

Detection of Two-Way Outliers in Multivariate Data and Application to Cheating Detection in Educational Tests

Yunxiao Chen, Yan Lu, and Irini Moustaki

London School of Economics and Political Science

Abstract

The paper proposes a new latent variable model for the simultaneous (two-way) detection of outlying individuals and items for item-response-type data. The proposed model is a synergy between a factor model for binary responses and continuous response times that captures normal item response behaviour and a latent class model that captures the outlying individuals and items. A statistical decision framework is developed under the proposed model that provides compound decision rules for controlling local false discovery/nondiscovery rates of outlier detection. Statistical inference is carried out under a Bayesian framework, for which a Markov chain Monte Carlo algorithm is developed. The proposed method is applied to the detection of cheating in educational tests due to item leakage using a case study of a computer-based nonadaptive licensure assessment. The performance of the proposed method is evaluated by simulation studies.

KEY WORDS: Bayesian hierarchical model, outlier detection, false discovery rate, compound decision, test fairness, item response theory, latent class analysis

1 Introduction

Factor models (Bartholomew et al., 2011) are widely used to analyse multivariate data, especially item-response-type data which involve individuals' responses to a set of items.

For example, in educational testing, unidimensional and multidimensional factor models, which are also known as Item Response Theory (IRT) models (Embretson and Reise, 2000; Reckase, 2009), are commonly used to model test takers’ responses to test items. In these applications, a latent factor is often interpreted as the ability that the test is designed to assess. In psychology, multidimensional factor models are typically used to describe respondents’ answers to items in a psychological questionnaire (Wirth and Edwards, 2007), where the factors are interpreted as psychological traits (e.g., personality traits). In political science, similar models, which are often known as idea point models, are used to describe voting behaviours (Bafumi et al., 2005), where the factors are typically interpreted as voters’ political standing.

It is often the case that real data contain outliers among both the individuals and the items. Such outliers can lead to a substantial deviation from a carefully specified factor model which may be supported by substantive theory and historical data. These outliers often provide insights about the data and it is thus of substantive interest to detect them. One example is Differential Item Functioning (DIF; Holland and Wainer, 1993; Millsap, 2012), a phenomenon that is widely observed in educational testing, psychological measurement, as well as other related areas. It happens when a subset of items do not measure subgroups (e.g., gender, race) in the same way. Specifically, in educational testing, it might be the case that a subset of items are “easier” or “more difficult” for a certain subgroup than the others. In this case, the subgroup of test takers and the subset of items may be viewed as outliers, for whom a factor model that fits the rest of the data may fail. A related, while more challenging, problem is the detection of latent DIF (Cho et al., 2016), for which not only the DIF items but also the group membership of individuals are not known a priori (two-way detection of outliers). One such example is the detection of cheating in educational tests due to item leakage that benefits test takers through the preknowledge of leaked items (Cizek and Wollack, 2017). Latent DIF may also exist in educational tests due to other reasons; see Cho et al. (2016) for a review. Similar problems also occur in

other areas besides educational testing. For example, in political science, it has been well recognised that roll call voting data of the United States Congress can largely be described by a liberal-conservative latent dimension with some minor deviations (Poole and Rosenthal, 1991; Poole et al., 1991). Through a two-way outlier-detection formulation, i.e., by detecting outlying legislators and roll calls that do not fit the unidimensional model, one may obtain a better understanding of the patterns of roll call voting that cannot be explained by the liberal-conservative dimension. Latent DIF may also exist in psychological measurement data in which a two-way outlier-detection formulation can facilitate the discovery of minor psychological traits and the relevant groups.

While statistical methods have been well established for the detection of DIF (Millsap, 2012), models and procedures for detecting latent DIF (two-way outliers) remain to be developed. We therefore propose a two-way outlier detection model to fill the gap that adds a latent class model component upon a factor model. The factor model component serves as a baseline model for data without outliers, and a latent class model component is used to capture the two-way outliers. Specifically, the proposed model imposes latent class structures among both the individuals and the items, rather than only assuming latent classes among the individuals as in the classical latent class analysis (Allman et al., 2009; Goodman, 1974; Lazarsfeld and Henry, 1968). The proposed model is closely related to, but also substantially different from, existing statistical models and methods for the detection of outliers in multivariate data (see e.g., Candès et al., 2011; Hadi, 1992; Mavridis and Moustaki, 2008,0; Reiser, 1996; Wang and Xu, 2015; Wang et al., 2018; Zhou et al., 2010)

Under the proposed model, statistical decision theory is established for the detection of two-way outliers. Motivated by compound decision theory for multiple testing (Benjamini and Hochberg, 1995; Efron, 2004,0,1; Efron et al., 2001; Robbins, 1951; Sun and Cai, 2007; Zhang, 2003), we propose the local False Discovery Rate (FDR) and local False Non-discovery Rate (FNR) as compound risk measures for the detection of two-way outliers. Decision rules are developed based on these measures, for which optimality results are established. The

statistical inference and decision making are performed under a fully Bayesian framework, for which a Markov chain Monte Carlo (MCMC) algorithm is developed. Since our model involves many discrete latent variables, standard MCMC algorithms such as Gibbs and Metropolis-Hastings can suffer from slow mixing (e.g., Celeux et al., 2000; Richardson and Green, 1997). We tackle this problem by applying the parallel tempering technique (Geyer, 2011).

The proposed method is applied to cheating detection based on data from the single administration of a non-adaptive test. It simultaneously detects outlying test takers and items as potential cheaters and compromised items. The proposed method uses item response data and item response time data, which are often collected in computer-based testing, for improving outlier detection. As shown via our real data analysis and simulation studies, incorporating response time information can improve outlier detection accuracy. Our simulation results further suggest that the proposed model is quite robust against various forms of model misspecification, even though it relies on some parametric assumptions that may not be satisfied perfectly in practice.

