Detection of two-way outliers in multivariate data and application to cheating detection in educational tests

Irini Moustaki
Co-authors: Dr Yunxiao Chen and Ms Yan Lu
London School of Economics and Political Science

Outline

- A brief introduction on latent variable modelling.
- Motivation.
- Proposed model specification
- Assumptions
- Compound Bayesian Decision Theory
- Real application
- Remarks and Extensions

Latent variables and measurement

Using statistical models to understand constructs better: a question of **measurement**

- Many theories in behavioral and social sciences are formulated in terms of theoretical constructs that are not directly observed

  attitudes, opinions, abilities, motivations, etc.

- The measurement of a construct is achieved through one or more observable **indicators** (questionnaire **items**, tests).
- The purpose of a measurement model is to describe how well the observed indicators serve as a measurement instrument for the constructs, also known as **latent variables**.
- **Measurement models** often suggest ways in which the observed measurements can be improved.

Latent variables and substantive theories

Using statistical models to understand relationships between constructs and covariates and to test **theories** about those relationships.

- Often measurement by multiple indicators may involve more than one latent variable.
- Subject-matter theories and research questions usually concern relationships among the latent variables, and perhaps also observed explanatory variables.
- Latent variables can be used as predictors for distal outcomes or as dependent variables explained by covariates.
- These are captured by statistical models for those variables: **structural models**.

Motivation of our work

- Cheating and leaked items are a high stake issue in testing.
- Latent differential item functioning.
- Other areas of application.

Notation

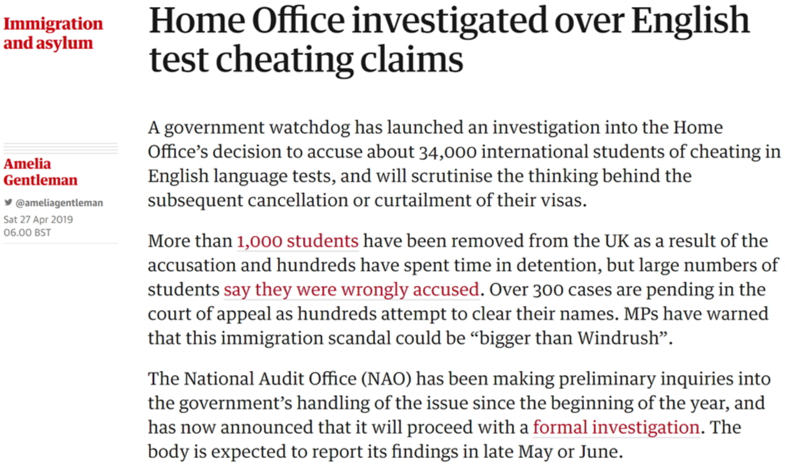
- Let $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$ be the vector of items/manifest/observed variables.
- Let $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_q)$ be the vector of continuous latent variables.
- Let ξ and η be discrete latent variables with a number of classes each.
- Let $\mathbf{x}' = (x_1, x_2, \ldots, x_k)$ be the vector of observed covariates.
- $$\begin{pmatrix} 1 & 0 & 1 & \ldots & 1 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \end{pmatrix}$$

Detection of Cheating and Compromised Items

SAT/ACT Prep
SAT/ACT Fee Waivers & Registration
BEST COLLEGES
2017 EDITION
ACT Coach English, Reading, and Writing
ACT Coach Science
ACT Coach Mathematics
triumphlearning
THE Real ACT Prep Guide
PETERSON'S
The Princeton Review 1,296 ACT Practice Questions
GRE 2011
10 ACT Practice Tests
The Princeton Review 1,296 ACT Practice Questions

