normal distribution (see Meijer & Sijtsma, 2001, for a discussion). Therefore, we used a bootstrapping procedure to estimate the 1% and 5% control limits for the l_z statistic similar to the procedure used for the CUSUMs.

The z statistic (Guo & Drasgow, 2010) is specifically designed for detecting misfits between unproctored and proctored tests. The formula for z is given by

$$z = \frac{\hat{\theta}_{\text{Un}} - \hat{\theta}_{\text{Pr}}}{\sqrt{SE^2_{\text{Un}} + SE^2_{\text{Pr}}}}. \tag{8}$$

z represents the difference between the estimated abilities from both tests, divided by a pooled standard error from both θ estimates. The z statistic is asymptotically standard normal when the θs are estimated using maximum likelihood and the IRT property of local independence assumption holds (Guo & Drasgow, 2010). However, the z scores in our data, albeit normally distributed, differed from the expected standard normal distribution. This problem might be related to the length of the tests used. Therefore, we also used bootstrap distributions to estimate the 1% and 5% control limits for the z statistic.

Our programs were written and executed in R (R Development Core Team, 2009).

Table 3. Detection Rates for Each Person-Fit Statistic

Statistic	1% Control limit	5% Control limit
C^{-}	2.9	6.9
C^{L}	2.4	6.0
C^{LR}	2.7	6.2
l_z	2.5	6.4
z	2.7	6.5

Note. Values are in percentages.

Table 4. Similarity Between Methods: Values for the Phi Coefficient

	C^{L}	C^{LR}	l_z	z Scores
C^{-}	.55	.35	.48	.55
C^{L}	—	.71	.69	.64
C^{LR}		—	.75	.49
l_z			—	.66

Results

CUSUMs, l_z, z

The detection rates, that is, the proportions of candidates whose answering patterns were flagged as inconsistent by each statistic, are summarized in Table 3. All results that we present next concern the detection rates using the 5% control limits.

For a given person-fit statistic, each candidate was coded 1 in case inconsistent behavior was detected and 0 otherwise. Thus, a binary vector of length $n = 850$ was available for each statistic. Table 4 summarizes the phi correlations between any pair of binary vectors. The correlations are between .35 and .75. The similarity between the likelihood-based statistics C^{L}, C^{LR}, and l_z was highest (correlations between .69 and .75). It is important to realize that different person-fit statistics detect different types of inconsistent response patterns. Hence, variation in the magnitude of the correlations between the statistics reflects the ability of different statistics to detect different types of inconsistencies. The researcher can use combined information from several person-fit statistics to decide which candidates should be flagged as inconsistent. In our analysis, we used three CUSUM person-fit statistics (C^{-}, C^{L}, and C^{LR}), the l_z, and the z statistics. The percentage of score patterns flagged by all five procedures was 2.0% (17 candidates). Other possibilities consist of looking at the score patterns flagged by all three CUSUMs (2.6%, i.e., 22 candidates), by all CUSUMs and the l_z (2.5%, i.e., 21 candidates), by all CUSUMs and z (2.0%, i.e., 17 candidates), or by at least one CUSUM plus both l_z and z (4.0%, i.e., 34 candidates).

