Discussion
Case Study
Simulations
Framework
Questions
Background

Application of a Sequential Procedure for Detecting Compromised Items to A CAT Licensure Exam

Danielle Lee & Hong Qian

NCSBN
Leading Regulatory Excellence

COTS
Conference On
Test Security

Discussion
Case Study
Simulations
Framework
Questions

Why detecting compromised items?

- Test items are reused over time in CAT.
- In high-stakes licensure exam, examinees may be motivated to seek access to the items.
- This can threaten the validity of the inferences from examination scores.

Background

How to prevent compromised items?

- Enlarge item pool.
- Change item pool frequently.
- Lower item exposure.
- Implement a monitoring system (e.g., Web monitor stolen item content.)

How to detect compromised items?

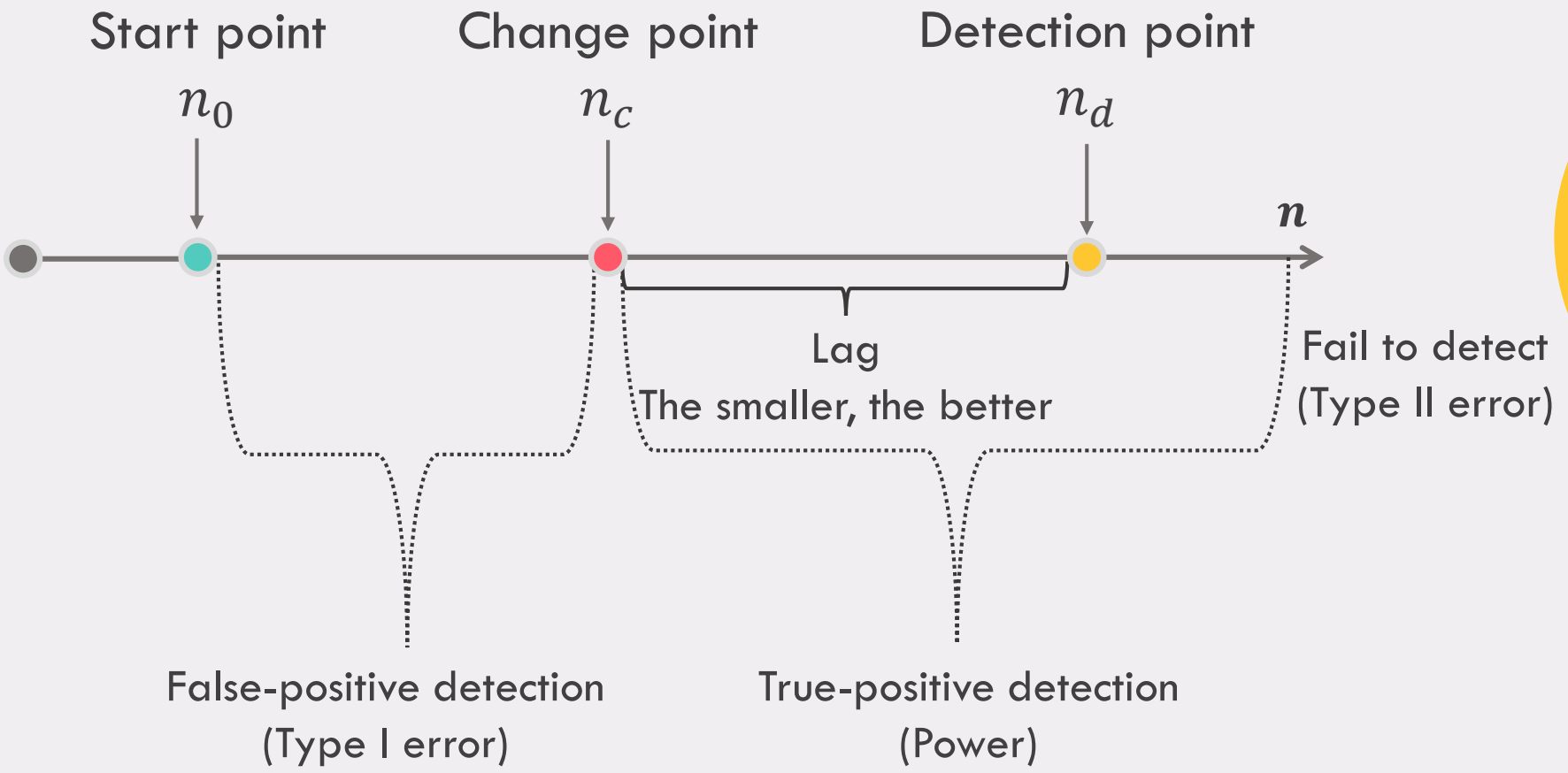
- A Sequential Procedure based on
  - Classical Test Theory (CTT) (Zhang, 2014)
  - Item Response Theory (IRT) (Zhang & Li, 2016)
- Use a series of statistical hypothesis tests.
- Examine whether the statistical characteristics of each item under inspection have changed significantly.

Research Questions

- Is the sequential procedure applicable to the variable length CAT exam?
  - Controlled Type I error
  - High Power
- Which factors in the procedure affect the detection of the compromised items?
- Do CTT-based and IRT-based procedures show similar results?
- Does the sequential procedure identify any compromised items in real item pool?

