Compromised Item Detection Using Item Response and Response Time

Chunyan Liu, Dan Jurich, & Kimberly Swygert

10/12/2018

NBME®

Introduction

Test security

- Item breach
- Item preknowledge
- Decrease in item difficulty
- Increase in examinee performance

Background

Response time (*RT*)

- Examination of test taker's motivation (Wise, 2006; Wise & Kong, 2005)
- Test form construction (van der Linden, 2011)
- Examination of test speededness (Shao, Li, & Cheng, 2016)
- Detection of item preknowledge (Meijer, & Sotaridona, 2006; Qian et.al, 2016; van der Linden & Guo 2008)

Background

Compromised Item Detection

- Sequential procedure (Zhang, 2013)
    - Computerized Adaptive Testing (CAT)
    - Change-point (n_c)
    - Item becomes easier at the changing-point

Background

Compromised Item Detection

- Sequential procedure (Zhang, 2013)
  - *moving sample*: most recent responses to an item up to *n*
  - *m*: size of the *moving sample*
  - n_c: changing point

Background

Sequential procedure (Zhang, 2013)

$$\hat{Z}_{nm} = \frac{\hat{p}_{nm} - \hat{p}_{n-m}}{\sqrt{\hat{p}_{n-m}(1 - \hat{p}_{n-m})}} \sqrt{\frac{m(n-m)}{n}}.$$

n: sequence number of the present examinee

m: moving sample size

$\hat{p}_{nm}$: item p-value of the moving sample at *n*

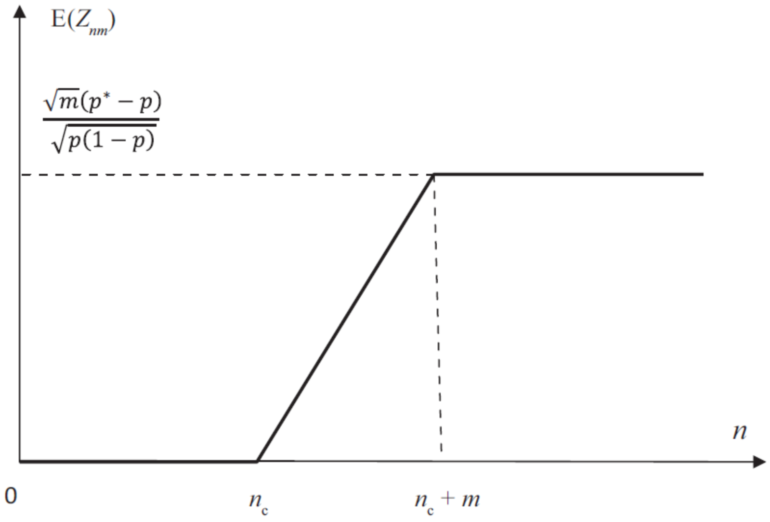
$\hat{p}_{n-m}$: item p-value of the first *n-m* responses

Note: $\hat{Z}_{nm}$ does not follow a normal distribution and a cutoff point (c_α) is used to flag items

Background

Sequential procedure (Zhang, 2013)

$$E[Z_{nm}] = \begin{cases} 0, & \text{if } m \le n \le n_c; \\ (n - n_c)(p^* - p)/\sqrt{mp(1-p)}, & \text{if } n_c < n < n_c + m; \\ \sqrt{m}(p^* - p)/\sqrt{p(1-p)}, & \text{if } n \ge n_c + m. \end{cases}$$



Background

Sequential procedure (Zhang, 2013)

- Applied in CAT Simulation
- Hasn’t been applied to operational data in continuously administered linear computer-based testing (CBT)
- Didn’t consider *RT*

Purpose of the Study

- ➢ Flag compromised items using the sequential procedure
  - o For operational data from a linear CBT
  - o For data from different countries
  - o Considering both item responses and item *RTs*
    - ✓ RT: change of average item latency of the moving sample

- ➢ Average examinee ability varies during the testing window (seasonal effect)

Method

Data

- Medical licensure examination in English
- Multiple test forms administered in a year
- Thousands of items
- > 35,000 test takers
- Four investigated countries (US, A, B, C)
- Seasonal effect across the year

Method

Sequential Procedure

- For US
  - Starting point n_0 = 500
  - m = 50
- For non-US
  - Starting point n_0 = 50
  - m = 25
- Cutoff point: c_α = 3.5 and 2.0