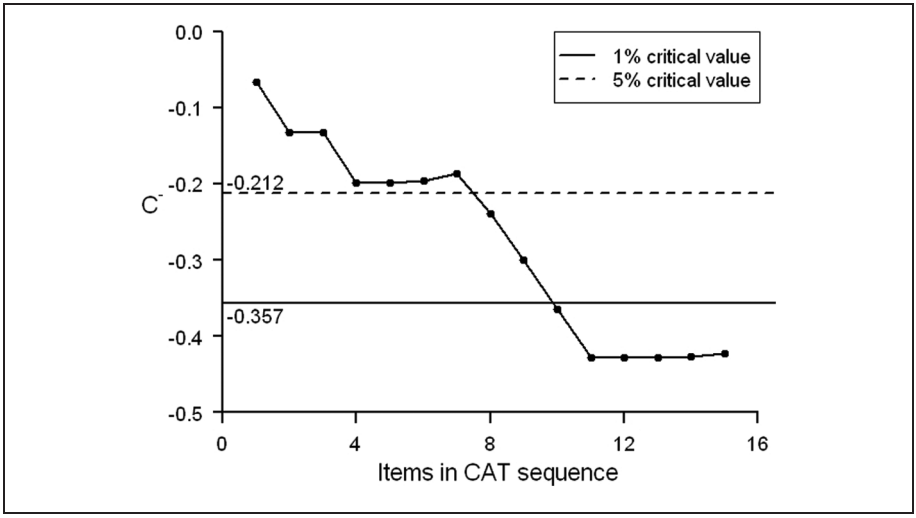

Figure 1. CUSUM chart for C^{-} statistic: Candidate 110

CUSUM Charts

Besides flagging inconsistent item score patterns, CUSUMs allow access to local information concerning item scores. It is, for example, possible to investigate whether there are items or strings of items on the confirmation test where the candidate performed better or worse than expected. We illustrate the use of CUSUM charts in Figures 1 through 4. These CUSUM charts display the variation of the values of the C^{-} statistic for candidates 110, 192, 563, and 577. Each chart represents the 1% and 5% control limits through horizontal lines; a response pattern is flagged as aberrant whenever the CUSUM series crossed a control limit. Candidates 110, 192, and 577 were flagged as inconsistent by the five person-fit statistics used in our analysis, whereas Candidate 563 was not flagged by any of the statistics.

Candidate 110 in Figure 1 was flagged by C^{-} as inconsistent at a 1% level. This candidate obtained an estimated ability from the unproctored test equal to $\hat{\theta}_{Un} = 1.54$. This estimate is high considering the easy items from the item pool of the confirmation test, see Table 1. This has practical consequences when computing the CUSUM sequences: Answering an item incorrectly results in a relatively large decrease in the CUSUM, whereas answering an item correctly increases the CUSUM only slightly, since the expected probability of answering the item correctly is large for each item. Candidate 110 failed three of the first four items. After the fourth item this candidate answered three items correctly, then four items incorrectly, and finally four items correctly. Some of the items answered incorrectly were easy, given the estimated ability $\hat{\theta}_{Un}$ from the unproctored test. For example, Items 8, 9, 10, and 11 with estimated difficulty $\hat{b} = 0.97$, 0.47, 0.29, and 0.00, respectively, were incorrectly answered. This

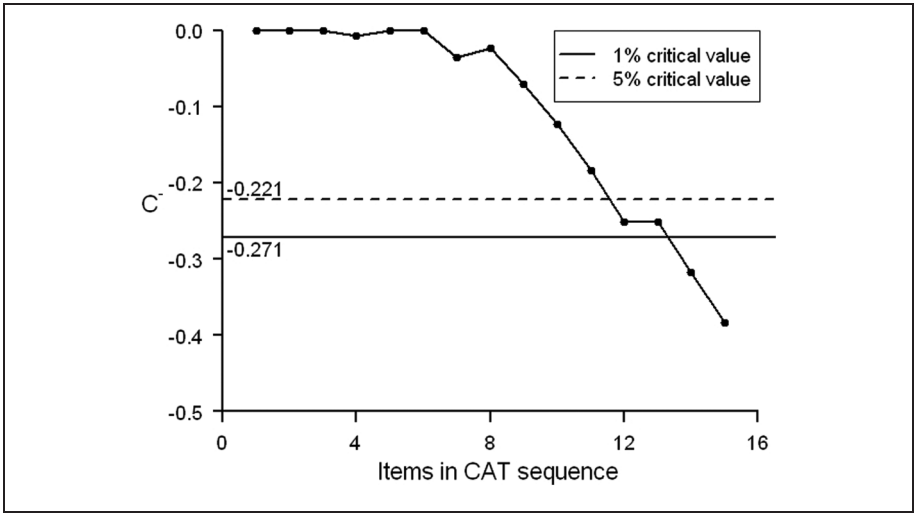

Figure 2. CUSUM chart for C^- statistic: Candidate 192

resulted in a CUSUM chart below the 1% control limit. Thus, the item score pattern of Candidate 110 was flagged as inconsistent because too many easy items (according to the ability estimate from the unproctored test) were answered incorrectly.

