Detecting Aberrant Behavior in CAT: The Lognormal Response Time Model

Xiaowen Liu
University of Connecticut
xiaowen.liu@uconn.edu

2019 IACAT
Minneapolis, Minnesota, USA.

6/12/2019

UCONN

Outline

- Background: Aberrant Behavior in Assessment and Response Time
- The Model and Person-Fit Statistics
- The Simulation Study and Results
- Conclusions, Limitations and Educational Implications

- Background: Aberrant Behavior in Assessment and Response Time
- The Model and Person-Fit Statistics
- The Simulation Study and Results
- Conclusions, Limitations and Educational Implications

Statewide Assessment

- Statewide assessment systems measure student achievement and growth as part of program evaluation and school, district, and accountability systems.

CSDE

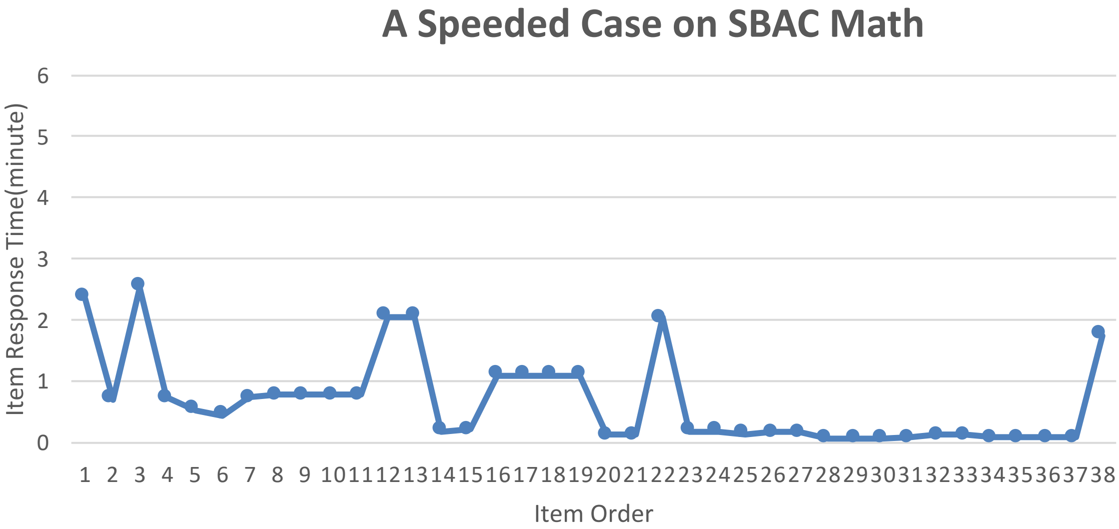
CONNECTICUT STATE
DEPARTMENT OF EDUCATION

- Connecticut has used Smarter Balanced Assessments for testing ELA and Mathematics since 2015.
- Smarter Balanced Assessments are CATs.
- Tests have no significant consequence for students.

Statewide Assessment

- Examinees with low motivation may not give their best effort

- These students show aberrant behavior in their responses (e.g., randomly select options with short response time.)


A Speeded Case on SBAC Math

- Students' aberrant response patterns need to be monitored, flagged, and explored to ensure the validity of results obtained from these testing programs.

Response Time

- With the advent of computerized test administration, information about response time is also available and is useful in detecting aberrant response behavior.
- Response time (RT) is an indicator that reflects information about respondents’ speed and mental activities, as well as item and test characteristics (Lee & Chen, 2011; Marianti et al., 2014).
- van der Linden (2006, 2007) investigated a lognormal response time model and the hierarchical model of speed and accuracy to estimate the parameters of person speed, item time intensity, and time discrimination.