Method

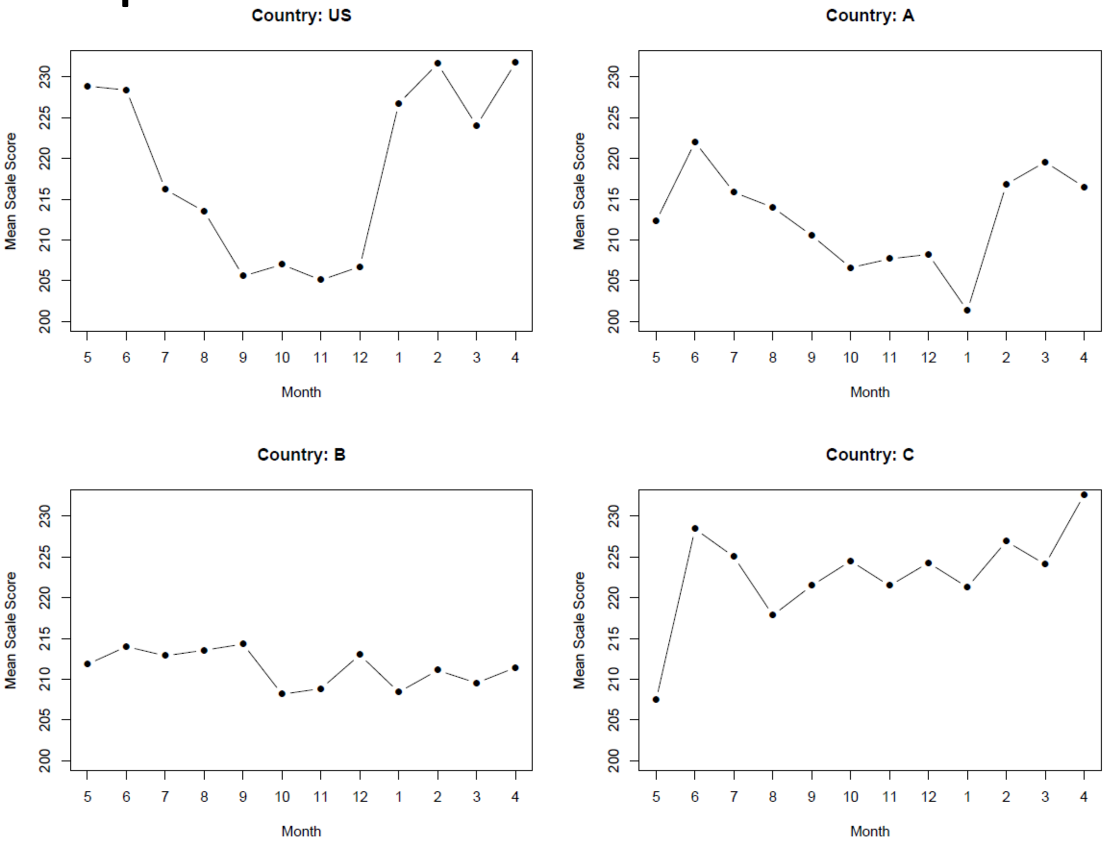
Assumptions

- Examinees’ test speed and examinees’ ability are not highly correlated

- Item response time decreases after it is breached

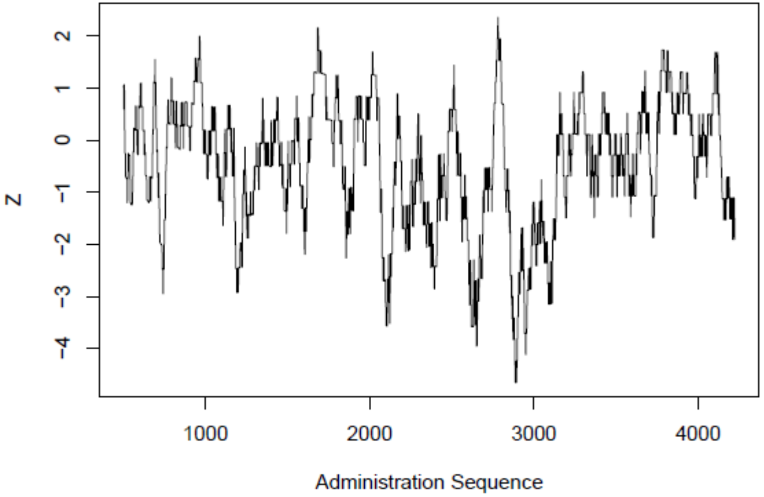
Data