Figure 2 provides another example of a CUSUM that reflects inconsistent response behavior (at 1% significance level). The estimated ability from the unproctored test equaled $\hat{\theta}_{Un} = 0.84$. This candidate performed well at the beginning of the test with only two incorrect scores on the first eight items. After the eighth administered item the candidate's performance deteriorated, and only one item of seven administered items was answered correctly. The bad performance in the second half of the test resulted in an inconsistent CUSUM chart. Observe that the increases and decreases in the CUSUM chart depend on the item difficulty. For example, the fourth item was answered incorrectly but the effect on the CUSUM was small because this was a difficult item ($b = 2.35$). In contrast, missing the last two items had a larger effect on the CUSUM because these were relatively easy items (difficulties equal to -1.00 and -1.24, respectively). Candidate 192 seems to have performed well only in the first half of the test, which may be interpreted as a result of plodding behavior (Meijer, 1996).

The CUSUM of Candidate 577 ($\hat{\theta}_{Un} = 1.03$) displayed in Figure 3 resulted from the updated version of the CA; the confirmation test had 21 items in total. This candidate's score pattern consisted of an alternation of correct and incorrect answers. Items answered incorrectly were heavily penalized by the CUSUM because the estimated ability $\hat{\theta}_{Un}$ was relatively high. At the end of the test, the 1% control limit was crossed, reflecting an accumulation of evidence favoring the hypothesis of an inconsistent answering pattern.

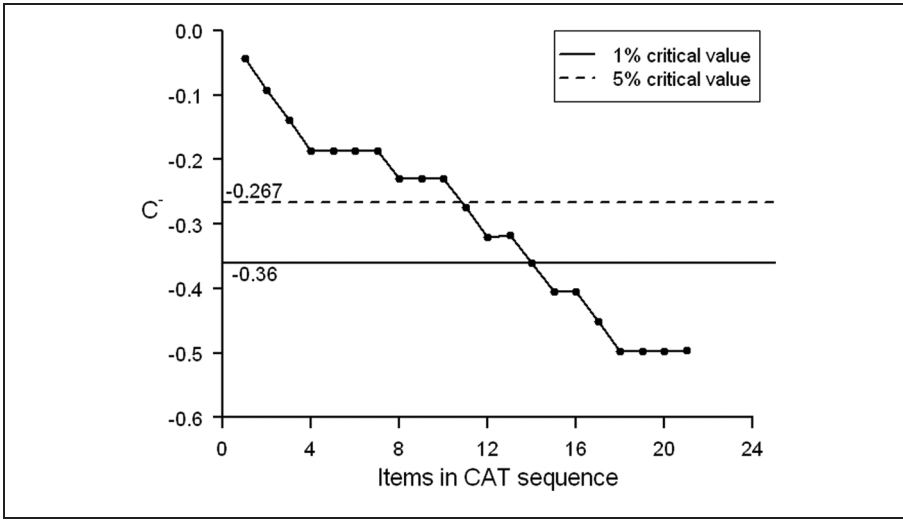

Figure 3. CUSUM chart for C^- statistic: Candidate 577

The score pattern of Candidate 563 ($\hat{\theta}_{Un} = 0.81$) displayed in Figure 4 was not flagged as inconsistent by the C^- CUSUM statistic. This candidate was also tested with the updated version of the CA. The response vector consisted of an alternation of correct and incorrect answers throughout the test. Some of the items answered incorrectly can be considered difficult, and as a consequence the CUSUM decreased moderately for these items (e.g., consider Items 3, 5, and 17 with estimated difficulty $\hat{b} = 2.35$, 1.66, and 1.31, respectively). Answering difficult items correctly, on the other hand, resulted in a pronounced increase of the CUSUM, as can be seen for Items 15 and 16 (difficulty $\hat{b} = 0.69$ and 1.60, respectively). The CUSUM did not flag the response vector as inconsistent because Candidate 563 did not fail, in general, on many items with difficulty well below the estimated ability $\hat{\theta}_{Un}$.