Theoretical Framework

Examinee Sequence n for Item i



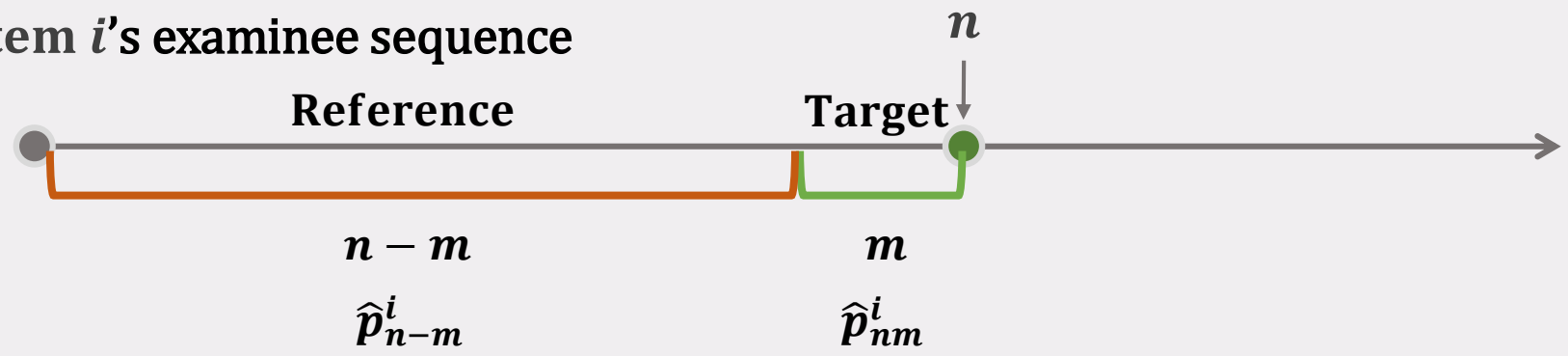
Statistics using CTT

- i: item index
- j: response or examinee index
- n: the sequence number of the examinees to whom item is administered in CAT
- m: moving sample size
- U_j^i: scored response j to item i (1 for a correct response, otherwise 0)

$$\hat{p}_{n-m}^{i} = \frac{1}{n-m}\sum_{j=1}^{n-m} U_j^i$$

$$\hat{p}_{nm}^{i} = \frac{1}{m}\sum_{j=n-m+1}^{n} U_j^i$$

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



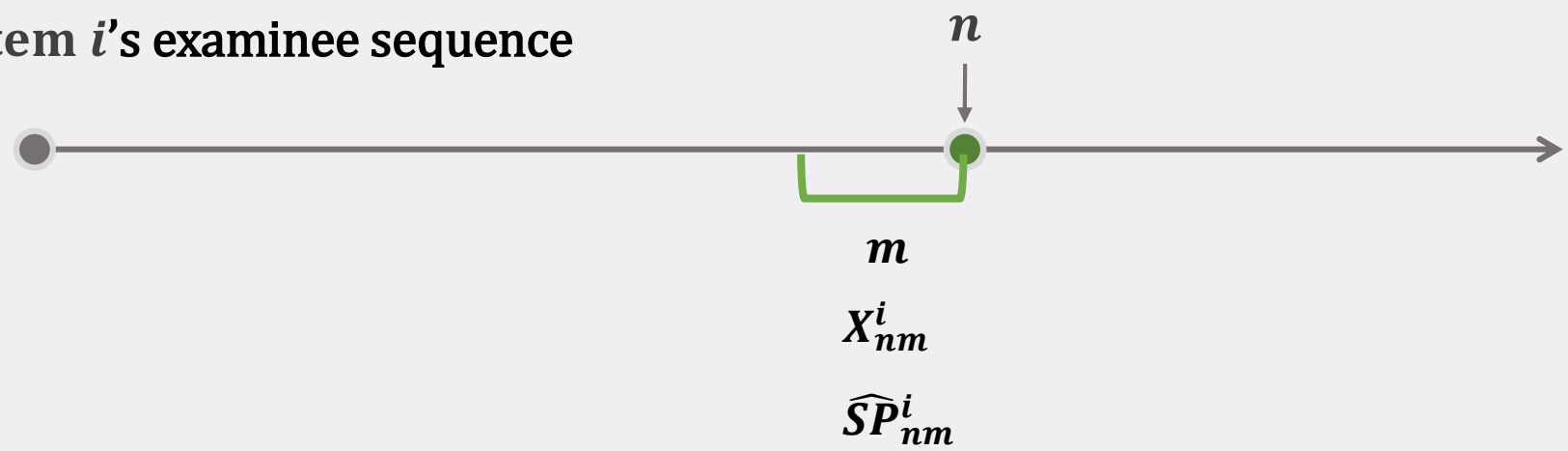
Statistics using IRT (Rasch Model)