1. Many tests are of high-stake.
2. Test fairness is a concern.
3. People cheat (e.g., US college cheating scandal)

Motivating Example: Sensitive Decisions



Immigration and asylum

Home Office investigated over English test cheating claims

Amelia Gentleman

@ameliagentleman

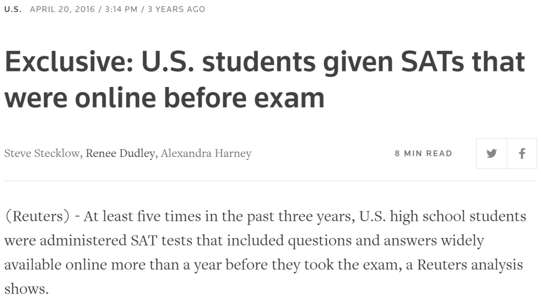
Sat 27 Apr 2019
06.00 BST

A government watchdog has launched an investigation into the Home Office's decision to accuse about 34,000 international students of cheating in English language tests, and will scrutinise the thinking behind the subsequent cancellation or curtailment of their visas.

More than 1,000 students have been removed from the UK as a result of the accusation and hundreds have spent time in detention, but large numbers of students say they were wrongly accused. Over 300 cases are pending in the court of appeal as hundreds attempt to clear their names. MPs have warned that this immigration scandal could be "bigger than Windrush".

The National Audit Office (NAO) has been making preliminary inquiries into the government's handling of the issue since the beginning of the year, and has now announced that it will proceed with a formal investigation. The body is expected to report its findings in late May or June.

Motivation: Fairness in Tests



U.S. APRIL 20, 2016 / 3:14 PM / 3 YEARS AGO

Exclusive: U.S. students given SATs that were online before exam

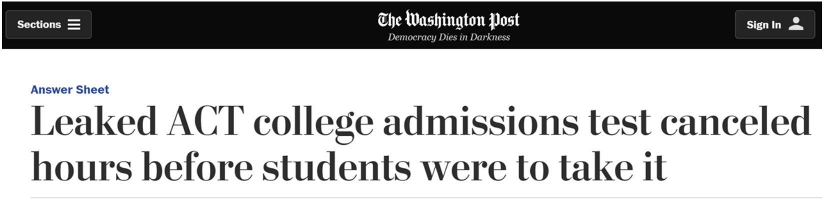
Steve Stecklow, Renee Dudley, Alexandra Harney

8 MIN READ

(Reuters) - At least five times in the past three years, U.S. high school students were administered SAT tests that included questions and answers widely available online more than a year before they took the exam, a Reuters analysis shows.

- Recycling test questions remains a major problem. Questions used in previous tests are available either because they were unofficially acquired by 'test prep' firms, or reconstructed and shared by test-takers on sites.
- The failures to secure the test questions threaten the validity of exam scores.

Motivation: Fairness in Tests



The Washington Post

Democracy Dies in Darkness

Answer Sheet

Leaked ACT college admissions test canceled hours before students were to take it

- Compromised items: leaked test questions
- Cheaters in this context: test-takers who have prior access to compromised questions

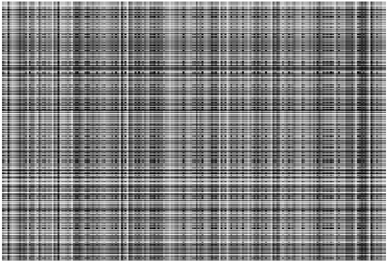
Questions of interest

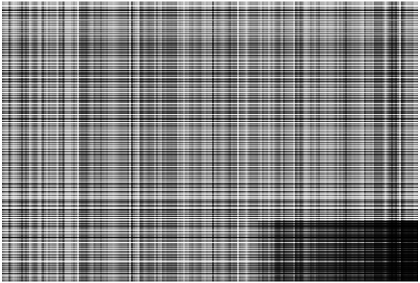
- Who may have cheated? How strong is the evidence?
- Which items may have been leaked? How strong is the evidence?
- Which items should we remove from the item pool? Which test takers should we flag in the database?

Many problems have a similar structure, e.g., detection of fake reviews in online-shopping /video-sharing/movie-rating websites.

Data-driven and Model-based Cheating Detection

Probabilities that test takers (y-axis) correctly answer questions (x-axis)





Data Preview: Real Data Example

- A benchmark dataset from the book *Handbook of Detecting Cheating on Tests*: 1636 test takers' responses to 170 scored items in a standardized test.

