Simultaneous Estimation of Compromised Items and Examinees with Item Preknowledge using Response Time Data

Cengiz Zopluoglu

College of Education, University of Oregon

April 22, 2022

Presented at the annual meeting of NCME, San Diego

UNIVERSITY OF OREGON

Background

- A new model to identify examinees with item preknowledge and compromised items in a single-stage analysis.

Deterministic Gated IRT Model (DG-IRT;Shu et al., 2013) + **Lognormal Response Time Model** (LNRT; van der Linden, 2006) + Additional Improvements

- The new model:
  - synthesizes the ideas from DG-IRT and LNRT
  - No assumption that the compromised items are known
  - Direct estimate of examinees with item preknowledge and comromised items
  - Marginalization of discrete parameters in the model

Model Description

$$t_{ij}^{*} | \tau_{ti}, \tau_{ci}, H_i, \beta_j, \alpha_j, C_j \sim \mathcal{N}(\mu_{ij}, \alpha_j^{-2})$$

t_{ij}^{*} is the observed logresponse time for examinee i on item j.

- **Person parameters**:
  - H_i: examinee item preknowledge status (1:yes, 0:no)
  - τ_{ti}: latent speed parameter for uncompromised items
  - τ_{ci}: latent speed parameter for compromised items
- **Item parameters**:
  - C_j: item compromise status (1:yes, 0:no)
  - β_j: time-intensity parameter
  - α_j: time-discrimination parameter
- **Gating mechanism**

$$\mu_{ij} = \begin{cases} \beta_j - \tau_{ci} & \text{, when } C_j = 1 \text{ and } H_i = 1 \\ \beta_j - \tau_{ti} & \text{, otherwise} \end{cases}$$

Breaking down the density for the distribution of the observed logresponse time

$$f(t^*_{ij};\tau_{ti},\tau_{ci},H_i,\beta_j,\alpha_j,C_j)$$

$$\downarrow$$

This density can be written as a sum of four terms that represent all possible combinations of H_i and T_j:

- An **examinee with item preknowledge** responds to a **compromised item** ($H_i = 1$, $C_j = 1$)

$$f(t^*_{ij};\tau_{ti},\tau_{ci},H_i=1,\beta_j,\alpha_j,C_j=1)=f(t^*_{ij};\tau_{ci},\beta_j,\alpha_j)\times P(H_i=1)\times P(C_j=1)$$

- An **examinee with item preknowledge** responds to an **uncompromised item** ($H_i = 1$, $C_j = 0$)

$$f(t^*_{ij};\tau_{ti},\tau_{ci},H_i=1,\beta_j,\alpha_j,C_j=0)=f(t^*_{ij};\tau_{ti},\beta_j,\alpha_j)\times P(H_i=1)\times P(C_j=0)$$

- An **examinee with no item preknowledge** responds to a **compromised item** ($H_i = 0$, $C_j = 1$)

$$f(t^*_{ij}; \tau_{ti}, \tau_{ci}, H_i = 0, \beta_j, \alpha_j, C_j = 1) = f(t^*_{ij}; \tau_{ti}, \beta_j, \alpha_j) \times P(H_i = 0) \times P(C_j = 1)$$

- An **examinee with no item preknowledge** responds to an **uncompromised item** ($H_i = 0$, $C_j = 0$)

$$f(t^*_{ij}; \tau_{ti}, \tau_{ci}, H_i = 0, \beta_j, \alpha_j, C_j = 0) = f(t^*_{ij}; \tau_{ti}, \beta_j, \alpha_j) \times P(H_i = 0) \times P(C_j = 0)$$

$$\begin{aligned} f(t^*_{ij};\tau_{ti},\tau_{ci},H_i,\beta_j,\alpha_j,C_j) = {} & f(t^*_{ij};\tau_{ci},\beta_j,\alpha_j) \times P(H_i=1) \times P(C_j=1) + \\ & f(t^*_{ij};\tau_{ti},\beta_j,\alpha_j) \times P(H_i=1) \times P(C_j=0) + \\ & f(t^*_{ij};\tau_{ti},\beta_j,\alpha_j) \times P(H_i=0) \times P(C_j=1) + \\ & f(t^*_{ij};\tau_{ti},\beta_j,\alpha_j) \times P(H_i=0) \times P(C_j=0) \end{aligned}$$