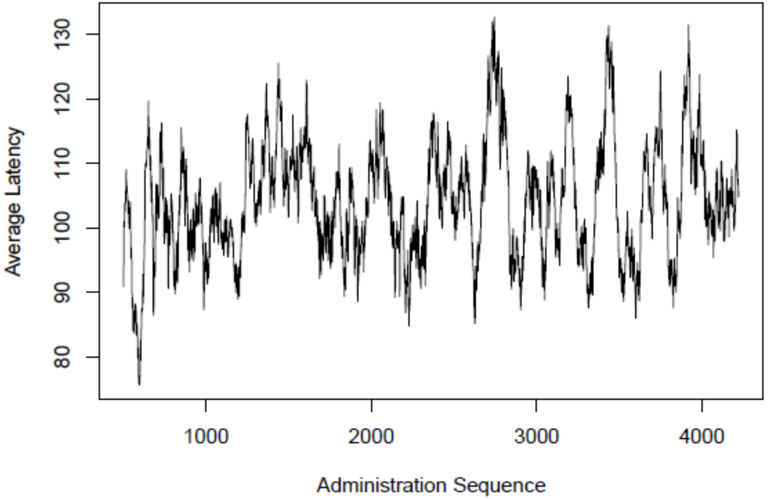
Examinee performance over time



Results

Z and average latency for an unflagged item(c_{α} = 3.5)





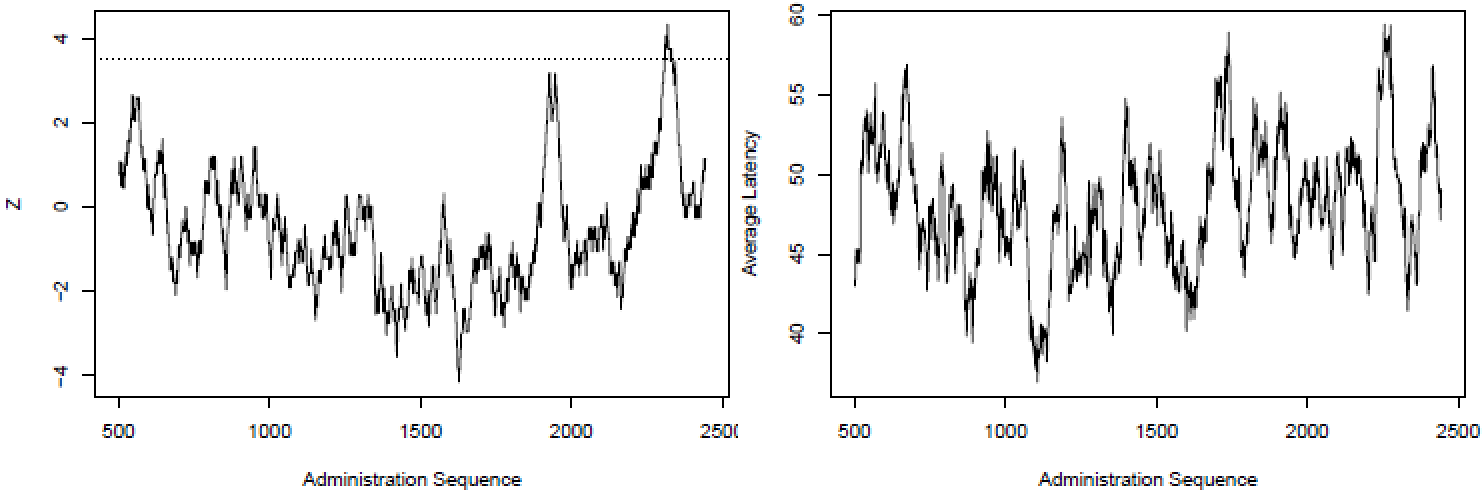
Results

Number of flagged items based on item responses only (c_{α} = 3.5)