The detection of test takers who benefit from item preknowledge (cheaters) and compromised items has received much attention among quantitative researchers in education. Specifically, McLeod et al. (2003) proposed a person-fit index for the detection of cheaters in computerised adaptive testing, under an IRT model. For non-adaptive testing, Belov (2013) proposed a person-fit index for characterising the outperformance of a student on the compromised items, assuming that the set of compromised items is known. Under a similar setting, Sinharay (2017a) proposed likelihood-ratio and score tests for the detection of cheaters, and Segall (2002) and Shu et al. (2013) proposed IRT models for item preknowledge and developed Bayesian classification procedures. For the detection of compromised items, O’Leary and Smith (2017) and Wang and Liu (2020) proposed methods based on data from the single administration of a non-adaptive test. These approaches require knowledge of a subset of non-compromised items to first identify a set of potential cheaters. The detection of

compromised items relies on the identified cheaters in the first stage. Under an online setting where data from multiple tests are sequentially collected, Veerkamp and Glas (2000), Zhang (2014), Chen and Li (2019), and Chen et al. (2020) formulated the detection of compromised items as a sequential change detection problem and proposed sequential procedures. We refer the readers to three edited volumes, Wollack and Fremer (2013), Kingston and Clark (2014) and Cizek and Wollack (2017), for a comprehensive review of related works. Note that most of the existing methods focus on the detection of either cheaters or compromised items, and often require prior information which is not always available, for example a given subset of non-compromised items. In contrast, the proposed method can simultaneously detect both test takers with item preknowledge and compromised items without such prior information.

The rest of the paper is organised as follows. In Section 2, we propose a statistical model for detecting two-way outliers in multivariate data and discuss its application to cheating detection. Statistical decision theory is developed under a Bayesian framework in Section 3, and Bayesian inference procedures are given in Section 4. The proposed method is applied to a real dataset from a licensure test in Section 5. Simulation studies are presented in Section 6 to further evaluate the performance of the proposed method under various situations. Concluding remarks are provided in Section 7. The appendix contains the proof of a theoretical result, the details of the developed MCMC algorithm, and additional simulation results.

2 Proposed Two-Way Outlier Detection Model

2.1 A Two-Way Outlier Detection Model for Multivariate Data

2.1.1 Background and Notation

Consider N individuals responding to J items. Let Y_{ij} be individual i 's response to item j . We focus on binary responses, i.e., $Y_{ij} = 0, 1$, where the two types of responses may

correspond to incorrect and correct answers in educational testing, and “no” and “yes” responses in psychological measurement, among others. We use $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})$ to denote the response vector from individual i and use $\mathbf{Y} = (Y_{ij})_{N \times J}$ to denote the response matrix. When item-response data are collected digitally rather than by paper and pencil, which is becoming more and more popular these days, response time data may also be collected. Let T_{ij} denote the amount of time individual i spends to answer item j and $\mathbf{T} = (T_{ij})$ denote the data matrix for response times.

In what follows, we discuss the two-way outlier detection model for (\mathbf{Y}, \mathbf{T}) . We introduce a latent binary variable ξ_i that takes the value 1 when individual i is an outlier and 0 otherwise. Similarly, η_j is a latent binary variable that takes the value 1 when item j is an outlier and 0 otherwise. Figure 1 illustrates how data are affected by the two-way outliers in the proposed model. (Y_{ij}, T_{ij}) are modelled with the outlier model if, and only if, both $\xi_i = 1$ and $\eta_j = 1$, which represents a typical phenomenon of latent DIF. In the cheating detection application, items with $\eta_j = 1$ correspond to the leaked/compromised items and individuals with $\xi_i = 1$ correspond to test takers who have preknowledge about all the compromised items before taking the test. In this context, a baseline model will capture the normal item-response behaviour, and the outlier model will capture the behaviour of the test takers with preknowledge when responding to the compromised items. In particular, the outlier model will allow test takers to have a higher probability of answering leaked items correctly and with a shorter response time (see, e.g. Wang et al., 2018). The proposed model is described below.

2.1.2 Proposed model

We start with a relatively more general model and then give specific examples. We introduce $\boldsymbol{\theta}_i$ and $\boldsymbol{\tau}_i$ as the person-specific parameters, also known as the factors, that drive the item responses and the response times, respectively. Both $\boldsymbol{\theta}_i$ and $\boldsymbol{\tau}_i$ can be unidimensional or multidimensional, but their dimensions are typically assumed to be much smaller than J .

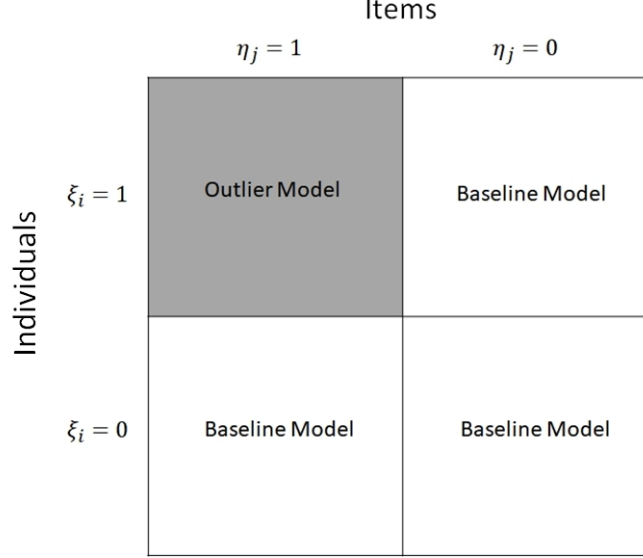


Figure 1: An illustration of the two-way outlier structure in the proposed model.

We also introduce β_j and α_j to denote the item-specific parameters for the item responses and the response times, respectively. To simplify the notation, we use $\Theta_i = (\theta_i, \xi_i, \tau_i)$ and $\Delta_j = (\beta_j, \eta_j, \alpha_j)$ to denote the person- and item-specific parameter vectors, respectively.

The proposed model consists of two submodels, one for binary item responses and one for continuous response times. The item-response submodel takes the following logistic form

$$P(Y_{ij} = 1 | \Theta_i, \Delta_j, \delta) := p(\Theta_i, \Delta_j, \delta) = \frac{\exp(h_1(\theta_i, \beta_j) + \xi_i \eta_j \delta)}{1 + \exp(h_1(\theta_i, \beta_j) + \xi_i \eta_j \delta)},$$

where δ is a non-negative parameter and $h_1(\cdot, \cdot)$ is a pre-specified function. Given $\xi_i = 0$ or $\eta_j = 0$,

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(h_1(\theta_i, \beta_j))}{1 + \exp(h_1(\theta_i, \beta_j))}$$

is the baseline item-response submodel for non-outlying item responses. When $\xi_i = \eta_j = 1$, the term $\xi_i \eta_j \delta \neq 0$ captures the deviation from the baseline model. In particular, for our application, the parameter δ is set to be non-negative to let the probability of providing a correct answer (i.e., $Y_{ij} = 1$) increase when individual i and item j are outliers. In this context, δ may be interpreted as the advantage that a test taker gains from item pre-knowledge.