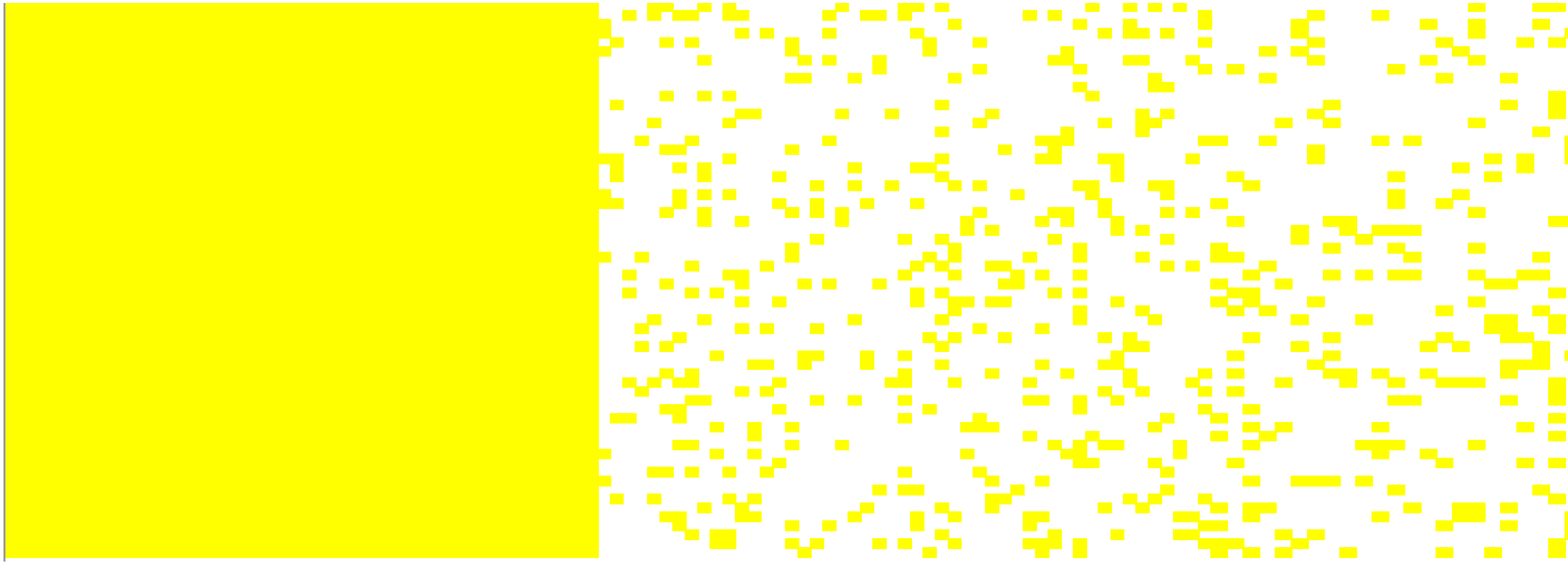
A very simplified version of corresponding model syntax in Stan:

```
for(i in 1:N)
  for (j in 1:I) {

      real p_t = beta[j] - tau_t[i];
      real p_c = beta[j] - tau_c[i];

      real lprt1 = log1m(pC[j]) + log1m(pH[i]) + normal_lpdf(Y[i,j] | p_t, 1/alpha[j]));
      real lprt2 = log1m(pC[j]) + log(pH[i])   + normal_lpdf(Y[i,j] | p_t, 1/alpha[j]));
      real lprt3 =  log(pC[j])  + log1m(pH[i]) + normal_lpdf(Y[i,j] | p_t, 1/alpha[j]));
      real lprt4 =  log(pC[j])  + log(pH[i])   + normal_lpdf(Y[i,j] | p_c, 1/alpha[j]));

      target += log_sum_exp([lprt1, lprt2, lprt3, lprt4]);
  }
}
```

Dataset description

- I used a random sample of 1000 examinees from Form A with 171 items.
- This subset of Form A
  - had 50 operational items that all 1000 examinees responded to, and
  - had 121 pilot items that a different set of 100-150 examinees responded to.
- Each examinee responded to 65 items (50 operational + 15 pilot items)

Data structure was relatively sparse due to missing data by design for the pilot items. Only 38% of the data matrix was complete.