- There are approximately 50 candidates, who were flagged by the testing company as likely cheaters. Candidates were flagged through a combination of statistical analysis and a careful investigative process which brought in other pieces of information.

- Based on both data forensics and a careful investigation, the testing program believes that 64 items were leaked.

Descriptive Analysis: Total Score and Item Correct Rates

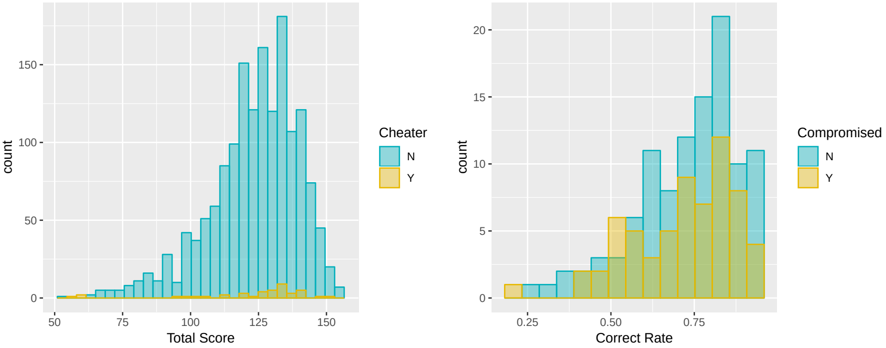


Figure : Panel (a): Histogram of test takers' total scores by the testing program's cheating labels. Panel (b): Histogram of items' correct rates by the testing program's compromisation labels.

Descriptive Analysis: Mean Log-Time

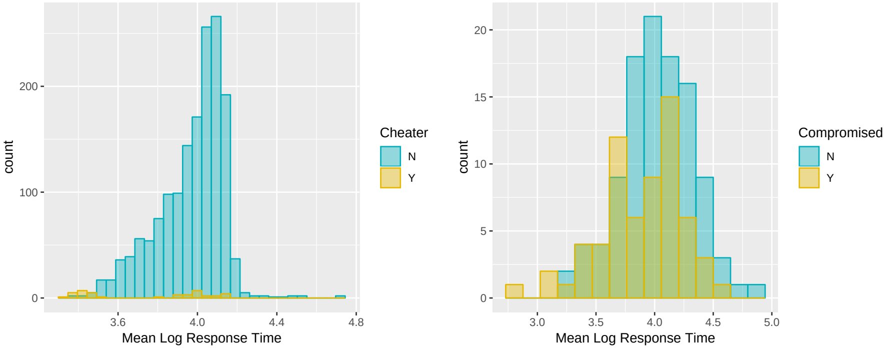


Figure : Panel (a): Histogram of test takers' mean log-time by the testing program's cheating labels. Panel (b): Histogram of items' mean log-time by the testing program's compromisation labels.

Descriptive Analysis: Credentialing dataset.

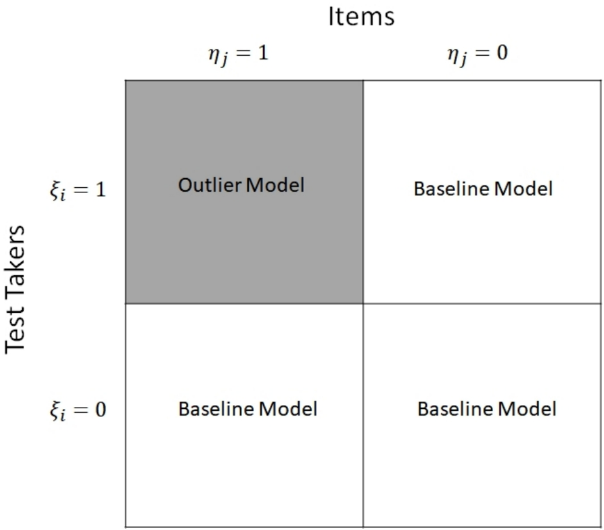
- Summary statistics do not have much information about the labels of the test takers and items.
- The area under the curves (AUC) of the corresponding ROC curves are 55.2% and 71.7% for the in-sample prediction of the cheating labels based on total score and mean log-time, respectively.
- Similarly, the corresponding AUCs for the classification of items are 52.4% and 60.6%, respectively.
- The proposed models are expected to substantially improve upon these benchmarks.