In other applications, this sign constraint can be removed if such prior information is not available. To keep the model parsimonious, the parameter δ is assumed to be the same across all the outlying individuals and items. As will be discussed in Section 2.2, this assumption can be relaxed.

The function h_1 should be chosen based on knowledge about the baseline model from substantive theory and/or historical data. We give two parametric examples of h_1 below, but point out that h_1 can also take a non-parametric form as in non-parametric IRT models (Douglas, 1997; Duncan and MacEachern, 2008; Ramsay and Winsberg, 1991).

Example 1. *The Rasch model (Rasch, 1960) is one of the most popular IRT models in educational testing, and is also widely used in many other areas. In particular, the licensure test to be studied in Section 5 is designed and scored under this model. The Rasch model assumes that both $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_j$ are unidimensional. With slight abuse of notation, we denote them by non-bold typeface θ_i and β_j , respectively. This model assumes that $h_1(\theta_i, \beta_j) = \theta_i - \beta_j$, which leads to*

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(\theta_i - \beta_j + \xi_i \eta_j \delta)}{1 + \exp(\theta_i - \beta_j + \xi_i \eta_j \delta)}. \quad (1)$$

In the context of educational testing, θ_i and β_j are interpreted as the ability of test taker i and the difficulty of item j , respectively. When there are no outliers, the probability of correctly answering an item is monotone increasing with one's ability θ_i and monotone decreasing with the item's difficulty β_j . When $\xi_i = 1$ and $\eta_j = 1$, $\text{logit}(P(Y_{ij} = 1 | \Theta_i, \Delta_j, \delta)) = \theta_i - \beta_j + \delta$. That is, the item response function still takes a Rasch form, but the log odds increases by a constant drift δ . This Rasch-type item-response submodel (1) will be further discussed in the rest of the paper, given its suitability for our case study in Section 5.

Example 2. *It may be the case that the J items simultaneously measure K factors, $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iK})$, for which a multidimensional factor model is needed. In that situation, we may set $h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j) = \beta_{j0} + \beta_{j1}\theta_{i1} + \dots + \beta_{jK}\theta_{iK}$, where $\boldsymbol{\beta}_j = (\beta_{j0}, \dots, \beta_{jK})$ contains $K + 1$ item-specific parameters. The item-response submodel then becomes*

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(\beta_{j0} + \beta_{j1}\theta_{i1} + \cdots + \beta_{jK}\theta_{iK} + \xi_i\eta_j\delta)}{1 + \exp(\beta_{j0} + \beta_{j1}\theta_{i1} + \cdots + \beta_{jK}\theta_{iK} + \xi_i\eta_j\delta)}. \quad (2)$$

When $\xi_i = 0$ or $\eta_j = 0$, (2) becomes the multidimensional two-parameter logistic model (Reckase, 2009) which includes the two-parameter logistic model (Birnbaum, 1968) as a special case when $K = 1$.

The response-time submodel is specified similarly to the item-response submodel. Specifically, we consider a log-normal model which assumes that

$$\log(T_{ij}) | \Theta_i, \Delta_j, \gamma, \kappa \sim N(h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j) - \xi_i\eta_j\gamma, \kappa),$$

where γ is another non-negative parameter that plays a similar role to δ in the item-response submodel, $h_2(\cdot, \cdot)$ is a pre-specified function, and $\kappa > 0$ is the variance of the normal distribution. In our application, the parameter γ is set to be non-negative to allow the response time for outlying individuals and items to be shorter (test takers with preknowledge tend to answer the compromised items faster). That is, when $\xi_i = 1$ and $\eta_j = 1$, the mean log-time is reduced from the baseline level $h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j)$ to $h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j) - \gamma$. In that context, γ may be interpreted as the reduction in response time due to item preknowledge. Similar to the discussion about parameter δ , the sign constraint on γ can also be removed if there is no such prior knowledge about the response times. We assume that the same γ and κ are shared by all the individuals and items for model parsimony, which can be relaxed.

The choice of function h_2 is similar to the choice of function h_1 in the item-response submodel. In what follows, we give a specific example, but also point out that other choices of h_2 are possible. In particular, one can choose h_2 so that the baseline response-time submodel is consistent with the one proposed in van der Linden (2007).

Example 3. *Similar to the Rasch-type model in Example 1, we let both $\boldsymbol{\tau}_i$ and $\boldsymbol{\alpha}_j$ be unidimensional and denote them by non-bold typeface τ_i and α_j . We let function h_2 take the*

form $h_2(\tau_i, \alpha_j) = \alpha_j - \tau_i$, which leads to

$$\log(T_{ij})|\Theta_i, \Delta_j, \gamma, \kappa \sim N(\alpha_j - \tau_i - \xi_i \eta_j \gamma, \kappa). \quad (3)$$

When $\xi_i = 0$ or $\eta_j = 0$, we obtain the baseline model for response times

$$\log(T_{ij})|\Theta_i, \Delta_j, \gamma, \kappa \sim N(\alpha_j - \tau_i, \kappa).$$

In the context of educational testing, τ_i can be interpreted as the speed factor of test taker i and α_j can be interpreted as the time-consumingness of item j . When there are no outliers, the mean response time is monotone increasing with the item-specific time-consumingness α_j and monotone decreasing with the person-specific speed factor τ_i . This response-time submodel will be applied to our case study in Section 5.

Like many other latent variable models, conditional independence assumptions are imposed. We first assume that (Y_{ij}, T_{ij}) , $j = 1, \dots, J$, are conditionally independent given Θ_i , Δ_j , δ , γ , and κ . Such a conditional independence assumption across items is often known as the local independence assumption. We further assume that Y_{ij} and T_{ij} are conditionally independent given Θ_i , Δ_j , δ , γ , and κ , meaning that all the person effects on the response and response time distribution are captured by the person parameters. Such conditional independence assumptions are commonly made in latent variable models for item responses and response times. We refer the readers to van der Linden (2007) for the substantive justifications.