Model Fitting

- The model was fitted using Bayesian estimation through the `rstan` package (Stan Development Team, 2018) in R (R Core Team, 2018).
- There were four chains using 1,000 iterations, and model parameter estimates from posterior densities were calculated using 750 iterations after 250 warm-up iterations.
- An informal post at the link - https://cengiz.me/posts/dglnrt2/ - includes
  - a more detailed description of the model,
  - parameter constraints necessary for model identification,
  - prior specifications, and
  - and two example analysis with all relevant R and Stan code.
- The code for the dataset used in this particular presentation can be found in the following Github repo:

  https://github.com/czopluoglu/dglnrt2/tree/main/R/ncme22/dglnrt

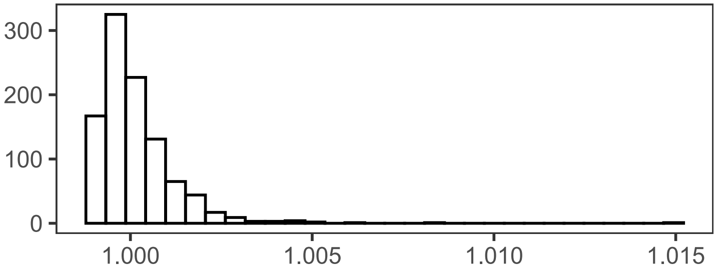
Results

Model Convergence

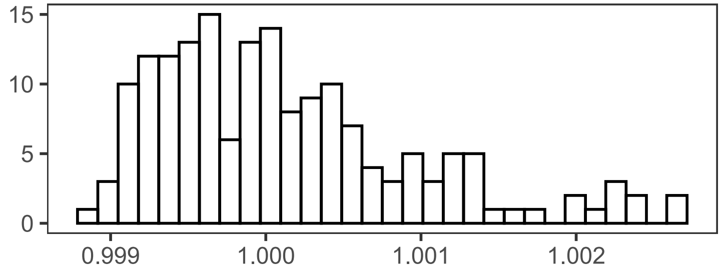
The model convergence was checked by visual inspection of the sampling chains and with the split-chain $\hat{R}$ statistic.

The split-chain $\hat{R}$ statistic was less than 1.01 for most parameters in the model, with a maximum value of 1.032 for one of the item parameters (β_{30}).

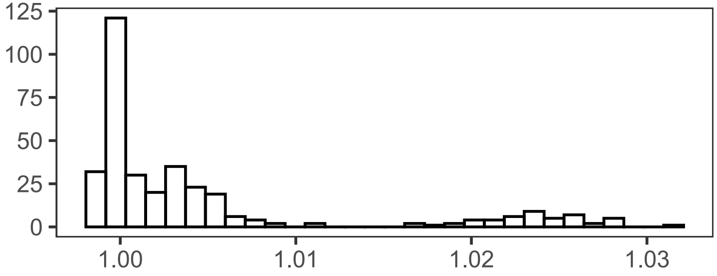
$\hat{R}$ statstics for P(H=1) parameter estimates



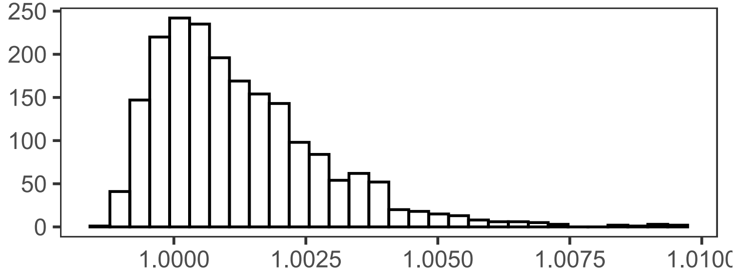
$\hat{R}$ statstics for P(C=1) parameter estimates