Cheating Detection based on Item Responses and Time information

- Data: Y_{ij}: individual i's responses to item j, $i = 1, ..., N$, $j = 1, ..., p$.
- Data: T_{ij}: individual i's response time to item j, $i = 1, ..., N$, $j = 1, ..., p$.
- We introduce two discrete binary latent variables: ξ_i to classify test takers to cheaters/non-cheaters and η_j to classify items to leaked/non-leaked.

- Some remarks:
  - Item response behavior on a well-designed test is usually explained well by a single-factor model (i.e., unidimensionality).
  - Cheaters are more likely to answer correctly on leaked items.
  - The proportion of cheaters and leaked items should not dominate the analysis.

Models for Cheating



	Items	
Test Takers	$\eta_j = 1$	$\eta_j = 0$
$\xi_i = 1$	Outlier Model	Baseline Model
$\xi_i = 0$	Baseline Model	Baseline Model

Baseline Model

- Rasch model (Rasch, 1960):
  - Item Response Probability:

$$P(Y_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

  - θ_i: an individual-specific latent ability (i.e. test taker's ability).
  - β_j: an item-specific parameter (i.e. item's difficulty level).
  - Assumption of Conditional Independence: : Y_{ij}s are all independent, given θ_1, ..., θ_N, b_1, ..., b_J.
  - Other IRT models such as the 2PL can be used as the baseline model.

Proposed Model: response items

A double mixture Rasch model (or IRT model in general):

- Two types of test takers indicated by a latent variable ξ_i: $\xi_i = 1$ if the test taker cheats and $\xi_i = 0$ otherwise.
- Two types of items indicated by a latent variable η_j: $\eta_j = 1$ if the item is leaked and $\eta_j = 0$ otherwise.
- Item response probability with an additional component δ capturing the effect of cheating:

$$P(Y_{ij} = 1|\theta_i, \beta_j, \xi_i, \theta_j, \delta) = \frac{\exp(\theta_i - \beta_j + \xi_i\eta_j\delta)}{1 + \exp(\theta_i - \beta_j + \xi_i\eta_j\delta)},$$

where $\delta > 0$.

Cheating Detection Model

- The cheating detection model assumes that normal test takers ($\xi_i = 0$) follow a standard Rasch model for all test items:

$$P(Y_{ij} = 1 | \theta_i, b_j, \xi_i = 0, \eta_j, \delta) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

- It also assumes that cheaters follow a standard Rasch model for non-leaked items:

$$P(Y_{ij} = 1 | \theta_i, b_j, \xi_i = 1, \eta_j = 0, \delta) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

- Cheaters follow a Rasch model with reduced difficulty on leaked item:

$$P(Y_{ij} = 1 | \theta_i, b_j, \xi_i = 1, \eta_j = 1, \delta) = \frac{\exp(\theta_i - (b_j - \delta))}{1 + \exp(\theta_i - (b_j - \delta))}.$$

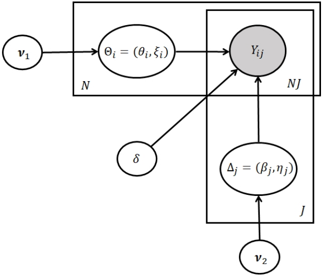
Cheating Detection Model

Hierarchical modeling assumption:

- (θ_i, ξ_i), $i = 1, \ldots, N$, are i.i.d. random vectors from a certain distribution with unknown parameters, denoted by $f_1(\theta, \xi | \boldsymbol{\nu}_1)$
- (b_j, η_j, δ), $j = 1, \ldots, J$, are i.i.d. random vectors from another distribution with unknown parameters, denoted by $f_2(b, \eta, \delta | \boldsymbol{\nu}_2)$.
- $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ are the only unknown parameters in the model (to be estimated from data).