We further adopt a Bayesian hierarchical modelling framework, under which parameters Θ_i , Δ_j , δ , γ , and κ are treated as random variables. Specifically, we let Θ_i , $i = 1, \dots, N$, be independent and identically distributed (i.i.d.) samples from distribution $g_1(\Theta|\boldsymbol{\nu}_1)$ and Δ_j , $j = 1, \dots, J$, be i.i.d. samples from distribution $g_2(\Delta|\boldsymbol{\nu}_2)$, respectively, where g_1 and g_2 characterise the population of individuals and the domain of items, respectively. Both g_1 and g_2 are taken to be parametric distributions and use $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ as generic notations

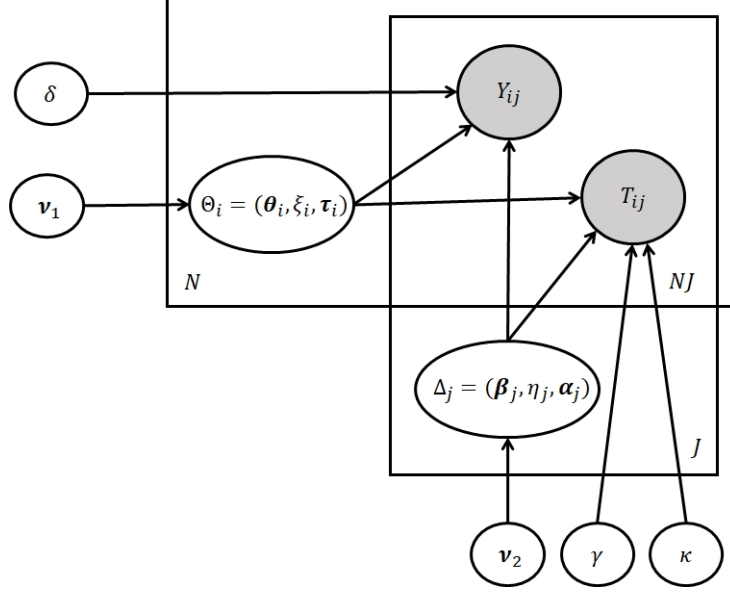


Figure 2: Graphical representation of the proposed model for the joint distribution of item responses and response times. The boxes are plates representing replicates. The two outer plates represent individuals and items, respectively, and the inner plates present item response and response time, respectively.

for the hyperparameters of the two distributions, respectively. This hierarchical modelling structure is visualised in Figure 2 using a graphical model representation. We showcase the specification of g_1 , g_2 , and the priors for ν_1 , ν_2 , δ , γ , and κ in Section 4.1 under the specific model with item-response submodel (1) and response-time submodel (3).

2.1.3 Model without response time data

Sometimes, response time information is not collected, for example, in paper-and-pencil-based educational tests. In that case, response times are missing completely at random and the proposed model reduces to a model for item responses. This reduced model only contains parameters from the item-response submodel and the corresponding hyperpriors. The graphical representation of this reduced model is given in Figure 3, where the reduced person and item parameters are denoted by $\Theta_i = (\theta_i, \xi_i)$ and $\Delta_j = (\beta_j, \eta_j)$, respectively, and the corresponding hyperparameters are still denoted by ν_1 and ν_2 , respectively.

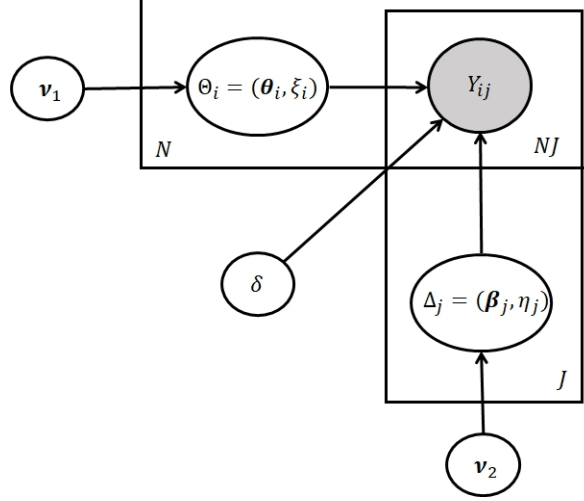


Figure 3: Graphical representation of the reduced model when all the response times are missing completely at random. The boxes are plates representing replicates. The two outer plates represent individuals and items, respectively, and the inner plate presents an item response.

2.1.4 Application to the detection of item preknowledge

The proposed model framework requires that the baseline model is correctly specified. Any deviation from it is solely attributed to item preknowledge and not to other aberrant situations, such as more than one latent dimension (multidimensionality) needed to explain the associations among the items. Although this assumption might appear to be strong, it can still be examined using historical test data with no leaked items and test takers from the same population.

Furthermore, the validity of the model interpretation also depends on the extent to which our parametric assumptions hold. Section 2.2 discusses how some of those parametric assumptions can be violated in practice and can be also relaxed. In addition, as shown via simulation studies in Section 6, the proposed two-way outlier detection model tends to be robust against several forms of model misspecification.

Finally, we emphasise that, given the sensitivity of decisions regarding cheating in tests and the relatively strong assumptions of the proposed model, the latent classes resulting from the two-way detection should be interpreted with caution (i.e. leaked items and test

takers with pre-knowledge). Results from our model can provide warnings to the test administrators, but the detected outlying cases should be further investigated and verified using additional sources of information.

2.2 Model Generalisations

Key components of the proposed two-way outlier detection model are the interaction terms $\xi_i\eta_j\delta$ and $\xi_i\eta_j\gamma$ in the item-response and response-time submodels, respectively. Specifically, the effects of the two-way outliers are characterised by the parameters δ and γ in the two submodels, respectively, and they are assumed to be the same across all the outliers. This assumption can be relaxed to allow for heterogeneity among the outliers. One way to relax this assumption is by assuming each drift parameter to be the sum of a person-specific parameter and an item-specific parameter. For example, one may replace $\xi_i\eta_j\delta$ by $\xi_i\eta_j(\delta_i + \delta'_j)$ in the item-response submodel, where δ_i and δ'_j are non-negative person- and item-specific drift parameters, respectively.