- B_j: examinee j's ability
- D_i: item i's difficulty
- $X_{nm}^i = \sum_{j=n-m+1}^{n} U_j^i$: sum of the scores of the moving sample in item i
- $\widehat{SP}_{nm}^i = \sum_{j=n-m+1}^{n} p_j^i$: estimated correct responses of the moving sample in item i

$$p_j^i = \frac{\exp(B_j - D_i)}{1 + \exp(B_j - D_i)}$$

$$\hat{Y}_{nm}^i = \frac{X_{nm}^i - \widehat{SP}_{nm}^i}{\sqrt{\sum_{j=n-m+1}^{n} p_j^i (1 - p_j^i)}}$$

item i's examinee sequence



Simulation Studies

- Type I Error Study
  - Determine Cut-off value (C_α)
    - To detect compromised items.
    - If statistic $> C\alpha$ at n, identify the item as a compromised.
- Power Study
  - Power Rate
  - LAG

Type I Error Study

- Data Simulation based on the Real Data
- Real Data
  - 65,753 examinees: $B_j \sim N(0.3, 0.44^2)$
  - 1,464 administered items: $D_i = (-2.2 \sim 2.2)$
  - May contain compromised items
- Simulation Data
  - 86,000 examinees: $B_j \sim N(0.3, 0.44^2)$
  - 1,464 items: $D_i = (-2.2 \sim 2.2)$
  - No compromised items

Type I Error Study

- Item Selection
  - Number of examinees > 150 (n_0)
  - 837 items
- Simulation Factors
  - Monitor starting point (n_0) = 150
  - Moving sample size (m) = {25, 50, 75, 100}
- Cut-off values

Model (α)	m=25	m=50	m=75	m=100
CTT (.05)	3.9	3.9	3.8	3.8
IRT (.05)	3.9	4.0	3.9	3.9
CTT (.01)	4.2	4.2	4.3	4.2
IRT (.01)	4.3	4.3	4.4	4.4

Power Study

- Simulation Data
  - 86,000 examinees: $B_j \sim N(0.3, 0.44^2)$
  - 619 items: $D_i = (-1.62 \sim 2.11)$
    - Number of examinees > 500
  - 80 compromised items selected randomly
  - 100 replications

Power Study

- Simulation Factors
  - Monitor starting point (n_0) = 150
  - Moving sample size (m) = {25, 50, 75, 100}
  - Item exposure degree
    - Low: Number of examinees (n_i) between 500 and 3,000
    - Middle: Number of examinees (n_i) between 3,000 and 6,900
    - High: Number of examinees (n_i) between 6,900 and 18,314
  - Item compromised point
    - Begin: $n_0 + 50u,\ u \sim$ Uniform [0, 1]
    - Middle: $\frac{n_i}{2} + 50u$
    - End: $n_i - n_0 + 50u$
  - Nominal significance levels (α) = {.05, .01}
  - p_j^i's increment = {0.1, 0.2, …until reaching 1.0}

