Detecting cheating in unproctored internet testing settings using CUSUMs (and more)

Jorge Tendeiro

University of Groningen

03 July 2012 / 8th ITC Conference

Outline

Detecting aberrant behavior

Main idea

Carefully, and routinely, evaluating the veracity of information obtained from tests, interviews,...

Is this really needed?

Sure! People do exaggerate, hide, make up, fake or even lie in tests and interviews.

What is 'aberrant behavior'?

Any type of behavior whose main purpose is distorting the assessment of a specific ability or trait.

Detecting aberrant behavior — is it really possible?

Are there definite ways of detecting specific types of aberrant behavior (e.g. cheating)?
NO (i.e., not using psychometric tools alone).

E.g., overperformance $\not\Rightarrow$ cheating for sure.
Other things could have happened:

- Luck, intense study, preknowledge of items ($\neq$ cheating);
- Too easy items;
- Scores were tampered by the teacher (e.g. Jacob and Levitt, 2003).

What we propose to do

- We try to identify misfits between scores on tests and true trait.
- We say nothing (or very little. . .) about how to interpret misfits (e.g., if the subject 'cheated' or 'was lucky').

Statistical Process Control (SPC)

Original idea

Supervise industrial production processes.

Features

- Assessing quality of production in (nearly) real-time: continuous *versus* final control.
- Allowing early interventions in the process once a malfunction is detected.
- Using charts to display results.
- Accessible interpretation of results for nonexperts.

Our focus within SPC

CUSUM charts: CUmulative SUM control charts (Page, 1954)

CUSUMs in IRT

CUSUMs in IRT? For person-fit purposes?
Yes.

How?
Think about CATs (Computerized Adaptive Testing)

- CATs: sequential and adaptive procedures.
- Regard CATs as 'industrial processes' to be monitored.
- Here,

  'out of control'

  means

  'misfit between item scores
  and
  IRT parameters (ability and item parameters)'.

CUSUMs in IRT — Literature

Bradlow, Weiss and Cho (1998)
Checking, after step i, whether the standardized absolute deviation of the number-correct score is unusually large.

Van Krimpen-Stoop and Meijer (2000)
Introduced upper and lower CUSUM statistics:

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

with $C_i^+ = C_i^- = 0$ and $T_i = f(X_i - p_i)$.

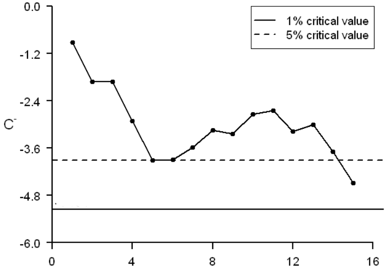
- C_i^- to detect underperformances, C_i^+ to detect overperformances
- Critical values $L_i(\alpha)$, $U_i(\alpha)$ need to be estimated;
  subject is flagged as aberrant if $C_i^- \leq L_i(\alpha)$, $C_i^+ \geq U_i(\alpha)$ for some i.

CUSUMs in IRT — Literature

Van Krimpen-Stoop and Meijer (2000) — example

From our empirical dataset; $\hat{\theta} = 0.84$.

i	X_i	a_i	b_i	p_i	$X_i - p_i$	C_i^-
1	0	2.60	-0.12	0.92	-0.92	-0.92
2	0	3.97	-0.70	1.00	-1.00	-1.92
3	1	3.83	-1.44	1.00	0.00	-1.92
4	0	2.81	-1.00	0.99	-0.99	-2.91
5	0	2.80	-1.24	1.00	-1.00	-3.91
6	1	3.10	-0.33	0.97	0.03	-3.88
...	...	...	...	...	...	...



CUSUMs in IRT — Literature

Armstrong and Shi (2009)

- CUSUM updates are estimated using logs of likelihood ratios.
  (exploring idea in Neyman and Pearson, 1933)
- Models for the probability of a correct response under sought aberrant behavior (p^L, p^U) are required.

Lower CUSUM	Upper CUSUM
$C_i^L = \min\left\{0, \gamma_i^L + C_{i-1}^L\right\}$	$C_i^U = \max\left\{0, \gamma_i^U + C_{i-1}^U\right\}$
$\gamma_i^L = \ln \frac{p_i^{x_i}(1-p_i)^{1-x_i}}{(p_i^L)^{x_i}(1-p_i^L)^{1-x_i}}$	$\gamma_i^U = \ln \frac{(p_i^U)^{x_i}(1-p_i^U)^{1-x_i}}{p_i^{x_i}(1-p_i)^{1-x_i}}$

- Upper and lower critical values must be estimated.
- Tendeiro and Meijer (2011) discuss improvements for this method.