Moreover, in the current framework, the outlier model is essentially unidimensional, as a result of the imposed two-way latent class structure. In the application to the detection of item preknowledge, it means that a test taker has preknowledge of either all or none of the leaked items. However, when there are multiple sources of item leakage and test takers with item preknowledge have access to one or more of those sources, this assumptions can be relaxed by assuming multiple latent classes among both the individual and item outliers.

2.3 Related Works

Factor analysis in the presence of outliers has received much attention in the literature, but mainly focuses on the detection of outlying cases/individuals rather than items as well. One line of research is on the robust estimation of factor models (Moustaki and Victoria-Feser, 2006; Pison et al., 2003; Yuan and Bentler, 1998, 2001). Another line of research focuses on the detection of outliers among the individuals who do not fit a baseline factor model, using

residual-based procedures (Reiser, 1996) or forward search procedures (Hadi, 1992; Mavridis and Moustaki, 2008,0). All these works only consider outlying individuals. The proposed two-way outlier detection method is among the very few attempts to simultaneously classify individuals and items as outliers.

Although several models that combine factor and latent class modeling have been proposed for detecting aberrant behaviours (Bolt et al., 2002; Boughton and Yamamoto, 2007; Goegebeur et al., 2008; Shu et al., 2013; Wang and Xu, 2015; Wang et al., 2018), none of them is about two-way classification of individuals and items.

Another feature of our model is that it does not require any prior knowledge about the outlying individuals and items (e.g. a subset of compromised items). Shu et al. (2013) proposed a Deterministic, Gated item response theory Model (DGM) for data consisting only of item responses. This model makes similar assumptions to our item-response submodel, except that (1) the DGM assumes the known status of each item (i.e., whether each item is compromised or not), and (2) the drift parameter for cheating (i.e., δ in the current model) is assumed to be person-specific in the DGM. Our model is more closely related to Wang and Xu (2015) and Wang et al. (2018) who also assume a mixture of log-normal distribution for response times from normal and aberrant response behaviours. Like the proposed method, these works also do not require prior knowledge about the test takers with preknowledge or the leaked items. The main difference is that Wang and Xu (2015) and Wang et al. (2018) focus on identifying person-item pairs for which aberrant behaviours are involved, rather than directly classifying test takers and items. Therefore, they allow aberrance in any person-item combination, by introducing a person-and-item specific latent variable to indicate the status of each response. By having person-and-item specific latent variables, the models of Wang and Xu (2015) and Wang et al. (2018) tend to be more flexible than the proposed model, in the sense that these models allow data to deviate from the baseline model along more directions. Consequently, these models may be preferred when data involve multiple types of aberrant behaviours, such as rapid guessing and cheating. On the other hand, unlike the

proposed method, the models of Wang and Xu (2015) and Wang et al. (2018) do not directly lead to classifications of the test takers and items, let alone quantifying the uncertainty of the classifications. To detect test takers with preknowledge and leaked items, follow-up analysis is needed based on the posterior distributions of the person-and-item specific latent variables. Therefore, these methods are not as straightforward as the proposed one, if the main goal is to perform the two-way detection of individuals with preknowledge and leaked items.

3 Statistical Decision Theory

In what follows, we provide statistical decision theory for the detection of two-way outliers under the proposed model, assuming the model is correctly specified. We start with the classical Bayesian decision theory and then develop compound decision rules for the detection of outlying individuals and items.

3.1 Bayesian Decision Theory

As individuals and items are essentially mathematically exchangeable, we only discuss the Bayesian decision theory for the detection of outlying individuals. We denote D_i as the decision on individual i , where $D_i = 1$ means flagging the individual as an outlier and $D_i = 0$ otherwise. A false positive error happens when $D_i = 1$ and $\xi_i = 0$ and a false negative error happens when $D_i = 0$ and $\xi_i = 1$. Decisions on the detection of outlying individuals involve a trade-off between these two types of errors, whose importance may be asymmetric. For example, in the application to the detection of item preknowledge, a false positive error corresponds to an innocent test taker being flagged as a cheater and a false negative error corresponds to a cheater not being flagged. These two types of errors have substantially different consequences (Skorupski and Wainer, 2017).

To apply Bayesian decision theory to this classification problem, we need to specify the

relative cost of a false positive error, denoted by $\zeta \in (0, 1)$, which further implies that the relative cost of a false negative error is $1 - \zeta$. Then the Bayes risk is defined as

$$\mathcal{R}(D_i) := \zeta P(D_i = 1, \xi_i = 0) + (1 - \zeta)P(D_i = 0, \xi_i = 1). \quad (4)$$

Following the classical Bayesian decision theory (see, e.g., Chapter 2, Shao, 2003), the optimal decision rule which minimises the Bayes risk is obtained by comparing the posterior probabilities with the relative cost ζ . That is, an individual is classified as an outlier if the posterior probability is larger than ζ .

This Bayesian decision rule depends on the relative cost ζ . However, this parameter may not be easy to specify in practice, as the relative importance of a false positive error is often hard to quantify. In what follows, we discuss how this parameter may be chosen adaptively based on a compound risk which is obtained by aggregating information from the entire set of individuals.