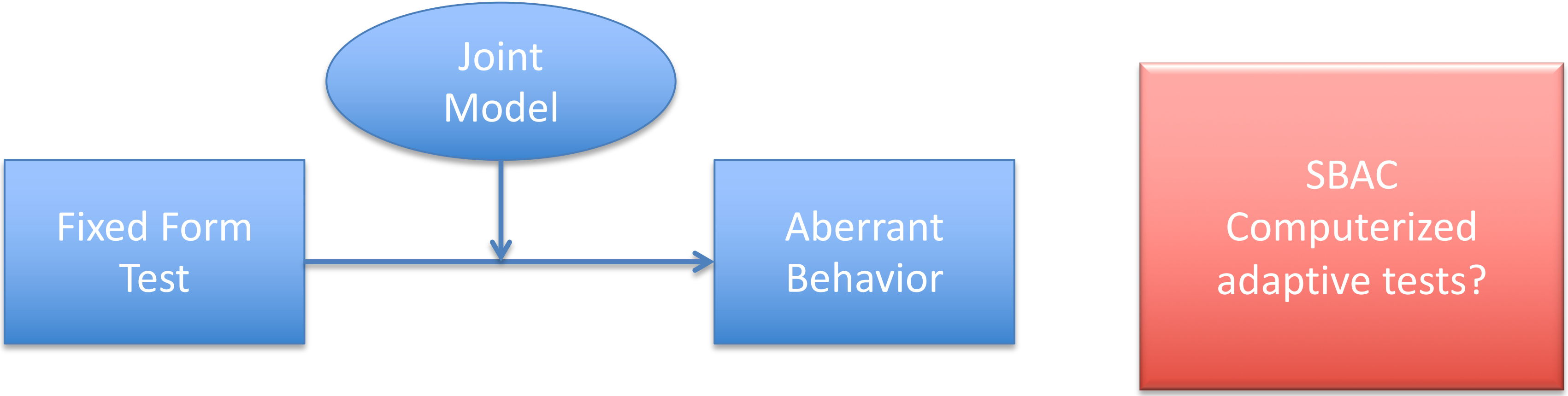
Person-Fit

- Marianti et al. (2014) applied the lognormal response time model and proposed a likelihood-based person-fit statistic to detect different types of aberrant test-taker response times.
- Recently, Fox and Marianti (2017) proposed person-fit statistics for joint model for accuracy and speed to detect aberrant responses accuracy and/or response time patterns.

Importance of Detecting Aberrances in CAT

- Aberrant behavior can impair test validity, especially in adaptive testing.
- Item selection and ability estimation heavily depend on the item response. If aberrant behavior emerges during the progress of the test, items would be inappropriately administrated, which leads to inaccurate performance estimation.
- Therefore, aberrant response patterns should be identified and further actions should be taken for irregular test-takers.

Research Question



Joint Model

Fixed Form Test

Aberrant Behavior

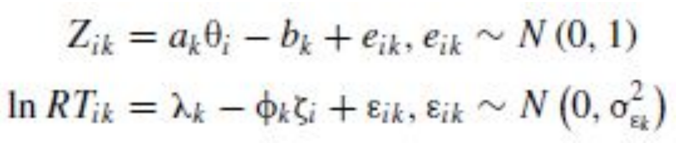
SBAC Computerized adaptive tests?

The current study extends the person-fit statistics for the joint model in Fox and Marianti (2017) to computerized adaptive testing to detect three types of aberrant test-taker responses and response times.

- Background: Aberrant Behavior in Assessment and Response Time
- The Model and Person-Fit Statistics
- The Simulation Study and Results
- Conclusions, Limitations and Educational Implications

Joint Model for Accuracy and Speed

The joint modeling approach for the latent continuous responses and RTs uses a two-parameter IRT model for binary responses and the log-normal model for response time:



$$Z_{ik} = a_k\theta_i - b_k + e_{ik}, e_{ik} \sim N\,(0, 1)$$

$$\ln RT_{ik} = \lambda_k - \phi_k\zeta_i + \varepsilon_{ik}, \varepsilon_{ik} \sim N\left(0, \sigma^2_{\varepsilon_k}\right)$$

For the two-parameter IRT model, the response is the indicator of the latent response variable Zik being positively truncated.
ζi (speed parameter)represents the working speed of student i*;*
*φk (*time discrimination parameter) represents the item-specific effect of working speed on the RT;
Λk (time intensity parameter) represents the average time needed to complete the item.