$\hat{R}$ statstics for item parameter estimates (α, β)



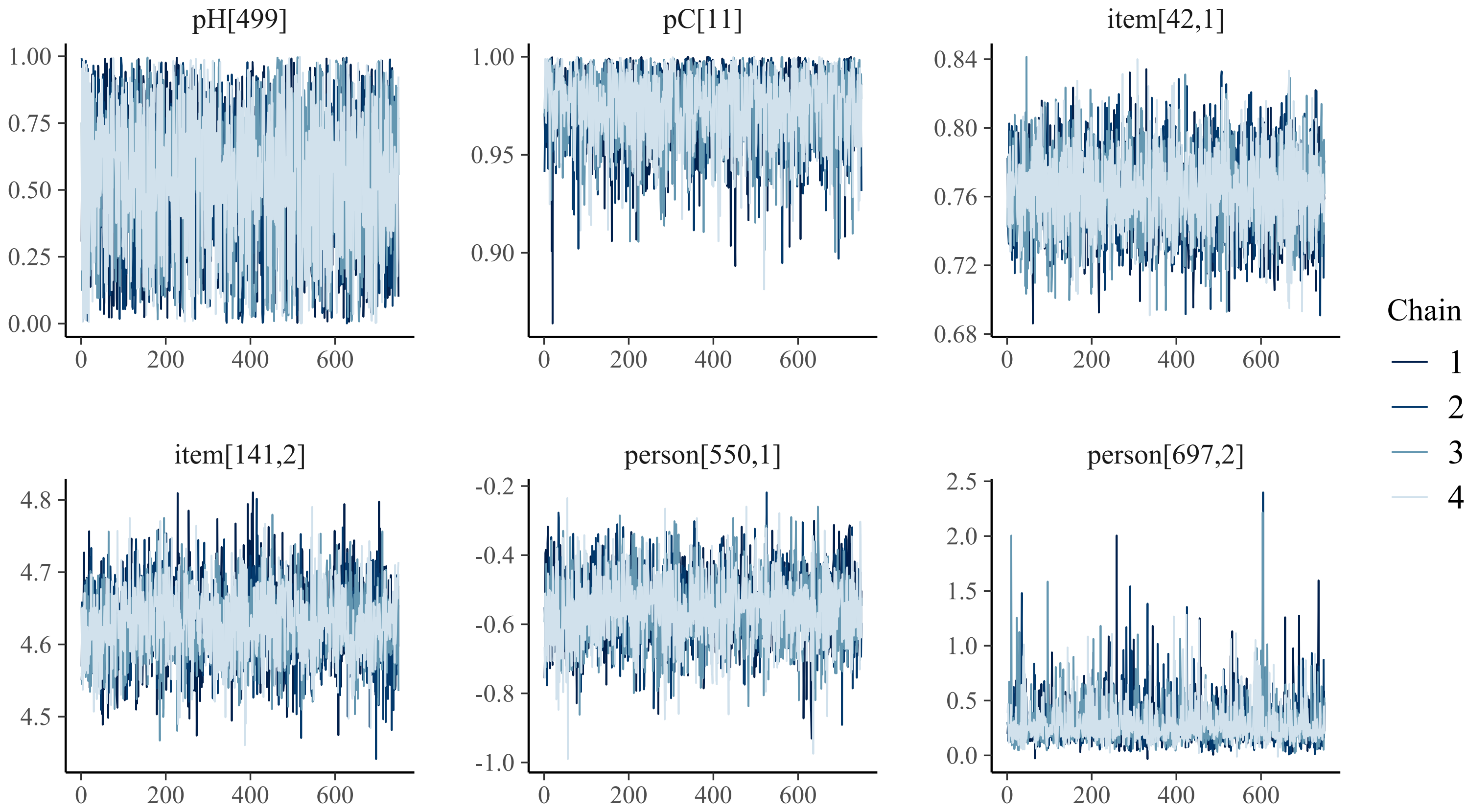
$\hat{R}$ statstics for person parameter estimates (τ_t, τ_c)