Bayesian Inference

- Parallel Tempering also known as Metropolis-coupled MCMC is used. Parallel Tempering simulates multiple MCMC chains simultaneously and improves the mixing of the low-tempered MCMC chains.
- Parallel Tempering (Geyer, 2011) provides a powerful took for exploring distributions with many local modes.
- A Metropolis-Hasting sampler is used from the MCMC sampling within a chain.


ν_1
$\Theta_i = (\theta_i, \xi_i)$
Y_{ij}
N
NJ
δ
$\Delta_j = (\beta_j, \eta_j)$
J
ν_2

Adding a Response Time Model to the Baseline Model

- Let T_{ij} denote the amount of time test taker i spends to answer item j, $i = 1, ..., N$, $j = 1, ..., p$.
- The baseline model now specifies the distribution of (Y_{ij}, T_{ij}) when there is no cheating.
- We assume that the distribution of Y_{ij} only depends on the parameters in the IRT model.
- The log-normal model is assumed as the baseline response time model,

$$\log(T_{ij})|\tau_i, \alpha_j, \kappa \quad \sim \quad N\left(\alpha_j - \tau_i, \kappa\right), \tag{1}$$

where τ_i as a person-specific speed factor, α_j captures the mean time for completing the item in log-scale and κ captures the variation in the response time across test takers.

Proposed Response Time Model, e.g. van der Linden (2007)

- The proposed item response model introduces the sub-model for response time:

$$\log(T_{ij})|\tau_i, \xi_i, \alpha_j, \eta_j, \gamma \quad \sim \quad N\left(\alpha_j - \tau_i - \xi_i \eta_j \gamma, \kappa\right), \tag{2}$$

  where ξ_i and η_j are the cheating indicators for test takers and items, respectively, and γ is a positive drift parameter, characterizing the reduction in time due to item pre-knowledge.

- The marginal dependence between Y_{ij} and T_{ij} will be introduced by the dependence between the ability and speed factors θ_i and τ_i, and the dependence between the item characteristics β_j and α_j

Item Responses and Time Information

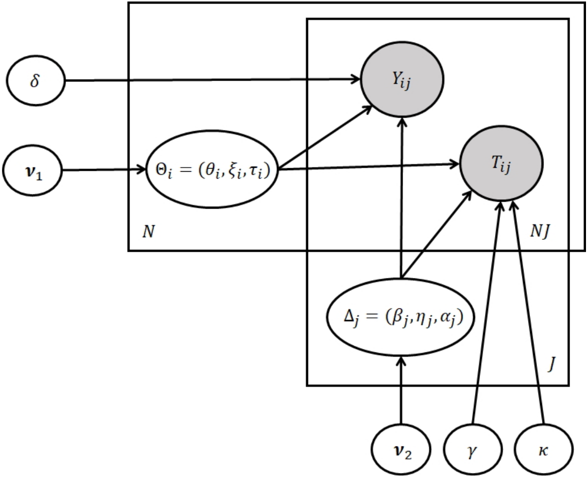


Figure : Graphical representation of the response time model

Bayesian Decision for Cheating Detection

- Assume $\mathbf{Z} = (\mathbf{Y}, \mathbf{T})$
- Which test takers may have cheated? How strong is the evidence?
    - Based on the posterior probabilities $P(\xi_i = 1|\mathbf{z}), i = \ldots, N$
- Which items may have been leaked? How strong is the evidence?
    - Based on the posterior probabilities $P(\eta_j = 1|\mathbf{z}), j = 1, \ldots, p$
- Denote with $\mathcal{D}_i = 1$ when test-taker i is flagged as a cheater and $\mathcal{D}_i = 0$ otherwise.
- Under the proposed model, let $\mathcal{D}_i = 1$ if $\Pr(\xi_i = 1 \mid \mathbf{z}) \geqslant \zeta$.
- Where ζ is a relative cost of a false positive error, $\zeta \in (0, 1)$.
- Bayes risk:

$$R(D_i) = \zeta \underbrace{P(D_i = 1, \xi_i = 0)}_{\text{FP}} + (1 - \zeta) \underbrace{P(D_i = 0, \xi_i = 1)}_{\text{FN}}$$