Fox& Marianti (2017)

Joint Model for Accuracy and Speed

The test takers are assumed to be randomly selected from a population and the ability and speed variables are assumed to have a multivariate normal population distribution

$$\begin{pmatrix}\theta_i\\ \zeta_i\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\theta\\ \mu_\zeta\end{pmatrix}, \begin{pmatrix}\sigma_\theta^2 & \rho_{\theta\zeta}\\ \rho_{\theta\zeta} & \sigma_\zeta^2\end{pmatrix}\right)$$

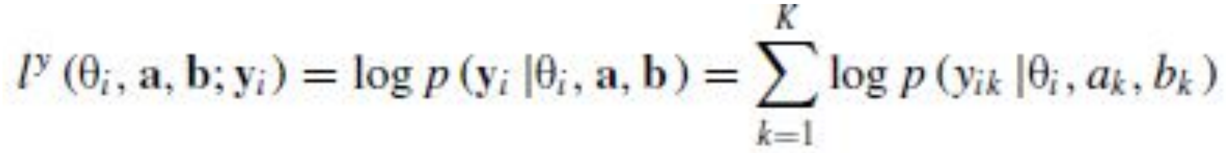
The population distribution of the item characteristics is a multivariate normal, which is given by

$$\begin{pmatrix}a_k\\ b_k\\ \phi_k\\ \lambda_k\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_a\\ \mu_b\\ \mu_\phi\\ \mu_\lambda\end{pmatrix}, \begin{pmatrix}\sigma_a^2 & \rho_{ab} & \rho_{a\phi} & \rho_{a\lambda}\\ \rho_{ab} & \sigma_b^2 & \rho_{b\phi} & \rho_{b\lambda}\\ \rho_{a\phi} & \rho_{b\phi} & \sigma_\phi^2 & \rho_{\phi\lambda}\\ \rho_{a\lambda} & \rho_{b\lambda} & \rho_{\phi\lambda} & \sigma_\lambda^2\end{pmatrix}\right)$$

Fox& Marianti (2017)

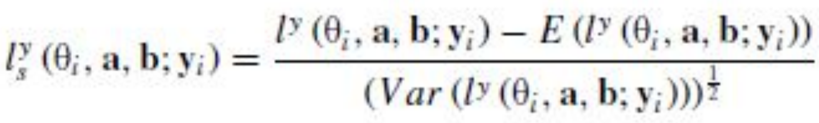
Person-Fit Statistics: Responses

Fox & Marianti (2017) used the l^y person-fit statistic based on the log-likelihood of the responses to evaluate the fit of an response (RA) pattern.



$$l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) = \log p(\mathbf{y}_i \,|\theta_i, \mathbf{a}, \mathbf{b}) = \sum_{k=1}^{K} \log p(y_{ik} \,|\theta_i, a_k, b_k)$$

The standardized version of this person-fit statistic is given by



$$l^y_s(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) = \frac{l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) - E(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i))}{(Var(l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i)))^{\frac{1}{2}}}$$

Fox& Marianti (2017)

Person-Fit Statistics: Responses

A Bayesian significance test with the MCMC algorithm was used to compute the extremeness of each RA pattern.

$$F_i^y = \begin{cases} 1 & \text{if } P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) > C\right) \\ 0 & \text{if } P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) \leq C\right) \end{cases}$$

The average for all MCMC iterations computing the status for F_i^y is used as an estimate of the posterior probability of an aberrant RA pattern.