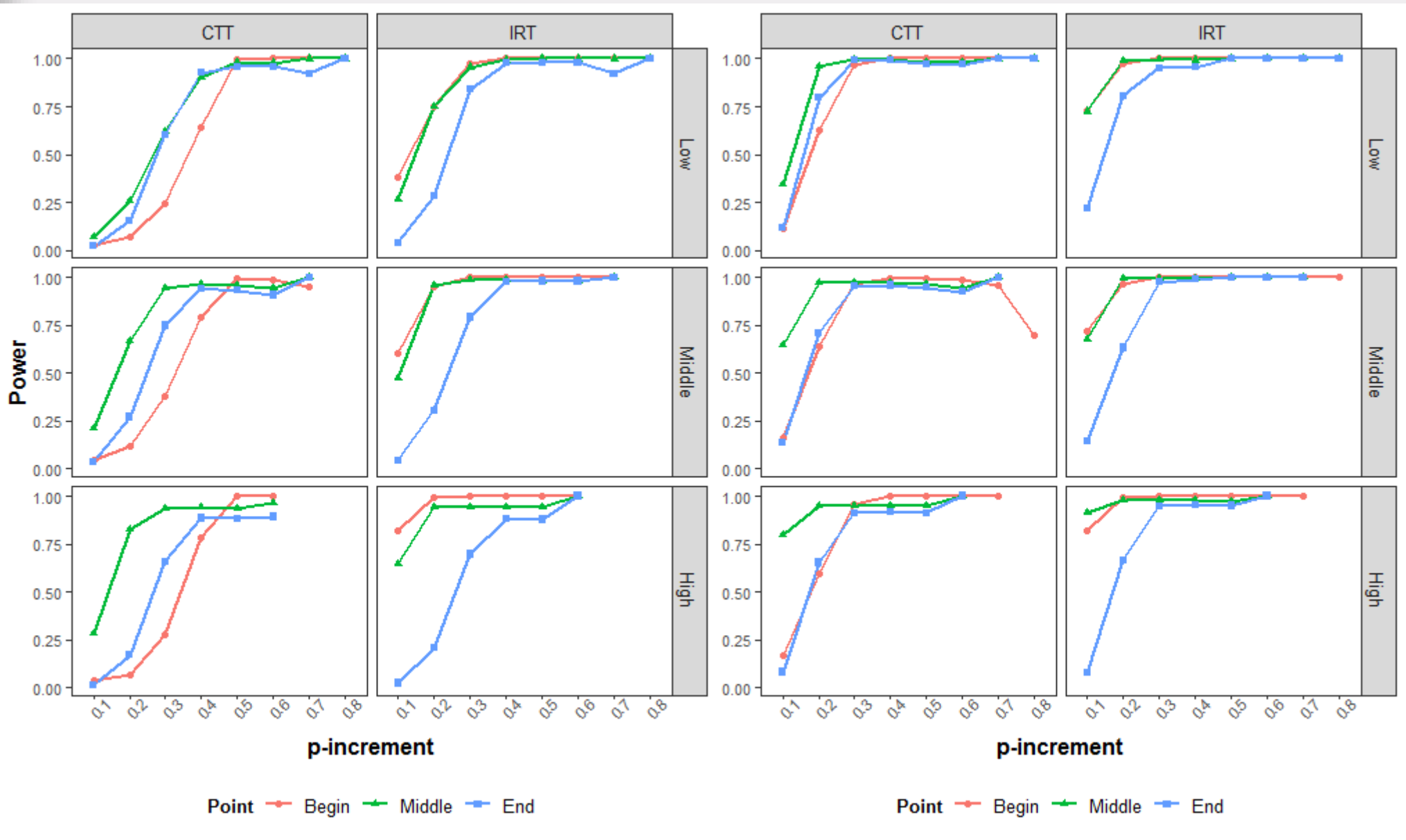
Power Study

- Response Manipulations
  - p_j^i's increment = 0.1
    - Reference p value: the proportion of all the correct responses before the compromised point
    - Increase the probability of the correct response from the reference p value by 0.1 until reaching 1.0
    - $U_j^i = 1$ if $u < p_j^i, u \sim$ Uniform[0,1] , otherwise $U_j^i = 0$

Power Study

- Response Manipulations
  - p_j^i's increment = 0.1
    - Reference p value: the proportion of all the correct responses before the compromised point
    - Increase the probability of the correct response from the reference p value by 0.1 until reaching 1.0
    - $U_j^i = 1$ if $u < p_j^i, u \sim$ Uniform[0,1] , otherwise $U_j^i = 0$