Country	US	A	B	C
N	92	2	4	0

Results - US

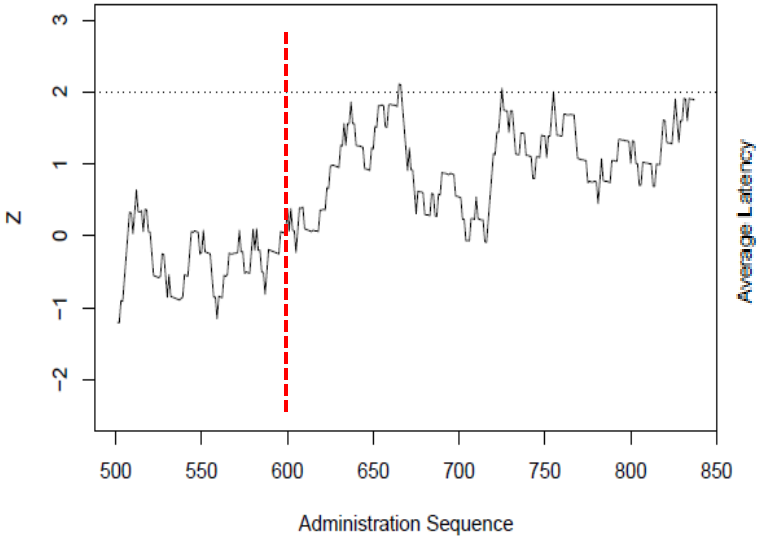
Example of Z and average latency for a flagged item based on item responses only (c_α = 3.5)

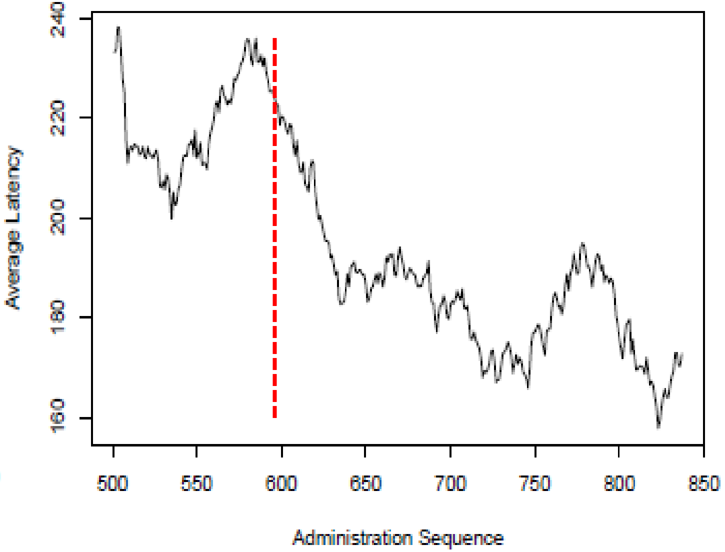


Possibly Type I error?

Results - US

Example of Z and average latency for a flagged item based on item responses (c_α =2.0) and *RTs*

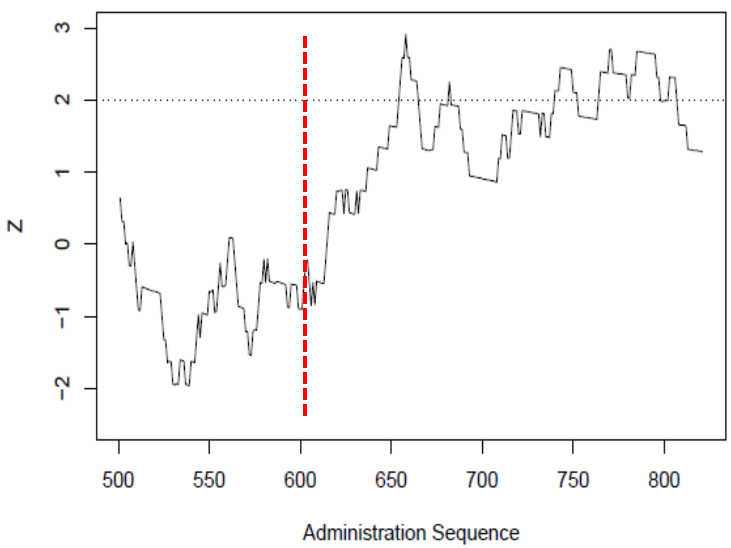


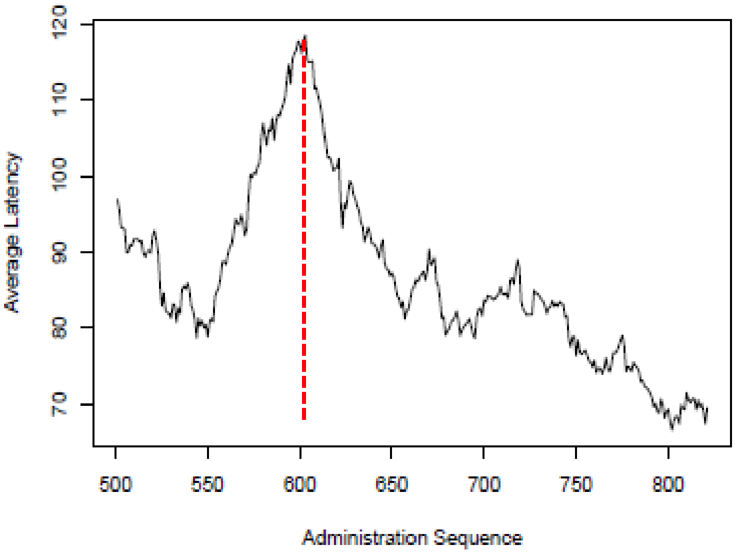


Potentially concerning?