Discussion

To identify cheating in UIT, Guo and Drasgow (2010) proposed to use a statistic in which global measures (estimated abilities) on unproctored and proctored tests are compared. Unusual response behavior is identified when scores are larger than some cutoff threshold. The aim of the present article was to discuss other psychometric methods that can be used to identify test-taker cheating in a test–retest context. In the psychometric literature there has been a proliferation of statistics that are aimed at identifying cheating inconsistent response behavior (Meijer & Sijtsma, 2001), but there are not many empirical studies that show their usefulness. Recently, a number of statistics were proposed to detect cheating (e.g., Belov, 2011; Sotaridona &

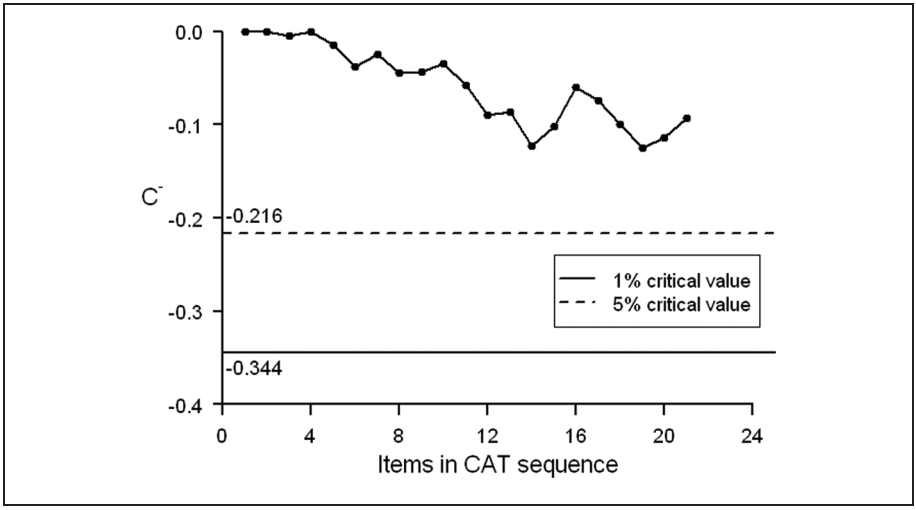

Figure 4. CUSUM chart for C^{-} statistic: Candidate 563

Meijer, 2003), however, these statistics mainly aim at detecting persons who copy item scores from their neighbors.

A different type of technique which allows to detect inconsistent answering behavior in UIT is based on the sequential probability ratio test (SPRT; Eggen, 1999; Eggen & Straetmans, 2000; Finkelman, 2008; Makransky & Glas, 2011; Wald, 1947). SPRT-based models, or more specifically their adaptations to confirmation testing (Makransky & Glas, 2011), allow to dynamically build the proctored test in a way that minimizes the number of items required until a decision can be made. We did not use SPRT models in our analysis because that would have required reformulating the CA adaptive testing procedure before collecting the data, which was not feasible.

An alternative to the cumulative sum strategy used in this article consists of estimating, for each candidate, two posterior distributions of ability using data from the unproctored and proctored tests. Both posteriors are then compared using the Kullback–Leibler divergence (KLD; see Belov, Pashley, Lewis, & Armstrong, 2007; Belov & Armstrong, 2010). Large values of the KLD indicate a significant change in performance between both tests. Critical values for the KLD at fixed levels of significance can be estimated using either simulation, approximating distributions such as the lognormal (Belov & Armstrong, 2010), or theoretical distributions which the KLD follows under specific conditions (Belov & Armstrong, 2011). As observed by an anonymous reviewer, the KLD approach takes into account all available information from the posterior distributions, unlike other statistics which rely only on the first moments of the posteriors (e.g., Guo and Drasgow's z statistic). We observe that the CUSUM technique is of a different nature than the KLD. CUSUMs are

sequential procedures which take into account the order in which the items are presented to each candidate. The KLD technique estimates posteriors from two sets of items (in our setting, the unproctored and proctored tests), but the order of the items within each set is not taken into account. In particular, psychometric information in the shape of CUSUM charts is not readily available for the KLD. Thus, the CUSUM and the KLD approaches can be regarded as two alternative ways for detecting aberrant response behavior. CUSUMs are specially suitable for situations where it is important to take into account a specific ordering of the items (e.g., the administration order).

Guo, Tay, and Drasgow (2009) showed, using simulated data, that CATs can protect against cheating when the frequency of test administration was held constant. CAT systems were better at resisting small-scale cheating than conventional test systems. For the conventional tests, lengthier tests and more test forms enhanced test security. Although there is not much evidence yet, the first empirical studies (Lievens & Burke, 2010; Nye, Do, Drasgow, & Fine, 2008) showed that mean scores on the proctored tests are higher than the scores on the unproctored tests, indicating that, in general, test takers do not cheat. The existing studies that report these results also recognize that there are test takers with large decreases in test scores. Furthermore, these studies used dichotomous decisions on whether or not people cheat, although this may be more subtle. As an alternative a researcher can calculate the likelihood of each pattern based on the estimation of the first unproctored trait, as we showed in this study.