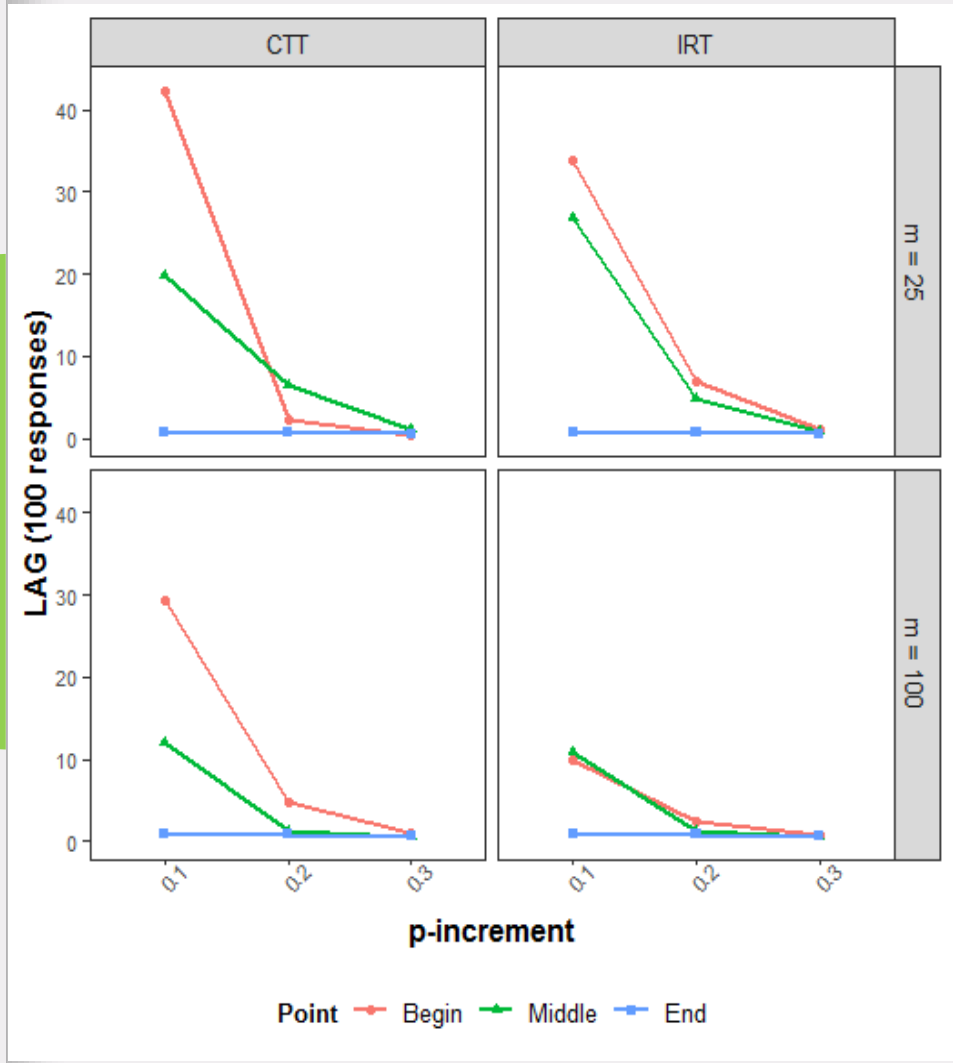
Results from Power Study: Power

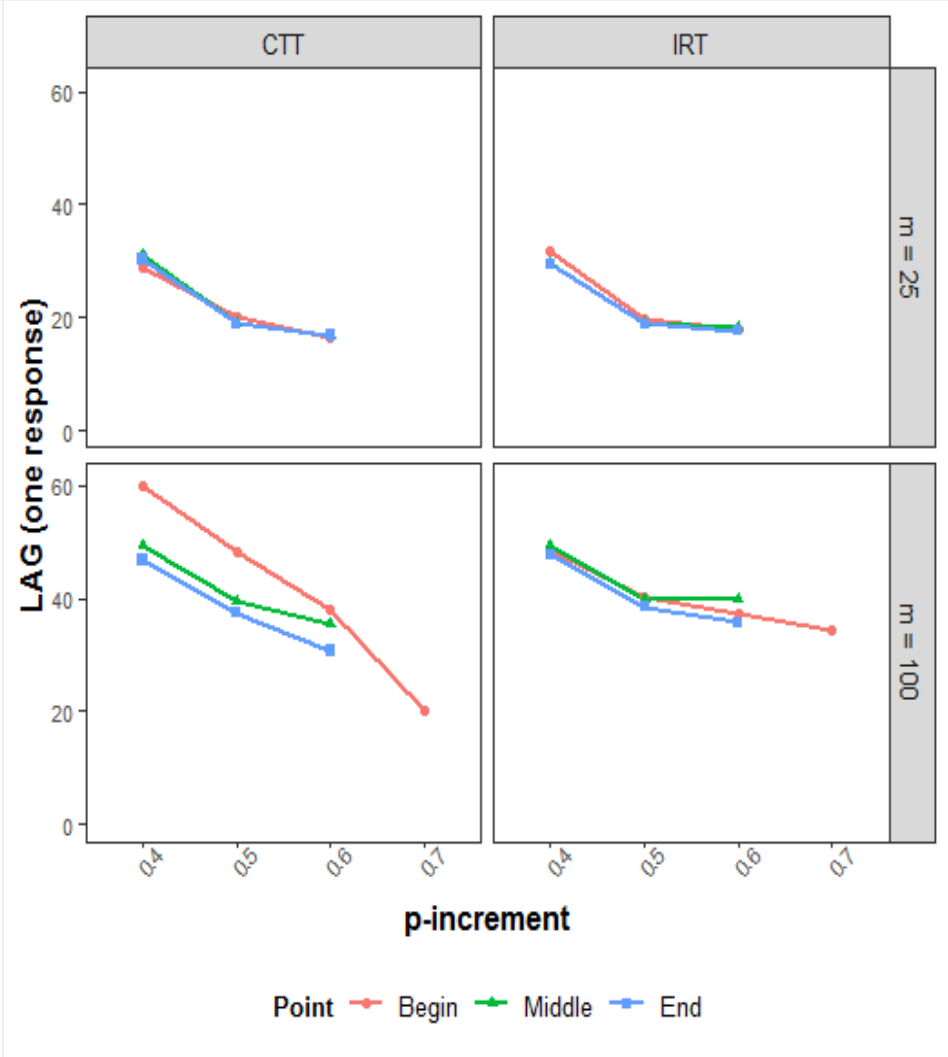


Results from Power Study: LAG

		m = 25		*m* = 100	
Point	Exposure	CTT	IRT	CTT	IRT
Begin	Low	96	157	83	88
	Middle	208	293	133	172
	High	760	707	525	211
Middle	Low	84	112	80	75
	Middle	170	213	115	143
	High	467	553	250	235
End	Low	32	33	48	49
	Middle	37	38	54	58
	High	44	44	59	61

Results from Power Study: LAG





Discussion
Case Study
Simulations
Framework
Questions
Background

NCSBN
Leading Regulatory Excellence

COTS
Conference On
Test Security

Results from Power Study

- Power increased as
  - *p*-increment increased.
  - Moving sample size (*m*) increased.
  - Exposure degree was higher.
- For CTT model, Power was higher when the item was compromised at the middle point of item administration (3,003th response) with small *p*-increment.
- For IRT model, Power was slightly higher when the item was compromised at the begin point of item administration (174th response) with small *p*-increment.
- IRT model detected compromised items better than CTT model when the item was compromised at the begin and middle point of item administration.

Results from Power Study

- LAG decreased as
  - *p*-increment increased.
  - Item exposure degree was low.
  - Moving sample size increased with smaller *p*-increment (0.1 or 0.2.)
  - Moving sample size decreased with larger *p*-increment (more than 0.3.)