Fox& Marianti (2017)

Person-Fit Statistics: Response Time

$$l^t = \sum_i^I Z_{ip}^2 = \sum_i^I (\frac{t_{ip}-u_{ip}}{\sigma_i})^2$$

where u_{ip} is the predicted response time calculated from the estimated parameters for each item and person.

The posterior distribution of the statistic can be used to examine whether a pattern of observed RTs is extreme under the model.

$$F_i^t = \begin{cases} 1 & \text{if } P\left(l^t\left(\zeta_i, \phi, \lambda; \mathrm{rt}_i\right) > C\right) \\ 0 & \text{if } P\left(l^t\left(\zeta_i, \phi, \lambda; \mathrm{rt}_i\right) \leq C\right). \end{cases}$$

Marianti et al., (2014) and Fox & Marianti (2017)

Person-Fit Statistic For RA and RT

Fox & Marianti (2017) proposed a classification variable $F_i^{t,y}$

$F_i^{t,y}$ equals 1 when the other classification variables are both equal to 1, i.e., F_i^y= 1 and F_i^t= 1, and equals 0 otherwise.

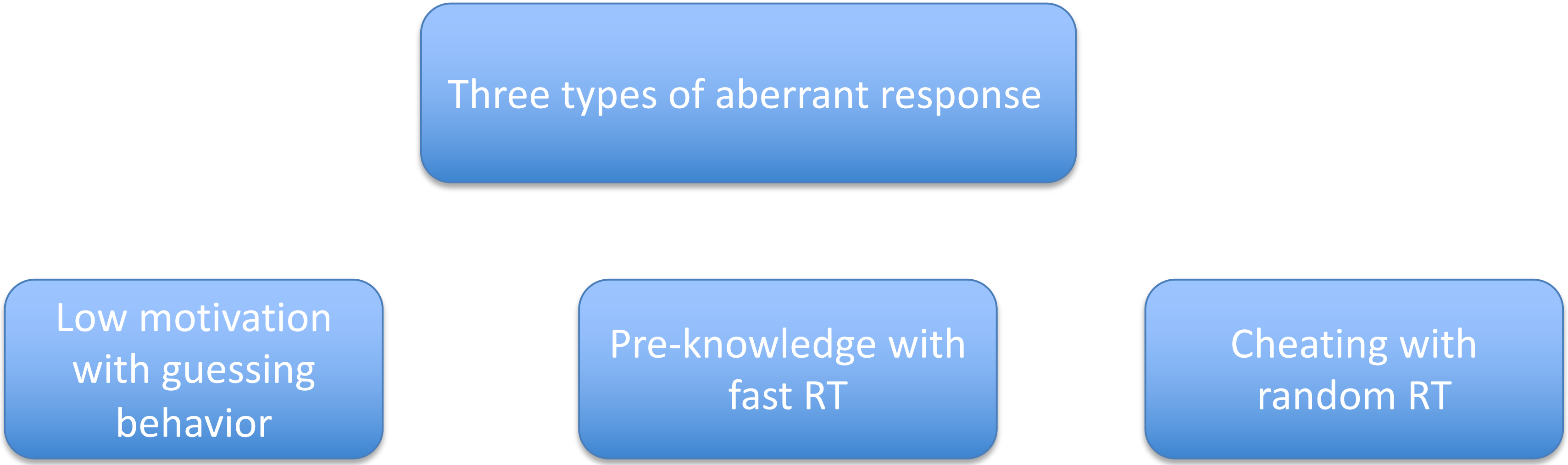
$$F_i^{t,y} = \begin{cases} 1 & \text{if } P\left(l^t(\zeta_i, \phi, \lambda; \mathbf{rt}_i) > C, l_s^y(\theta_i, \alpha, \beta; \mathbf{y}_i) > C\right) \\ 0 & \text{if } 1 - P\left(l^t(\zeta_i, \phi, \lambda; \mathbf{rt}_i) > C, l_s^y(\theta_i, \alpha, \beta; \mathbf{y}_i) > C\right) \end{cases}$$

The response pattern is flagged as aberrant when the observed statistic value is larger than the critical value at alpha equal to .05.