How should we use different sources of psychometric information with respect to cheating? An interesting study, although not in employment setting, that combined different types of statistical evidence to detect cheating was presented in Jacob and Levitt (2003). In that study teacher cheating was identified by making use of suspicious test score fluctuations and misfitting item score patterns. To improve student achievement, several states and districts in the United States have recently implemented programs that use student scores to punish or reward schools. States are required to test elementary students each year, rate schools on the basis of test performance, and intervene in schools that do not make sufficient improvement. Although these policies may be effective, one of the drawbacks is that teachers and administrators may cheat. Teachers may change the responses on answer sheets, providing correct answers to students. Documented cases of this kind of cheating have been described by Kolker (1999) and Hofkins (1995). Jacob and Levitt (2003) were in particular interested in the identification of unlikely answer strings. As a first measure they used counts concerning the number of identical answers within a class room. As a second measure they used the variance of student responses on an item within a classroom. A third measure used the degree of correlations across questions within a classroom and a fourth measure focused on how much a student's response pattern differed from the other response patterns in the classroom.

In our opinion, the Jacob and Levitt (2003) study is an excellent example of how information from inconsistent response pattern combined with other (statistical)

information can be of great help to identify suspicious test scores of (groups of) individuals. In general, we think that in psychological and educational computer-based testing it is relatively easy to routinely check differences in test scores and combine this information with, for example, information about unexpected item score patterns. Additional context-specific information will probably always be needed to ''have a case,'' but the routine to check scores and item score patterns and, perhaps, communicate with candidates that this is being done may help obtain valid scores.

In this article, we presented strategies based on cumulative sum statistics that can be used to detect inconsistent item score patterns in test–retest contexts. We exemplified CUSUM techniques by means of empirical data. Comparisons with alternative methods were considered. We discussed several advantages associated to CUSUM techniques. One important advantage is the fact that CUSUMs allow interpreting individual item scores or sequences of item scores, which contrasts with person-fit techniques that are based on total score measures. Another advantage of using CUSUMs is the easiness and richness that CUSUM charts add to the analysis of the results.

In general we believe that CUSUM procedures can provide an additional view on person-fit in IRT. A future step in this field could be to consider implementing CUSUMs in real-life settings such as personnel recruitment, together with other well-known psychometric techniques. Integrating several person-fit methods together can benefit interpretation of results and can better substantiate the final outcomes of the analysis.

Acknowledgments

The authors thank two anonymous reviewers and the editor for their help with improving this article.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*, 391-410.

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement, 35*, 495-517.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement, 34*, 379-392.

Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback-Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64*, 291-309.

Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback-Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7-14). Tokyo, Japan: Universal Academy Press.

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. 397-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910-919.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125-144.

Burke, E. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 35-38.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC Press.

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.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

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

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18*, 351-364.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing, 9*, 283-309.

Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement*. New York, NY: Springer-Verlag.

Hofkins, D. (1995, June 16). Cheating ''rife'' in national tests. *Times Educational Supplement*, p. 1.

Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91-114). Mahwah, NJ: Lawrence Erlbaum.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher teaching. *Quarterly Journal of Economics, 118*, 843-877.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Kolker, C. (1999, April 14). Texas offers hard lessons on school accountability. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1999/apr/14/news/mn-27348

Lievens, F., & Burke, E. (2010). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84*, 817-824.

Maij-de Meij, A. M., Schakel, L., Smid, N., Verstappen, N., & Jaganjac, A. (2008). *Connector Ability 1.1, professional manual*. Utrecht, Netherlands: PiCompany.

Makransky, G., & Glas, C. A. W. (2011). Unproctored internet test verification: Using adaptive confirmation testing. *Organizational Research Methods, 14*, 608-630.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3-8.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2010). Detecting person misfit in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 315-329). New York, NY: Springer.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75-106.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150-162.

Neyman, J., & Pearson, E. S. (1933). On the problems of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289-337.

Nye, C. D., Do, B.-R., 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.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100-115.

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

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53-69.

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36,* 420-442.

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.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*, 189-225.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detection of 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.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.