3.2 Compound Decision for Detecting Outlying Individuals

We evaluate decision making at an aggregated level for all individuals. This involves solving N decision problems simultaneously and thus is called a compound decision problem (Robbins, 1951; Sun and Cai, 2007; Zhang, 2003). Given a decision rule, hypothetically, the results can be classified into four categories, as summarised in Table 1, where N_{00} , N_{01} , N_{10} , and N_{11} denote the numbers of true negative, false positive, false negative, and true positive, respectively. The quality of decisions can be quantified by two quantities. One is the False Discovery Proportion (FDP) $N_{01}/\max\{N_{\cdot 1}, 1\}$, which is the proportion of non-outliers among the detections. In the application to cheating detection, this gives the proportion of innocent test takers among those who are flagged as cheaters. The denominator is chosen so that this proportion is well-defined even when $N_{\cdot 1} = 0$. The other quantity is the False Non-discovery Proportion (FNP) $N_{10}/\max\{N_{\cdot 0}, 1\}$, which is the proportion of outliers

	Not flagged as outlier	Flagged as outlier	Total
Non-outlier	N_{00}	N_{01}	$N_{0\cdot}$
Outlier	N_{10}	N_{11}	$N_{1\cdot}$
Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

Table 1: A summary of the outcome of detecting outlying individuals. Note that this table is hypothetical, as in real applications, outliers and non-outliers are directly observable.

among the non-detections. It is worth noting that, however, the FDP and FNP cannot be directly used because the outliers are not directly observable. As an alternative, we use the posterior means of the FDP and FNP, which are known as the local FDR and local FNR, respectively. Similar measures have been proposed for solving compound decision problems in Efron et al. (2001), Efron (2004,0,1), among others. Given data and a decision rule, the local FDR and local FNR are completely determined under the proposed model.

Suppose that the consequence of a false positive error is more severe than that of a false negative error, which may be the case for the detection of cheaters in educational testing. Then a sensible decision criterion is to minimise the local FNR, while controlling the local FDR to be below a pre-specified threshold ρ . Given the practical meaning of local FDR, the threshold ρ should be much easier to specify than the relative cost in the Bayesian decision rule discussed previously. For instance, by setting $\rho = 0.01$, we approximately control the proportion of non-outliers to be below 1% among those who are detected as outliers.

Now consider a Bayesian decision rule with a relative cost ζ . We discuss how the optimal ζ is determined by the above decision criterion based on the local FDR and local FNR with a given threshold ρ . For ease of exposition, we use \mathbf{Z} as a generic notation for the data, where $\mathbf{Z} = \mathbf{Y}$ when only item responses are collected and $\mathbf{Z} = (\mathbf{Y}, \mathbf{T})$ when both item responses and response times are available. Specifically, given relative cost ζ , the Bayesian decision for each test taker i can be written as

$$D_i(\zeta) = 1_{\{P(\xi_i=1|\mathbf{Z})>\zeta\}}. \quad (5)$$

Under our fully Bayesian setting and given threshold ζ , the local FDR becomes

$$\text{fdr}_\zeta(\mathbf{Z}) = \frac{\sum_{i=1}^N D_i(\zeta) P(\xi_i = 0|\mathbf{Z})}{\max\{\sum_{i=1}^N D_i(\zeta), 1\}}, \quad (6)$$

which only depends on the posterior probabilities $P(\xi_i = 1|\mathbf{Z})$, $i = 1, \dots, N$. The local FNR can be obtained similarly as

$$\text{fnr}_\zeta(\mathbf{Z}) = \frac{\sum_{i=1}^N (1 - D_i(\zeta)) P(\xi_i = 1|\mathbf{Z})}{\max\{\sum_{i=1}^N (1 - D_i(\zeta)), 1\}}.$$

As summarised in Proposition 1 below, the optimal relative cost is given by $\zeta^* = \inf\{\zeta : \text{fdr}_\zeta(\mathbf{Z}) \leq \rho\}$. That is, the decision rule given by $D_i(\zeta^*)$, $i = 1, \dots, N$, minimises the local FNR under the constraint that the local FDR is below ρ . The proof of Proposition 1 is given in Appendix A.

Proposition 1. *Given data $\mathbf{Z} = \mathbf{Y}$ or (\mathbf{Y}, \mathbf{T}) , the local FDR $\text{fdr}_\zeta(\mathbf{Z})$ as a function of ζ is non-increasing and left-continuous, and the local FNR $\text{fnr}_\zeta(\mathbf{Z})$ is non-decreasing in ζ . Thus,*

$$\zeta^* = \inf\{\zeta : \text{fdr}_\zeta(\mathbf{Z}) \leq \rho\} \quad (7)$$

solves the optimisation

$$\min_{\zeta} \text{fnr}_\zeta(\mathbf{Z}), \quad \text{s.t.} \quad \text{fdr}_\zeta(\mathbf{Z}) \leq \rho. \quad (8)$$

The corresponding optimal decision rule is $D_i(\zeta^) = 1_{\{P(\xi_i=1|\mathbf{Z}) > \zeta^*\}}$.*

Given posterior probabilities $P(\xi_i = 1|\mathbf{Z})$, the optimal decision is easy to obtain. The computation is described in Algorithm 1.

Algorithm 1 (Optimal compound decision). *Let the posterior probabilities $P(\xi_i = 1|\mathbf{Z})$ and threshold ρ for local FDR be given. The optimal relative cost ζ^* is given by the following steps.*

1. Sort the posterior probabilities in an increasing order. Denote the sorted values as

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}.$$

2. Compute the cumulative means $c_{(0)} \leq c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(N)}$, where

$$c_{(0)} = 0, \text{ and } c_{(i)} = \frac{\sum_{j=1}^i p_{(j)}}{i}, i = 1, \dots, N.$$

3. Let $i^* = \max \{i : c_{(i)} \leq \rho\}$.

Then the optimal relative risk is given by $\zeta^* = p_{(i^*)}$, if $i^* > 0$, and $\zeta^* = 0$, if $i^* = 0$.

This local FDR control procedure can be viewed as the Bayesian version of the well-known Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) for multiple hypothesis testing. The BH procedure is designed to control the FDR, which is defined as the unconditional expectation of the FDP. Unlike the proposed procedure that is based on posterior probabilities, the BH procedure achieves the control of FDR using p-values from multiple testing. Under the proposed Bayesian framework, it seems more straightforward to control local FDR as in the proposed procedure. Also note that the FDR is automatically controlled by controlling local FDR, due to the relationship between conditional and unconditional expectations.

Another possible decision criterion is to control the posterior probability of making at least one false discovery, which corresponds to the Family-Wise Error Rate (FWER) under the frequentist setting. This FWER-type criterion exerts a more stringent control over false discovery than the proposed one by its definition. Therefore, the proposed procedure has greater power at the cost of increased rates of false positives. In this sense, the proposed local FDR control procedure is more suitable when having a large number of individuals and thus a large number of decisions need to be made simultaneously.