Case Study

- Real Data
  - 65,753 examinees: $B_j \sim N(0.3, 0.44^2)$
  - 837 administered items: $D_i = (-2.2 \sim 2.2)$
  - May contain compromised items
- Conditions
  - Monitor starting point $(n_0) = 150$
  - Moving sample size $(m) = 50$
  - Cut-off values (C_α)

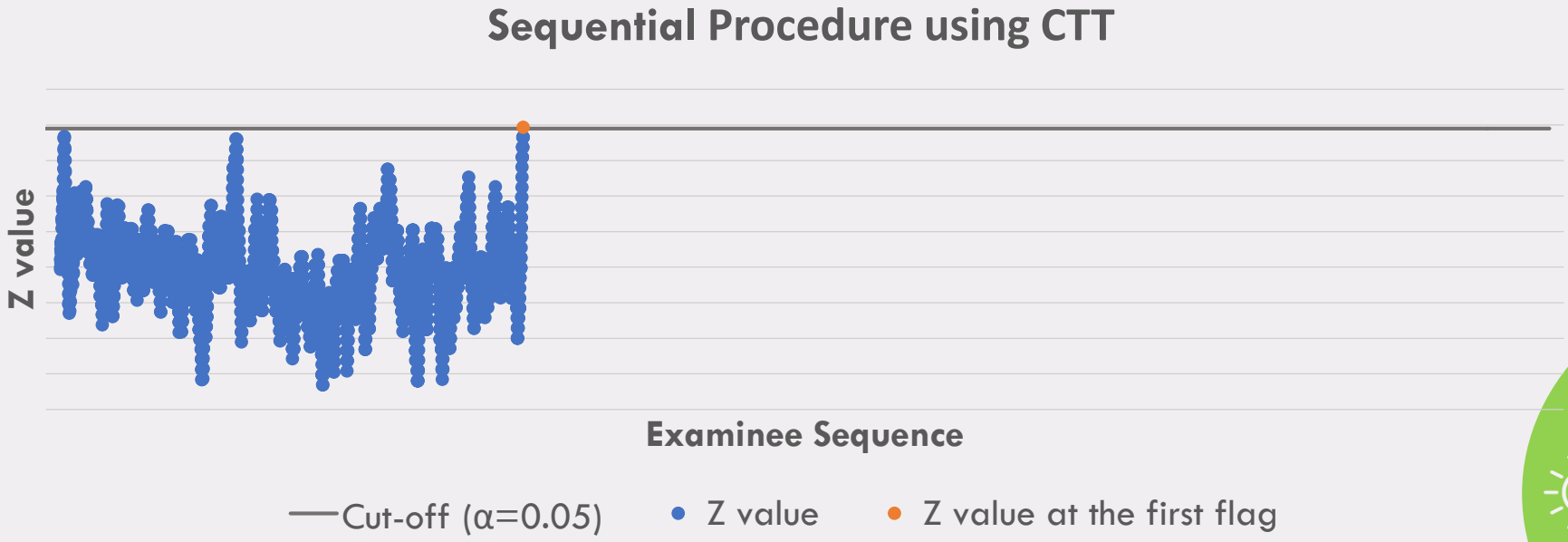
Model (α)	CTT (.05)	IRT (.05)	CTT (.01)	IRT (.01)
C_α	3.9	4.0	4.2	4.3