Compound Decision: General remarks

- How should a test company determine (ζ) the threshold for flagging test takers as potential cheaters?
- What threshold is big enough?
- Trade-off: We would like to detect as many cheaters as possible, but do not want to make many mistakes (flagging innocent test takers as cheaters).
- Possible solution: Compound decision theory (Robbins, 1950)– Using information from all test takers to make individual decisions.
- Solving N decision problems simultaneously.

Compound Decision: Local FDR Control for Persons (1)

- Results can be classified as follows:

	Not flagged as cheater	Flagged as cheater	Total
Non-cheater	N_{00} (TN)	N_{01} (FP)	$N_{0\cdot}$
Cheater	N_{10} (FN)	N_{11} (TP)	$N_{1\cdot}$
Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

Table : A summary of the outcomes of detecting cheaters.

We focus on two quantities:

- The false discovery proportion: $N_{01}/max\{N_{\cdot 1}, 1\}$ - (FDP, the proportion of innocent test takers among the detected ones) estimated by the posterior probability $E(FDP \mid \mathbf{Y}, \mathbf{T})$ and is known as local False Discovery Rate (FDR).
- The false non-discovery proportion: $N_{10}/max\{N_{\cdot 0}, 1\}$ - (FNP, proportion of cheaters among the non-detected ones) estimated by $E(FNP \mid \mathbf{Y}, \mathbf{T})$ and is known as local False Non-Discovery Rate (FNP).

Compound Decision: Local FDR Control for Persons (2)

- Given data and a threshold ζ we can compute both measures.
- The smaller the threshold ζ, the more detections we make, the larger the local FDR.

Compound Decision: Local FDR Control for Persons (3)

- For detecting cheaters, the consequences of false positives (i.e. flagging an innocent test-taker as a cheater) is more serious than that of false negatives (i.e. failing to flag a cheater).
- A sensible decision rule is to minimise the local FNR while controlling the local FDR to be below a pre-specified threshold ρ.
- Intuition: Control the proportion of innocent test-takers misclassified as cheaters.

Compound Decision: Local FDR Control for Persons (2)

- Given a relative cost ζ, the decision on each test taker i is given by $\mathcal{D}_i(\zeta) = 1_{\{P(\xi_i=1|\mathbf{z})>\zeta\}}$.
- The local FDR

$$\mathrm{fdr}_\zeta(\mathbf{z}) = \frac{\sum_{i=1}^{N} D_i(\zeta) P(\xi_i = 0 \mid \mathbf{z})}{\max\left\{\sum_{i=1}^{N} D_i(\zeta), 1\right\}},$$

  which depends only on $P(\xi_i = 0|\mathbf{z})$.
- The local False Nondiscovery Rate (FNR), which is defined as the posterior mean of $N_{10}/N_{\cdot 0}$, can be obtained similarly

$$\mathrm{fnr}_\zeta(\mathbf{z}) = \frac{\sum_{i=1}^{N}(1 - D_i(\zeta)) P(\xi_i = 1|\mathbf{z})}{\max\left\{\sum_{i=1}^{N}(1 - D_i(\zeta)), 1\right\}}$$

.

Compound Decision: Local FDR Control for Persons (3)

The optimization problem: Control the local FDR to be lower than a pre-specified level ρ (e.g. 1%, 5%, 10%), while minimising the local FNR:

$$\min_{\zeta} \mathrm{fnr}_{\zeta}(\mathbf{z}), \quad \textit{subject to.} \;\; \mathrm{fdr}_{\zeta}(\mathbf{z}) \leq \rho.$$

- The local FNR $\mathrm{fnr}_{\zeta}(\mathbf{z})$ is non-decreasing in ζ. Then the optimal threshold
$$\zeta^{*}(\mathbf{z};\rho) = \inf\{\zeta : \mathrm{fdr}_{\zeta}(\mathbf{z}) \leq \rho\}$$
- ρ is easier to specify than the relative cost.