Results

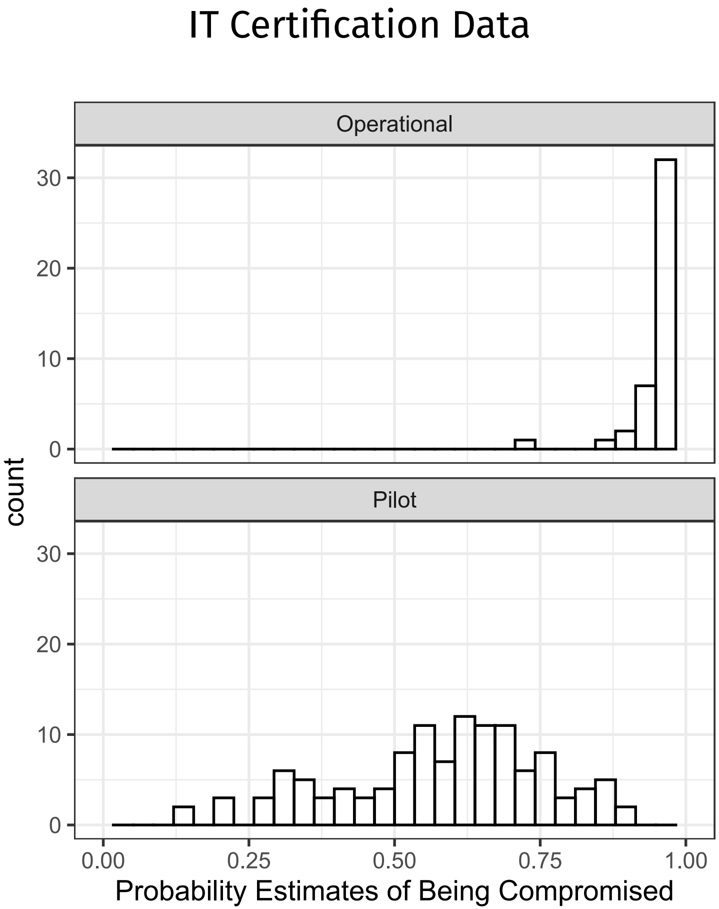
Model Convergence

The inspection of trace plots didn't indicate any pathological behavior during sampling. Below are the trace plots for a random selection from each parameter type.