Results - US

Example of Z and average latency for a flagged item based on item responses (c_{α} =2.0) and *RTs*

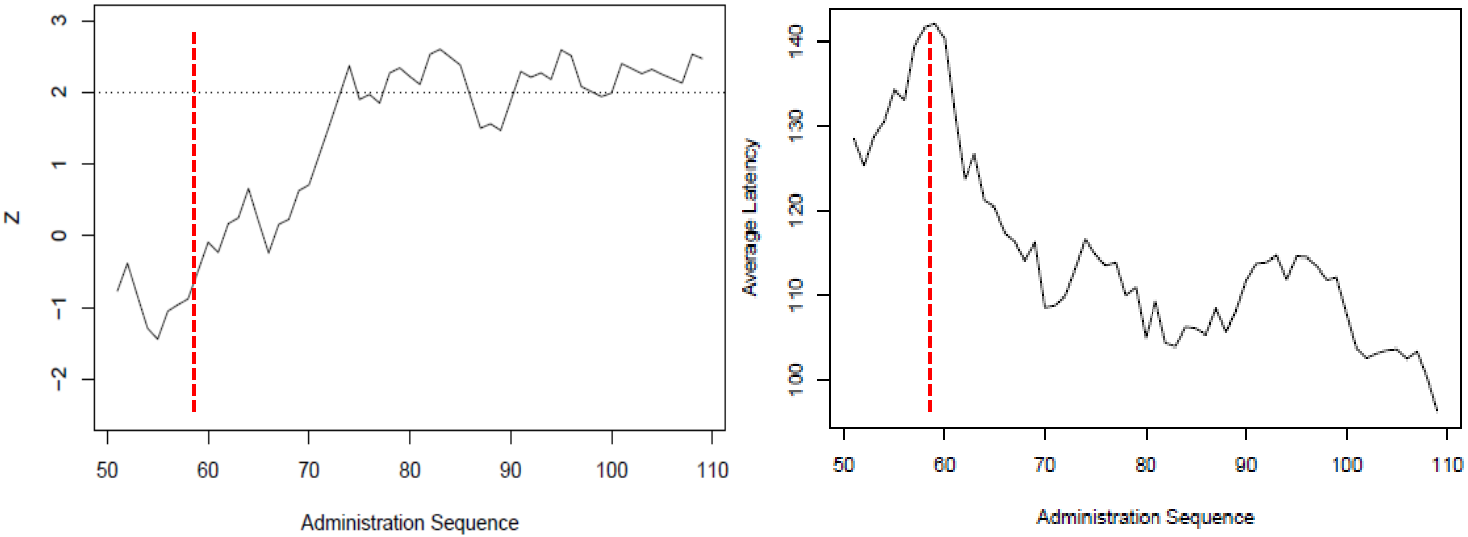




Potentially concerning?