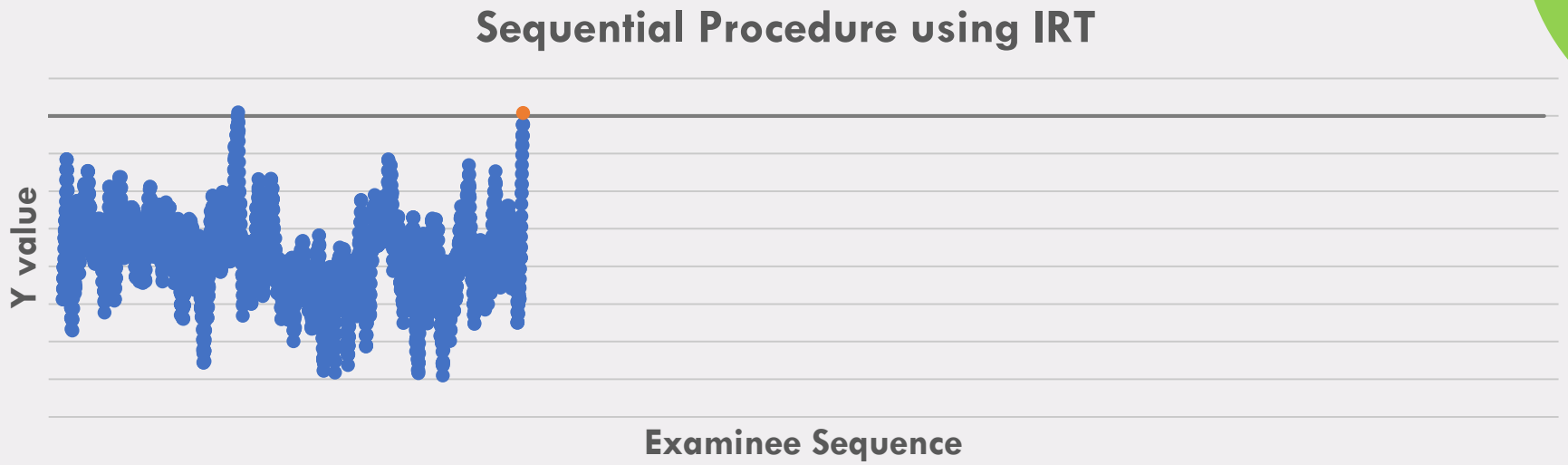
Results from Case Study

- Number (Flag Rate) of flagged items as a compromised item

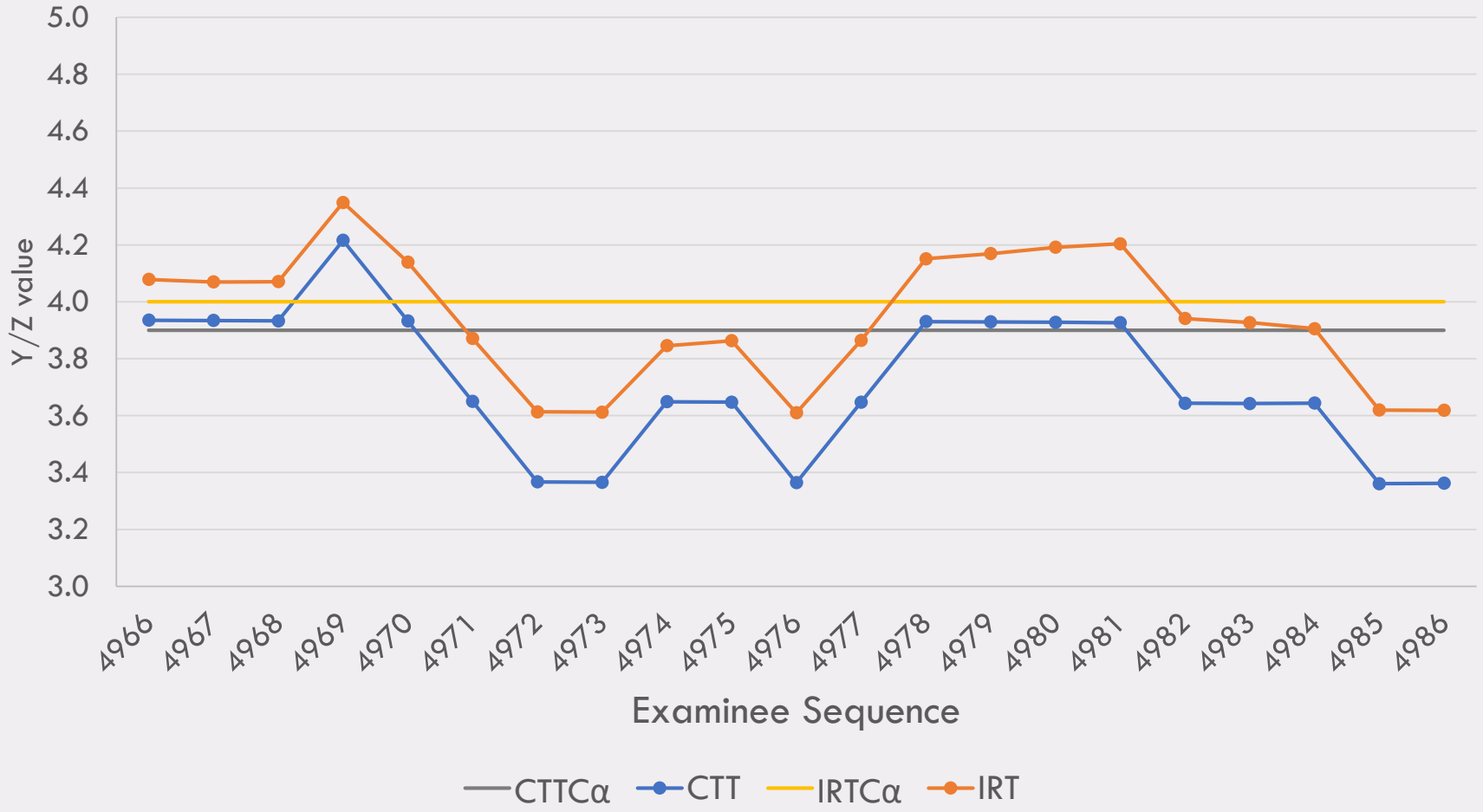
Model	α = .05 (Flag Rate)	α = .01 (Flag Rate)
CTT	45 (.054)	15 (.018)
IRT	82 (.098)	48 (.057)
CTT and IRT	17 (.023)	10 (.012)