Fox& Marianti (2017)

- Background: Aberrant Behavior in Assessment and Response Time
- The Model and Person-Fit Statistics
- The Simulation Study and Results
- Conclusions, Limitations and Educational Implications

Simulation Study



Three types of aberrant response

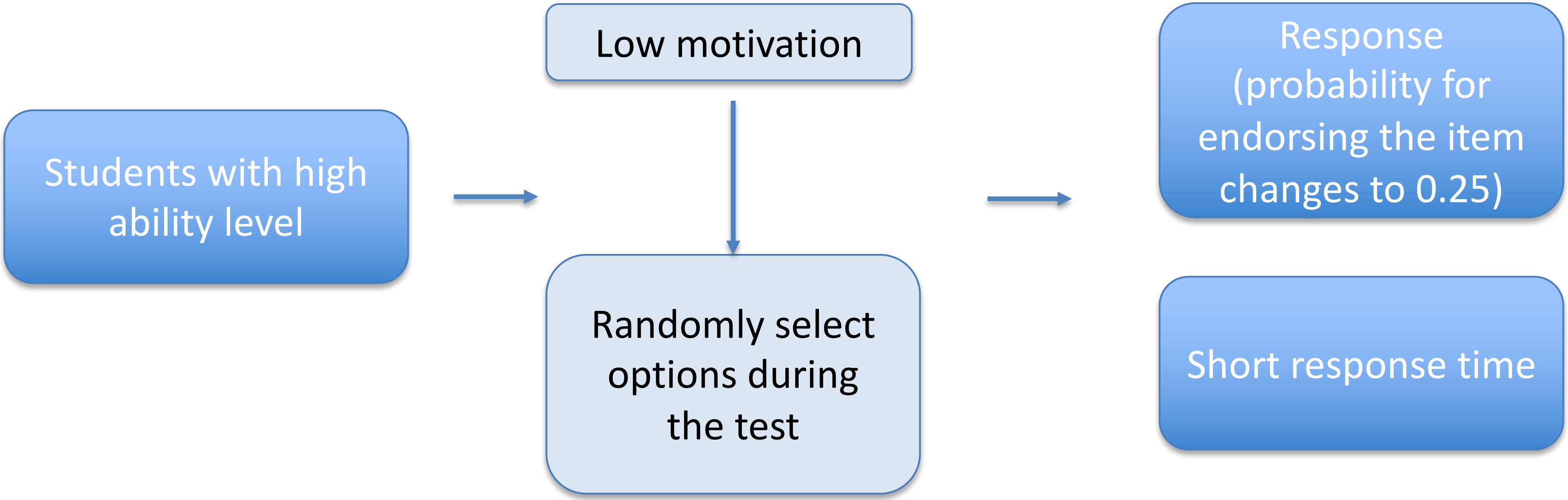
Low motivation with guessing behavior

Pre-knowledge with fast RT

Cheating with random RT

Simulation Study

Low motivation with guessing behavior



Low motivation

Students with high ability level