Compound Decision: Local FNR Control for Items

- For detecting compromised items, false negatives (i.e. failing to flag a compromised item) is typically worse than false positives (i.e. a non-compromised items flagged as compromised).
- Intuition: Control the quality of the remaining items, while not removing too many items.
- The optimization problem: Control the local FNR to be below a pre-specified level ρ (e.g. 1%, 5%, 10%) and in the meantime minimise the local FDR.

Real Data Example

- A benchmark dataset from the book *Handbook of Detecting Cheating on Tests*: 1636 test takers' responses to 170 scored items in a standardized test.

- We removed 12 test takers who had zero response times in regard to one or more items. 5 out of 12 people with a response time of zero were flagged as cheaters and the rest 7 were labelled as non-cheaters according to the test company.

- The remaining 1624 test takers' item responses were used. 41 out of 1624 examinees were flagged as potential cheaters (2.52%), and 64 of 170 items were suspected to get leaked (37.6%).

Bayesian Inference

- Compare models $\mathcal{M}_1$ (only for item responses and a cheating component) and $\mathcal{M}_1^0$ (only item responses) by DIC.
- The DIC values for the two models are 138,282.6 and 218,308.4 respectively.
- Item pre-knowledge is likely to exist among test takers.

Model $\mathcal{M}_1$: Posterior Means, response items only

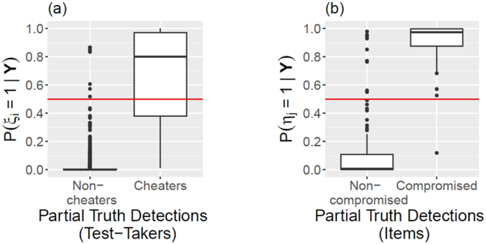


Figure : Boxplots of the posterior means of (a) ξ_i for the cheating and non-cheating groups and (b) η_j for the compromised and non-compromised items (defined by the testing program)

Model $\mathcal{M}_1$, response items only

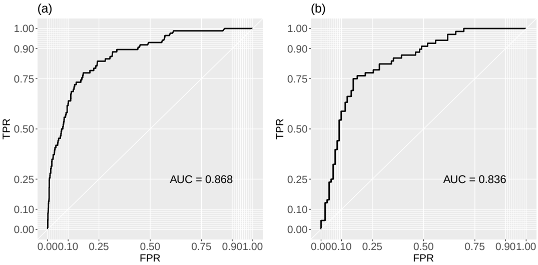


Figure : Prediction of cheaters and compromised items by the posterior means of ξ_i and η_j respectively (labeled by the testing program).

$\mathcal{M}_1$: Local FDR and local FNR plots for (a) individuals and (b) items

The number of detections increases, the local FDR increases and the local FNR decreases.

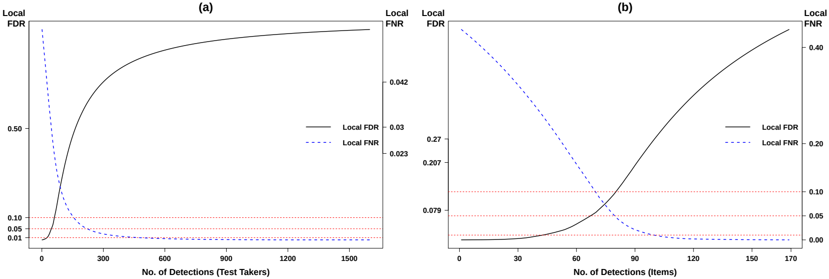


Figure : The local FDR and the local FNR as functions of the number of detections.