Results

Probability Estimates of Being Compromised for Operational and Pilot Items

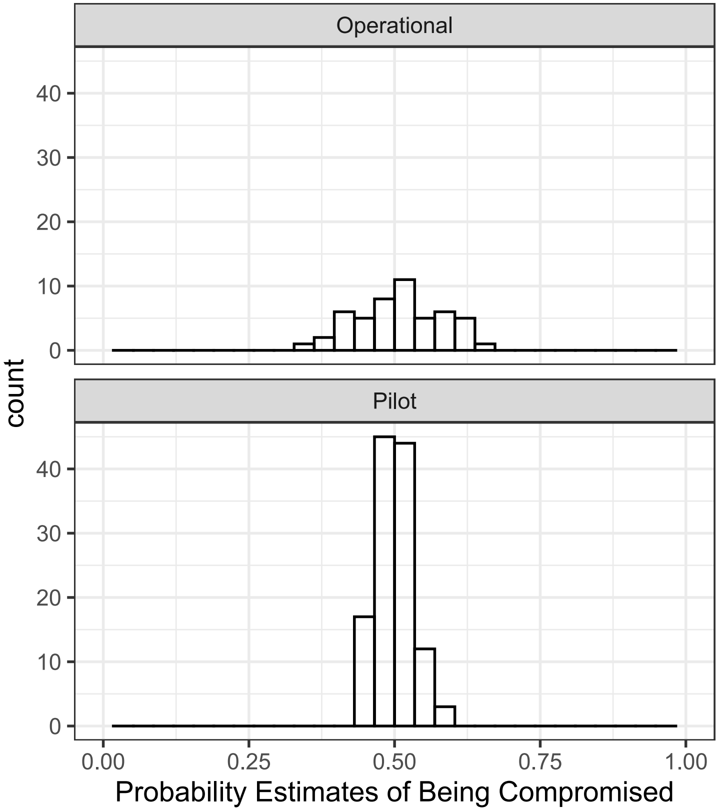


- The probability estimates of being compromised ranged
  - from 0.71 to 0.99 with a mean of 0.96 for **50 operational items**.
  - from 0.13 to 0.91 with a mean of 0.58 for **121 pilot items**.
- If we use 0.91 as a cut-off point, the model indicated that 47 out of 50 operational items were potentially compromised.

Results

Probability Estimates of Being Compromised for Operational and Pilot Items

Simulated Data - Null Condition



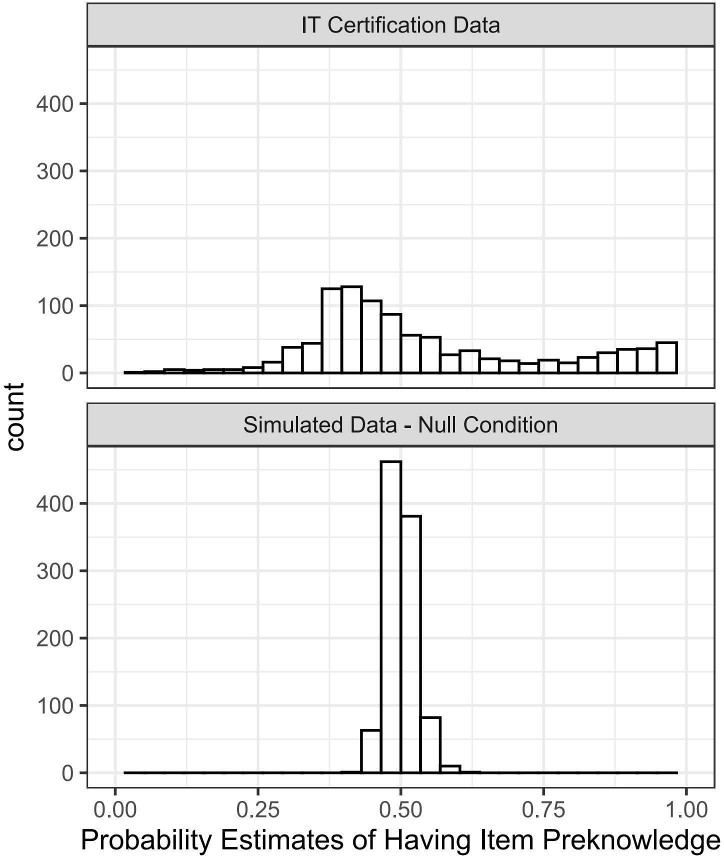
- To observe the model behavior for the same set of parameters of being compromised, we simulated data with no item preknowledge.

- While simulating data, we used the estimates from real data analysis for the item parameters (β, α) and true latent speed parameters (τ_t). The exact structure of missingness was replicated by replacing values with NAs.

- The distribution of the estimates of being compromised for simulated null condition indicated that the model was picking up some signals in the real data for the operational items.