For certain applications, false negatives may have a more significant consequence than false positives. Then it may be more suitable to minimise the local FDR while controlling

local FNR. As the definitions of local FDR and local FNR are mathematically symmetric, the above procedure can be easily adapted.

3.3 Compound Decision for Detecting Outlying Items

The compound decision theory developed above can be adapted to the detection of outlying items, based on the posterior probabilities for the item-specific binary indicators η_j . In the application to the detection of item preknowledge, when there is sufficient evidence suggesting that an item is compromised, it should be removed to maintain the quality of the item pool. This decision problem faces a trade-off between the financial cost for item pool replenishment and the need of maintaining the quality of the item pool. For high-stake tests, test fairness is usually the first priority and thus false negatives may have a more significant consequence than false positives. In that case, it becomes more sensible to minimise the local FDR, under the constraint that the local FNR is below a suitable threshold (e.g., 1%).

4 Bayesian Inference

4.1 Prior and Hyperprior Specification

4.1.1 Prior and hyperprior specification

We showcase the specification of prior and hyperprior distributions under the specific model with item-response submodel (1) and response-time submodel (3). We start with the specification of g_1 , the joint distribution of $\Theta_i = (\theta_i, \xi_i, \tau_i)$. It is assumed that (θ_i, τ_i) follows a bivariate normal distribution $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{2 \times 2}$. Note that a person's ability and speed are typically correlated, which is why we assume a bivariate normal distribution for (θ_i, τ_i) . Similar settings are adopted in existing models for item responses and response times; see e.g., van der Linden (2007). We further assume that the latent indicator ξ_i is independent of (θ_i, τ_i) , following a Bernoulli distribution $\text{Bern}(\pi_1)$. This independence as-

sumption can be relaxed by modelling the conditional distribution of ξ_i given (θ_i, τ_i) , for example, by a logistic regression model. This relaxation is left for future investigation.

We now specify g_2 , the joint distribution of $\Delta_j = (\beta_j, \eta_j, \alpha_j)$. Similar to that of g_1 , we first let (β_j, α_j) follow a bivariate normal distribution $N(\boldsymbol{\mu}, \Omega)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\Omega = (\omega_{ij})_{2 \times 2}$. It is further assumed that η_j is an independent Bernoulli random variable, $\text{Bern}(\pi_2)$.

It remains to specify the prior for positive parameters δ , γ , κ , as well as the priors for hyperparameters π_1 , π_2 , μ_1 , μ_2 , Ω , and Σ .

1. As δ is positive, we assume a half-Cauchy prior distribution with scale 2.5. This is regarded a weakly informative prior, following the suggestions given in Gelman (2006), Gelman et al. (2008) and Polson and Scott (2012).
2. γ is assumed to follow the same half-Cauchy prior distribution as δ .
3. κ is assumed to follow an inverse Gamma distribution, $\text{IG}(0.001, 0.001)$, where the shape and scale parameters are both set to 0.001. This choice follows the suggestion in Chapter 5, Lunn et al. (2012).
4. π_1 and π_2 are assumed to be i.i.d., following a beta distribution $\text{Beta}(2, 2)$. This prior distribution can be regarded as a weakly informative prior given the sample and item sizes in our application. It is suggested in Agresti and Coull (1998) and Carlin and Louis (2000), Chapter 2, as the prior for a proportion parameter. We choose this distribution rather than a uniform distribution, because π_1 and π_2 are believed to not locate on the boundaries of the interval $[0, 1]$.
5. μ_1 and μ_2 are assumed to be i.i.d., following a normal distribution $N(0, 5^2)$. The standard deviation 5 is chosen based on the scales of μ_1 and μ_2 in the current application, under which this prior may be regarded as weakly informative.

6. Σ and Ω are assumed to be i.i.d., following an inverse Wishart distribution where the scale matrix, $\text{IW}(\Psi, \nu)$, $\Psi = ((2, 0)^\top, (0, 2)^\top)$, and the degree of freedom $\nu = 2$. This choice follows the suggestion in Chapter 6, Lunn et al. (2012). Under this prior distribution, σ_{11} , σ_{22} , ω_{11} , and ω_{22} marginally follow an inverse Gamma distribution $\text{IG}(1/2, 1)$.

4.1.2 Induced priors and hyper-priors for reduced model

For the reduced model of item response data, the priors and hyperpriors are induced by those for the full model. For completeness, we list the induced priors and hyperpriors below.

1. For $\Theta_i = (\theta_i, \xi_i)$, θ_i and ξ_i are independent, following normal distribution $N(0, \sigma_{11})$ and Bernoulli distribution $\text{Bern}(\pi_1)$, respectively.
2. Similarly, for $\Delta_j = (\beta_j, \eta_j)$, β_j and η_j are independent, following normal distribution $N(\mu_1, \omega_{11})$ and Bernoulli distribution $\text{Bern}(\pi_2)$, respectively.
3. δ follows a half-Cauchy prior distribution with scale 2.5.
4. π_1 and π_2 are i.i.d., following a beta distribution $\text{Beta}(2, 2)$.
5. μ_1 follows a normal distribution $N(0, 5^2)$.
6. σ_{11} and ω_{11} are i.i.d. $\text{IG}(1/2, 1)$.

4.2 Bayesian Inference

4.2.1 Computation

Statistical inference is carried out under a full Bayesian setting. An MCMC algorithm is developed for the computation¹. This computation is non-trivial, due to the presence of many discrete variables and the interactions between them. More specifically, the model involves

¹The R code for the MCMC algorithm can be found on <https://github.com/YanLu-stats/OD2WIRT>.

person- and item-specific binary latent variables ξ_i and η_j . The complexity of simulating these variables by MCMC is similar to the simulation of discrete systems like mixture models and Ising-type models. Such systems typically involve many well-separated local modes and thus suffer from the issue of slow mixing (e.g., Celeux et al., 2000; Katzgraber et al., 2006; Richardson and Green, 1997). Tempering methods (Geyer, 2011) provide a powerful tool for exploring distributions with many local modes. Specifically, parallel tempering, which is also known as the Metropolis-coupled Markov chain Monte Carlo, is chosen for the computation of the proposed model. More precisely, parallel tempering simulates multiple MCMC chains simultaneously. A Metropolis-Hastings sampler is used for the MCMC sampling within each chain. The target distributions of these chains are obtained by tempering the original posterior density, i.e., raising the original posterior density to different powers $T^{-1} \in [0, 1]$, where T is known as the ‘temperature’. The original posterior density is included by setting $T = 1$. A chain corresponding to a higher temperature tends to have a flatter target distribution, for which the MCMC sampler is less likely to be trapped at local modes and thus has fast mixing. In contrast, when the temperature is low, the MCMC chain is more likely to be trapped and thus suffers from slow mixing. Parallel tempering improves the mixing of the low-tempered MCMC chains, by exchanging information between chains with adjacent temperatures. That is, at each iteration, a pair of chains with adjacent temperatures is randomly chosen and a Metropolis-Hastings update is used to decide whether to swap their parameter states.