Unproctored internet testing (UIT)

What is it?

- UIT: testing procedure under which tests are given to examinees via the web
- They can, in theory, be solved anywhere, 24/7
- Convenient for both parts: it saves time & money
- UITs are wide-spreading (Tippins, 2009; Pearlman, 2009)

Problems inherent to UIT (e.g., Nye *et al.*, 2008)

- Access to internet/up-to-date PC/WWW (in)experience bias
- Test security; reliability (unstandardized testing environment)
- Examinee identification
- Cheating (validity issue), e.g.:
  - using surrogate
  - accessing non-allowed sources (books, websites)

Unproctored internet testing (UIT)

How to avoid these problems?

The most common way, proposed by the International Testing Commission, consists in using a second test:
confirmation/ verification/ proctored test

About the confirmation test

- Taken in a secured, supervised environment.
- Uses all, or only best, candidates from the UIT.
- Made as small as possible (strive for efficiency).
- Main purpose: confirm/reject the results of the UIT, not to replace UIT scores (Lievens & Burke, 2010, defend differently).

Our data

About the applicants

- 850 applicants (67% male, 28% female, 5% unknown)
- Context — applying for jobs requiring
  - MA educational level (82%)
  - BA educational level (14%)
  - Upper Vocational educational level (4%)
- Age: 52% $18 \leq \cdot \leq 29$, 48% ≥ 30
- 69% autochthon applicants, 6% western-minorities, 11% non-western minorities, and 14% unknown ethnic background

Our data

About the tests

- The Connector Ability (Maij-de Meij et al., 2008) CAT procedure was used.
- It consists of three parts: series of numbers, figures, and matrices.
- Designed to measure cognitive abilities (easiness and speed when tackling problems).
- A general intelligence factor is estimated as a weighted combination of the ability estimates of the three subtests.
- Administered in two stages
  - first administration: UIT
    # items between 30 and 45 (mean= 37.0, SD= 5.1);
  - second administration: proctored
    # items 15 (50%) or 21 (50%).

Our data

About the IRT model

- 2PL model used (Birnbaum, 1968).
- Abilities estimated using MLE method.

About the item pools

- Two separate item pools were used.
- First pool larger than second pool.
- Discrimination larger in second pool.

Main question of interest

"Which examinees suffered a notorious decrease in performance from the first (unproctored) to the second (proctored) test?"

Our methodology

Implementing CUSUMs

- Statistics used: C_i^-, C_i^L and C^{LR}.
- We used:
  - item parameters (a, b) and item scores from confirmation test;
  - $\hat{\theta}_{\text{Un}}$ from first test.
- p_i^L and p_i^U (for C_i^L, C^{LR}) estimated as in Armstrong and Shi (2009), with adjustments as in Tendeiro and Meijer (2011).
- Control limits were estimated per CUSUM statistic per examinee.

l_z statistic (Drasgow, Levine and Williams, 1985)

- Standardized logarithm of the likelihood function evaluated at the MLE of θ (Levine and Rubin, 1979).

z statistic (Guo and Drasgow, 2010)

- Standardized difference between abilities estimated from both tests.

Some results

Detection rates

Statistic	5% control limit
C^-	6.9
C^L	6.0
C^{LR}	6.2
l_z	6.4
z	6.5

Similarity between methods

	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

Some results

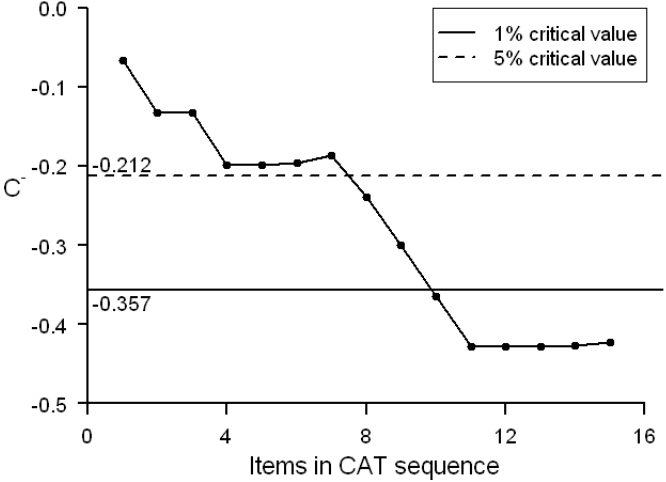
Final shortlist of aberrant examinees

There is no written-in-stone kind of rule...