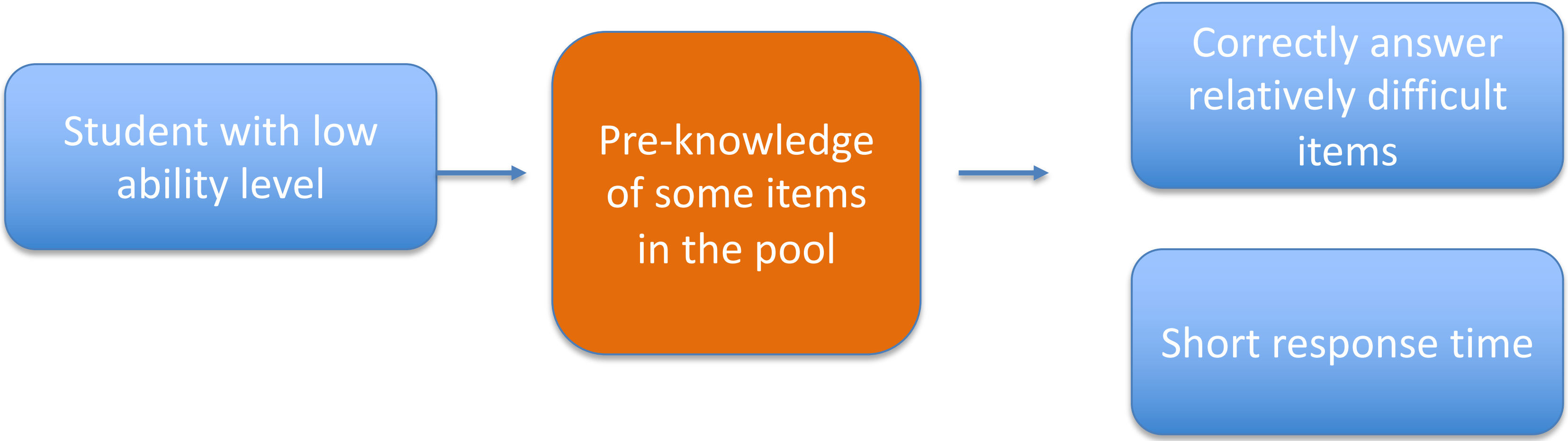
Randomly select options during the test

Response (probability for endorsing the item changes to 0.25)

Short response time

Simulation Study

Pre-knowledge with fast RT



Student with low ability level

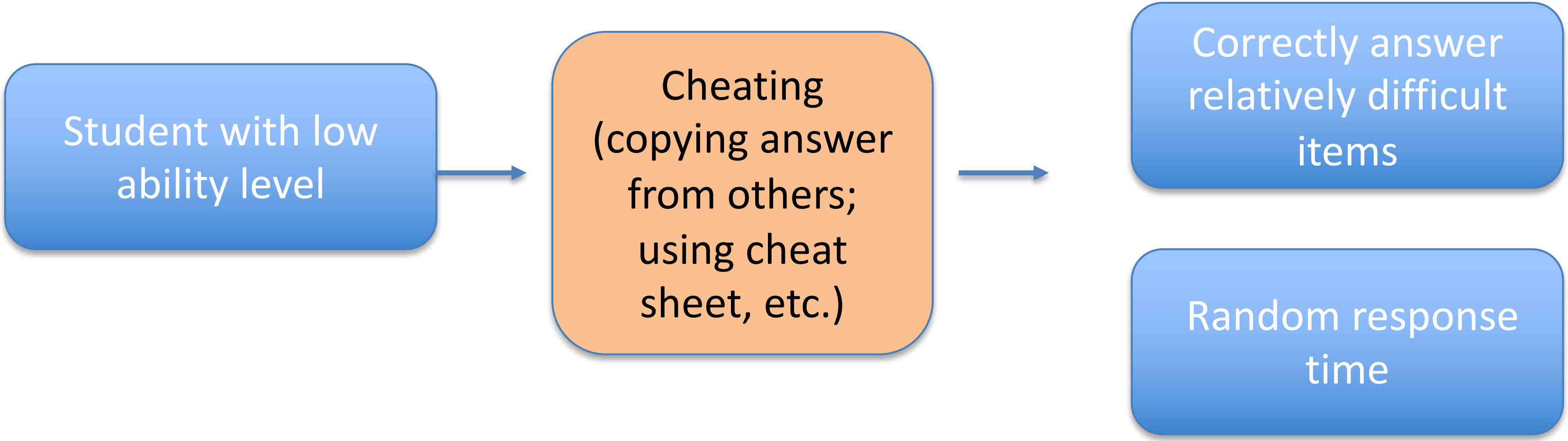
Pre-knowledge of some items in the pool

Correctly answer relatively difficult items

Short response time

Simulation Study

Cheating with random RT



Student with low ability level

Cheating (copying answer from others; using cheat sheet, etc.)