$\mathcal{M}_1$: Number of detections

Table : The first row shows the numbers of detections for test takers, when controlling the corresponding local FDR at 1%, 5%, and 10% levels, respectively. The second row shows the numbers of detections for items, when controlling the corresponding local FNR at 1%, 5%, and 10% levels, respectively.

	1%	5%	10%
Test takers	25	46	61
Items	100	91	71

Model $\mathcal{M}_1$: posterior means of parameters, item responsed only

Table : The row labelled "EAP" shows the posterior means of the global parameters, where EAP represents for Expected A Posteriori, and the row labelled "95% CI" provides the corresponding 95% credible intervals.

	σ_{11}	π_1	π_2	ω_{11}	μ_1	δ
EAP	0.285	0.028	0.401	0.685	-1.004	0.895
95% CI	(0.261, 0.319)	(0.020, 0.036)	(0.387, 0.433)	(0669, 0.854)	(-1.237, -0.912)	(0.758, 0.959)

- Test Company Results: 2.8% examinees were flagged as potential cheaters and 40.1% of items were suspected to get leaked.
- $\hat{\delta} = 0.895$. The odds ratio of correctly answering a compromised item is about $\exp(0.895) = 2.447$ when comparing a cheater and a non-cheater with the same ability level.

Model $\mathcal{M}_2$: Detection based on Item Responses and Response Times

- DIC for three models: We compare model $\mathcal{M}_2$ (response time is included) and model $\mathcal{M}_2^0$ where only cheating is taken into account. The corresponding DIC values are $\mathcal{M}_2$ =176,935.2 and $\mathcal{M}_2^0$=214,201.3.
- The ROC curves based on the posterior means of ξ_i and η_j have AUC 0.892 and 0.867, respectively, where the AUC values are slighter higher than from $\mathcal{M}_1$.
- More detections tend to be made under the model $\mathcal{M}_2$, especially at the lower thresholds 1% and 5%. This is likely due to that the posterior distributions tend to be more concentrated under model $\mathcal{M}_2$ as it utilizes more information.

Table : Applying model $\mathcal{M}_2$ to credentialing dataset. The first row shows the numbers of detections for test takers, when controlling the corresponding local FDR at 1%, 5%, and 10% levels, respectively. The second row shows the numbers of detections for items, when controlling the corresponding local FNR at 1%, 5%, and 10% levels, respectively.

	1%	5%	10%
Test takers	26	47	65
Items	101	89	74

Posterior means and 95% credible intervals for the global parameters of model $\mathcal{M}_2$

- The estimated correlation between the ability and speed factors is as high as 0.410. This result indicates that test takers with higher ability tend to answer the items faster.
- The correlation between the two item-specific parameters is 0.237. This positive correlation suggests that solving more difficult items tends to take more time, which is consistent with our intuition.

	σ_{11}	π_1	π_2	ω_{11}	μ_1	δ
EAP	0.289	0.027	0.410	0.699	-0.867	0.807
95% CI	(0.259, 0.298)	(0.022,0.036)	(0.365, 0.432)	(0.626, 0.789)	(-0.993, -0.795)	(0.732, 0.852)

	σ_{22}	σ_{12}	ω_{22}	ω_{12}	μ_2	γ
EAP	0.248	0.110	0.397	0.125	-0.472	0.620
95% CI	(0.213,0.285)	(0.0986, 0.139)	(0.334, 0.427)	(0.082, 0.132)	(-0.879, -0.291)	(0.451, 0.907)

	κ
EAP	0.802
95% CI	(0.589, 1.037)

Remarks and Extensions

- The model is robust to certain model mis-specifications (linear predictor, (θ_i, ξ_i) being correlated, different drift δ for items).
- The model applies to one type of cheating behavior (preknowledge due to item leakage).
- Add covariates.
- Apply to other data from education as well from psychological measurement.
- Explore other types of estimation.
- Model fit and model selection.
- Relax some of the assumptions.

References

- Chen, Y., Lu, Y. and Moustaki, I. (2022) Detection of two-way outliers in multivariate data and application to cheating detection in educational tests.*Annals of Applied Statistics*.

Thank you for your attention