Flagged by:	
All statistics	17 (2.0%)
All CUSUMs	22 (2.6%)
All CUSUMs $\oplus$ l_z	21 (2.5%)
All CUSUMs $\oplus$ z	17 (2.0%)
At least one CUSUM $\oplus$ l_z, z	34 (4.0%)

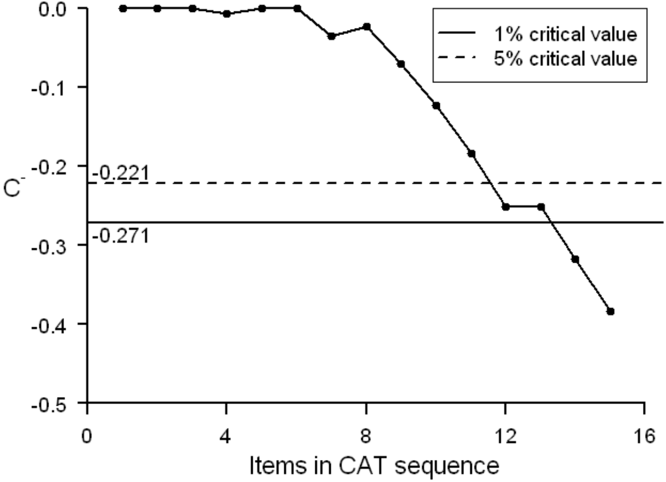
Some results — CUSUM charts

Examinee # 110 ($\hat{\theta}_{\text{Un}} = 1.54$): too many wrong easy items…



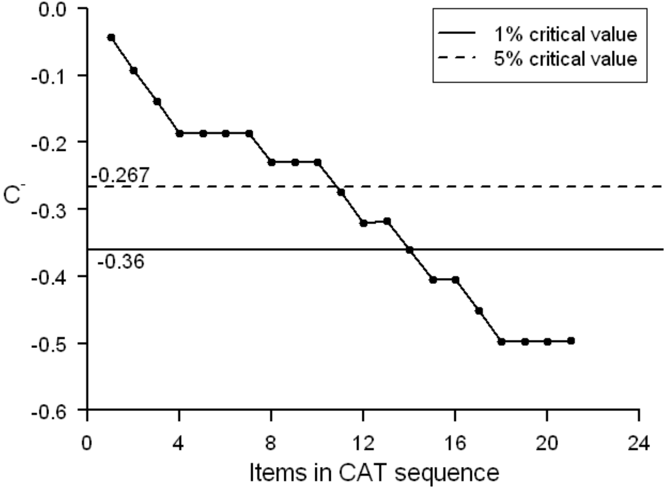
Some results — CUSUM charts

Examinee #192 ($\hat{\theta}_{\text{Un}} = 0.84$): starts well, but then…



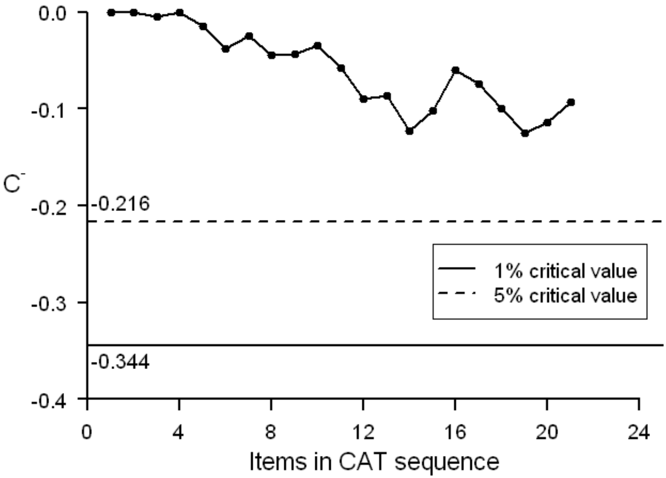
Some results — CUSUM charts

Examinee # 577 ($\hat{\theta}_{\text{Un}} = 1.03$): alternating 1's and 0's...



Some results — CUSUM charts

Examinee # 563 ($\hat{\theta}_{\text{Un}} = 0.81$): normal behavior…



Discussion

CUSUMs

- CUSUMs were applied in the setting of confirmation tests following UIT.
- Bootstrapping was used to estimate the control limits.
- Interpretation of CUSUM charts is very rich.