One item flagged as the compromised

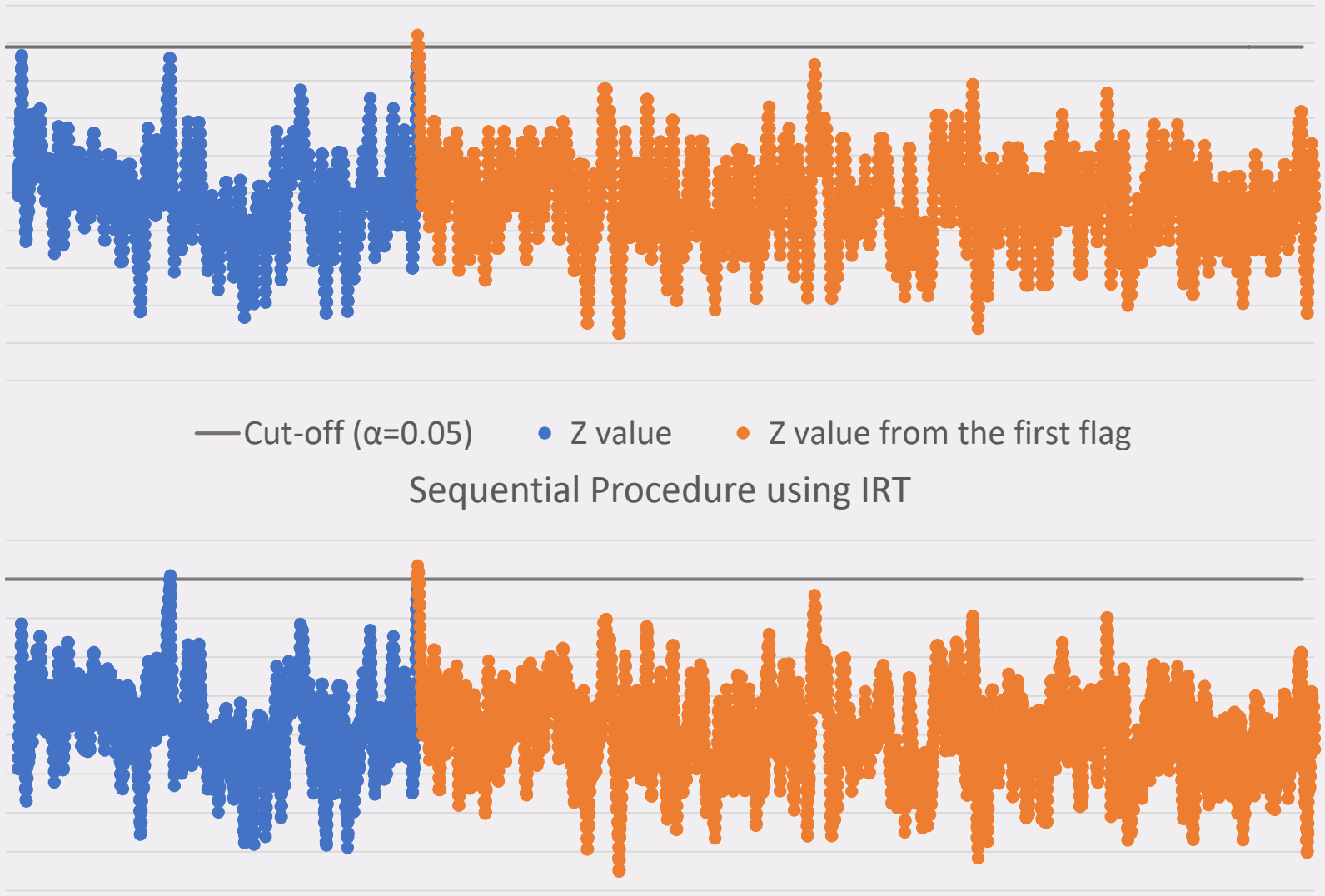




One item flagged as the compromised



One item flagged as the compromised



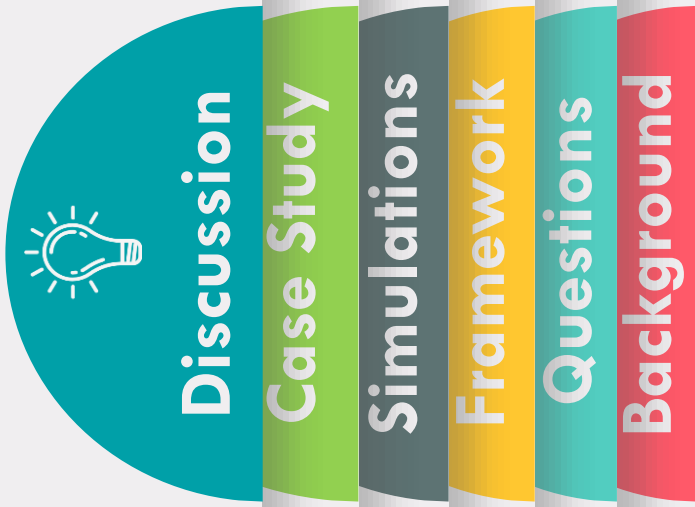
Discussion

- The sequential procedure can be applicable to the variable length CAT exam.
- The severity of the compromisation, compromised point, moving sample size, and item exposure degree were significant factors for Power and LAG.
- The results of the CTT and IRT models were slightly different in Power under different factors, but similar in LAG.
- 17 (.023 at $\alpha = .05$) items were flagged as a compromised items by both CTT and IRT.

Thank You!

Danielle Lee

dlee@ncsbn.org



NCSBN
Leading Regulatory Excellence

COTS
Conference On
Test Security