Results

Probability Estimates of Item Preknowledge for Examinees

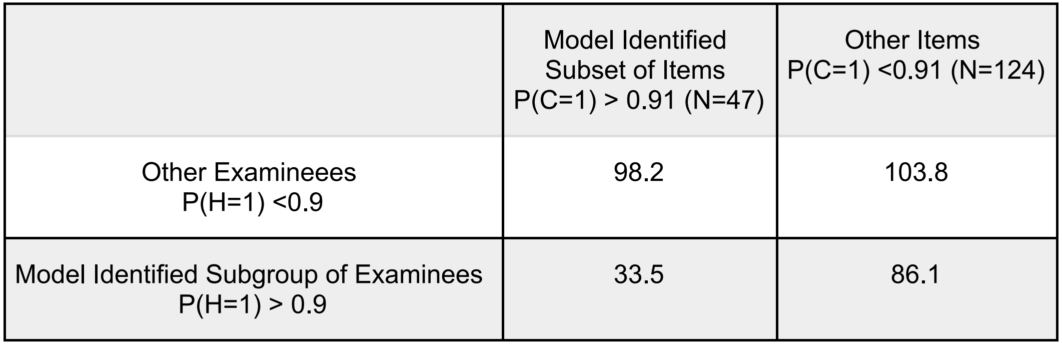


- The probability estimates of having item preknowledge
  - from 0.04 to 0.98 with a mean of 0.54 for 1000 real examinees in the IT Certification dataset.
  - from 0.43 to 0.64 with a mean of 0.58 for 100 simulees in the null condition.
- If we use 0.9 as a cut-off point, the model indicated that 96 out of 1000 examinees had potentially accessed some compromised items before the test.

Results

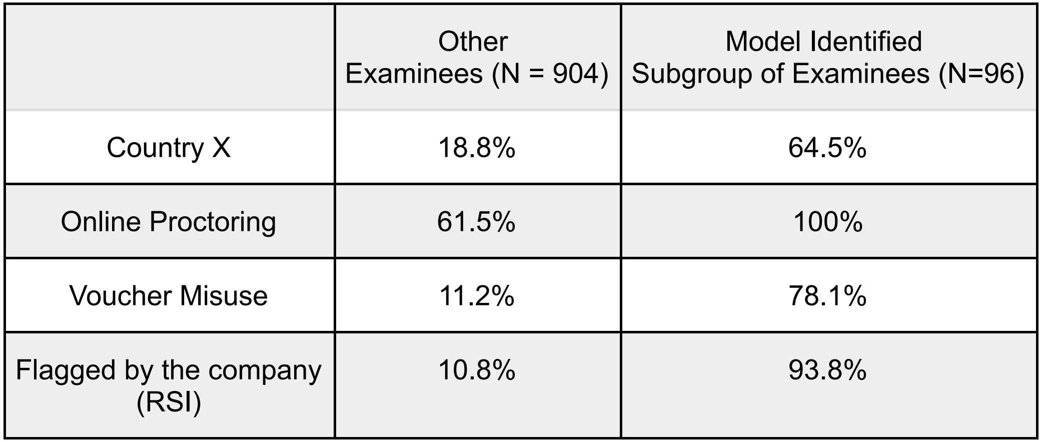
Comparison of Model Identified Subgroup (N = 96) and Others (N=904)

- Average Response Time



	Model Identified Subset of Items P(C=1) > 0.91 (N=47)	Other Items P(C=1) <0.91 (N=124)
Other Examineees P(H=1) <0.9	98.2	103.8
Model Identified Subgroup of Examinees P(H=1) > 0.9	33.5	86.1

- Other characteristics: country, online proctoring, voucher misuse, flagged by company (RSI)



	Other Examinees (N = 904)	Model Identified Subgroup of Examinees (N=96)
Country X	18.8%	64.5%
Online Proctoring	61.5%	100%
Voucher Misuse	11.2%	78.1%
Flagged by the company (RSI)	10.8%	93.8%

Concluding Remarks

- I can argue that a certain subgroup of examinees responded significantly faster to operational items than the rest of the group.
- The proposed model
  - is designed to pick such a signal and were successfully fitted to a random sample from the dataset,
  - successfully separated a particular group of examinees in the data from the rest of the group by estimating the probability of item preknowledge for each examinee,
  - successfully separated operational items from the pilot items by estimating a probability of being compromised for each item.
- In the context of this presentation, I tend to interpret faster response times as an indication of item preknowledge. If there are other plausible explanations for faster response times in operational items for this subgroup of examinees, this inference is void.
- The idea can be extended and used for response accuracy data (work in progress!)
- Response time and response accuracy pieces can be combined. It becomes a very complex model but can potentially yield the highest performance (work in progress!)

Limitations

- There are some, but I am sorry I am running out of time :)

Thank you!

Questions --> cengiz@uoregon.edu