Results – Country C

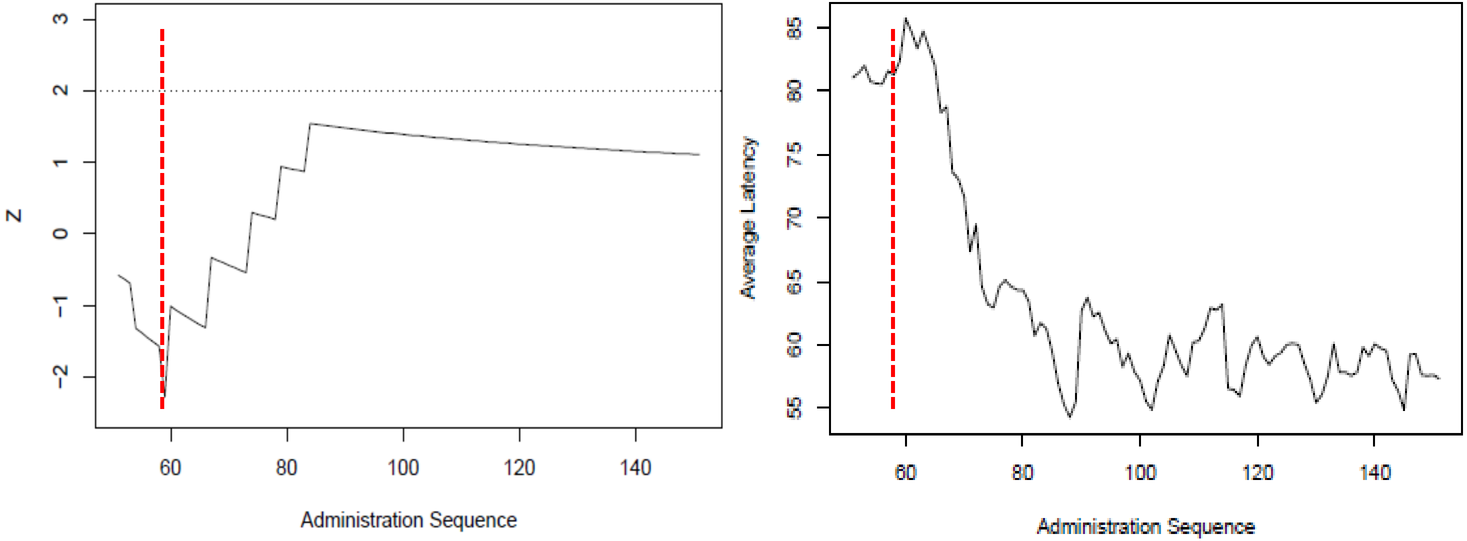
Example of Z and average latency for a flagged item based on item responses (c_α =2.0) and *RTs*



Potentially concerning?

Results – Country C

Example of Z and average latency for a flagged item based on item responses and *RTs*



Potentially concerning?

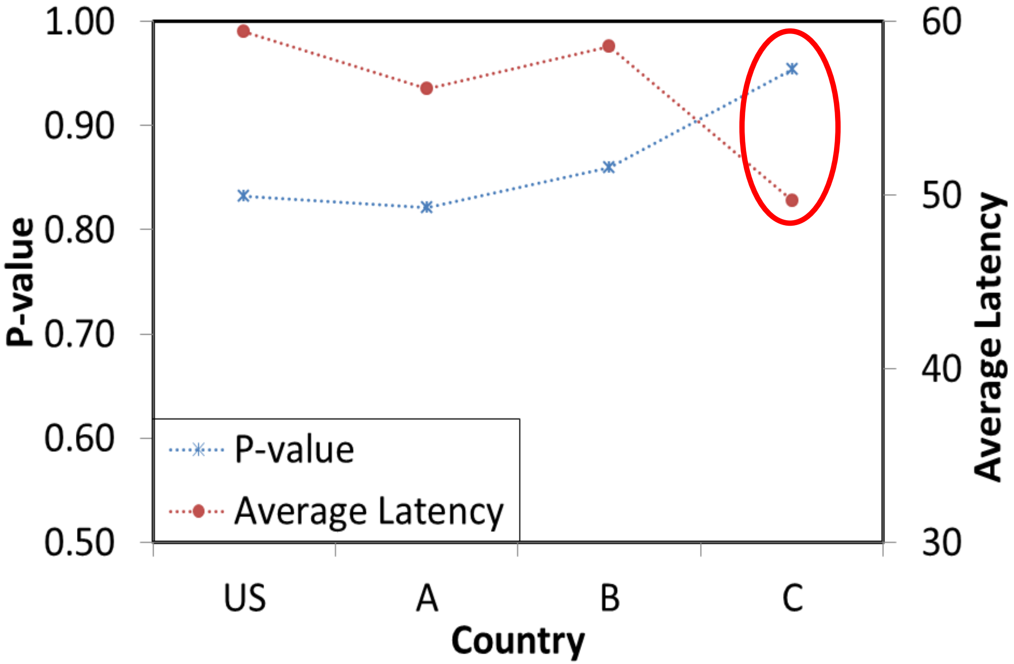
Results

Number of flagged items based on *RTs* and item responses

Country	US	A	B	C
N	5	1	0	4

Results

Overall item p-value and average item latency across different countries



Take Home

- A lot of items were likely falsely flagged using item responses only

- For the current dataset, only 10 items were flagged using *RTs* and item responses, and 4 of them may need more attention/monitoring

Thank You!

cliu@nbme.org