Correctly answer relatively difficult items

Random response time

Simulation Study

Aberrance Type	Ability Level	Aberrant ability level	Response time
Guessing (high ability with less motivation)	High	Low	Short
Preknowledge with fast RT	Low	High	Short
Cheating with random RT	Low	High	Random aberrant RTs were generated from a log-normal distribution with large SD of RTs

Simulation Study

- Two different proportions of test-takers with aberrant responses and response times were generated: 5% and 20%.
- The percentage of items with aberrant responses and response times were set at either 25% or 50%.
- The sample size was 1000 examinees and the test length was 40 items (80-item bank)
- The current study did post-hoc CAT simulation using the R package catIrt (Nydick, 2014). The R package LNIRT (Fox, Klotzke, and Entink, 2018) was used in fitting the joint model for accuracy and speed and computing the person-fit statistics. Missing data by design were ignored in the parameter estimation and person-fit statistics calculation.

Results: Hit and False Positive Rates

	Low Motivation—Guessing Fast RT		Pre-knowledge with Fast RT		Cheating with Random RT	
Hit Rate	25% Item	50% Item	25% Item	50% Item	25% Item	50% Item
5% Student	0.71	0.66	0.72	0.52	0.59	0.56
20% Student	0.40	0.30	0.43	0.17	0.42	0.24
False Positive						
5% Student	0.01	0.01	0.01	0.01	0.01	0.01
20% Student	0.01	0.02	0.00	0.02	0.00	0.02

Potential Reasons

Compared to Fox and Marianti (2017), the results of the current study show lower power in detecting aberrance.

The adaptive test design--

- Longer tests: 40 items (compared to 20-item test in Fox and Marianti (2017)).
- This could lead to more aberrant responses of the items in the test, which will impair the overall accuracy of item calibration(time parameter) (e.g., in the 25% aberrant item condition, we designed 25% aberrance. But different person took different items. Overall there are more than 25% aberrant items in the test.)

- Background: Aberrant Behavior in Assessment and Response Time
- The Model and Person-Fit Statistics
- The Simulation Study & Results
- Conclusions, Limitations and Educational Implications

Conclusions

- The current study applied the joint modeling of responses and response times to detect aberrant test-takers by simulating three types of aberrant responses and response time behaviors in the context of adaptive testing.

- The person-fit statistics performs well when there is a low percentage of aberrance in the response pattern. As the percentage of test-takers with aberrant responses increased, the hit rates decrease. This result shows that when there is a large number of students with aberrant responses, many of them will not be flagged.

- Further, when half of the items are simulated as aberrant in the test, the hit rate drops. Since more aberrant responses and responses time appear in the test, the normal behavior and aberrant behavior couldn’t be differentiated for each response pattern.

Limitations

The detection rates for the condition of high percentage of aberrance in students and items are low. The use of joint model for ability and response time should be considered carefully in low stake tests. More comprehensive studies need to be done for further investigation of this person fit statistic.

Educational Implications

- Response time can be used as an indicator for the detection of aberrant response behaviors. Combining response time with responses in person fit statistics provides more information in identifying aberrances.

- Missing by design has been taken into account in the current study, which makes the joint model feasible in detecting aberrance in computerized adaptive testing.

- Aberrant behavior can impair test validity, especially in adaptive testing. With procedures for accurate identification of aberrant responding, such test-takers can be removed during calibration of the test items to improve item parameter estimation.

References

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, J. P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. Journal of educational measurement, 54(2), 243-262.

Fox, J. P., Klotzke, K., and Entink, R. K. (2018). R-package LNIRT.

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359-379.

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of educational and behavioral statistics*, *39*(6), 426-451.

Nydick, S. W. (2014). catIrt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.5-0.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251-265.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365-384.

Questions