The use of the algorithm requires some tuning, including (a) the step sizes of random-walk Metropolis-Hastings updates within each chain, (b) the number of temperature levels, and (c) the temperature values. For (a), we follow the suggestion given by Roberts and Rosenthal (2001); that is, we tune the step sizes to achieve an acceptance rate around 2.3. For (b) and (c), it is suggested to follow the theoretical guidance given in Atchadé et al. (2011). Further details of this algorithm are given in Appendix B. For the implementation of the decision procedures in Section 3, the posterior distributions of ξ_i and η_j are approximated by the

posterior samples from this MCMC algorithm.

Instead of taking a fully Bayesian setting, it is also possible to adopt an empirical Bayes framework (Casella, 1985; Efron, 2014; Robbins, 1956), under which $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta, \gamma$, and κ are treated as fixed parameters and estimated by maximum likelihood estimation, while the person- and item-specific parameters Θ_i and Δ_j are treated as random variables. However, due to the complex structure of the current model, the expectation-maximisation algorithm, which is the standard tool for empirical Bayes inference, is computationally infeasible. A tailored stochastic optimisation algorithm needs to be developed.

4.2.2 Model comparison

We use model comparison methods to answer the following questions that may be of substantive interest in specific applications. That is, does our item response data show evidence of the existence of outlying individuals and items? If so, does item response time information help detect these outliers? These questions may be answered by Bayesian model comparison.

To answer the first question, we compare the proposed model for item responses with the baseline item-response model which does not contain outliers. These two models are the same, including the specification of the priors and hyperpriors, except that the hyperparameters π_1 and π_2 are set to be 0 and thus $\xi_i \eta_j \delta = 0$ for all i and j in the baseline model. The preference of the proposed model against the baseline model suggests the existence of the outliers.

To answer the second question, we compare the proposed model for item responses and response times with a null model for the same data. This null model is the same as the proposed model, except that the drift parameter γ in the response-time submodel is set to be zero. When $\gamma = 0$, it means that there is no difference between the outlying and non-outlying individuals in their response-time distributions. If the proposed model is preferred to the null model, it suggests that response times contain information about the outliers thus incorporating response time information may improve the detection of outliers.

These models are compared by the deviance information criterion (DIC; Spiegelhalter et al., 2002), more specifically, a marginalised DIC in which the person- and item-specific parameters are treated as latent variables and integrated out. In comparing two models, the one with a smaller DIC value is preferred. We choose the marginalised DIC, instead of the conditional DIC that incorporates the person- and item-specific variables in the focus of the analysis, because the marginalised DIC often performs better in comparing hierarchical models (e.g., Quintero and Lesaffre, 2018). The marginalised DIC is computed by MCMC sampling.

While computationally convenient, the DIC suffers from several caveats, including the lack of model selection consistency and the possibility of selecting overfitted models (see Spiegelhalter et al., 2014, and references therein). Therefore, we note that the results based on the DIC need to be interpreted with caution. In future research, other model comparison criteria will be investigated and compared with the DIC, including the Bayes factor (Kass and Raftery, 1995) and the Bayesian information criterion (BIC; Schwarz, 1978) that approximates the logarithm of the Bayes factor. The Bayes factor and BIC may be theoretically more attractive because they yield consistent model selection under suitable regularity conditions. However, their computation tends to be less straightforward than the DIC. Algorithms remain to be developed for computing them under the current modelling framework.

5 An Application to Licensure Test

We apply the proposed method to a dataset from a computer-based non-adaptive licensure test. The test is designed and operated under the Rasch model, which is consistent with the proposed baseline item-response submodel. This dataset has also been analysed for the detection of cheating in several chapters of Cizek and Wollack (2017) and journal articles including Sinharay (2017a) and Sinharay (2017b). We point out that the methods in these analyses require prior information about the items' statuses. For example, Sinharay (2017a)

and Sinharay (2017b) require to know the compromised items a priori and focus on the detection of cheaters. Unlike these existing analyses, we focus on the simultaneous detection of cheaters and compromised items, without requiring such prior knowledge.

The test contains 170 binary-scored items ($J = 170$), for which test takers' item responses and response times are available. The dataset is preprocessed by removing 12 test takers with zero response time in one or multiple items, which is believed to be data recording error. This leads to a final dataset containing 1,624 test takers ($N = 1,624$). The testing program flagged 41 among these 1,624 test takers as likely cheaters, through a combination of data analysis and a careful investigative process which brought in other pieces of information. By a similar investigation process, the testing program also believed that 64 among the 170 items were compromised. These labels will be used as partial truth for validating our data analysis results, but *they are not used in our model*. It is worth noting that these labels are not the ground truth and it is possible that there were test takers and items which ought to have been flagged but were not (Chapter 1, Cizek and Wollack, 2017). It is believed that the given labels are of good quality, so that they can be used for the evaluation of detection methods. On the other hand, evaluation criteria based on these labels are not perfect, due to possible labelling errors.

The purpose of this analysis is twofold. First, it is used to show the effectiveness of the proposed method, through a comparison between our results and the partial truth given by the testing program. Second, it is used to demonstrate the use of the proposed method in real tests, which may be of interest to practitioners.

5.1 Descriptive Analysis

We start with descriptive analysis to give an overview of the dataset. Panel (a) of Figure 4 shows the histogram of test takers' total scores by the testing program's cheating labels. Similarly, Panel (b) of Figure 4 gives the histogram of items' correct rates by the testing program's compromisation labels. Similarly, the two panels of Figure 5 show the histograms

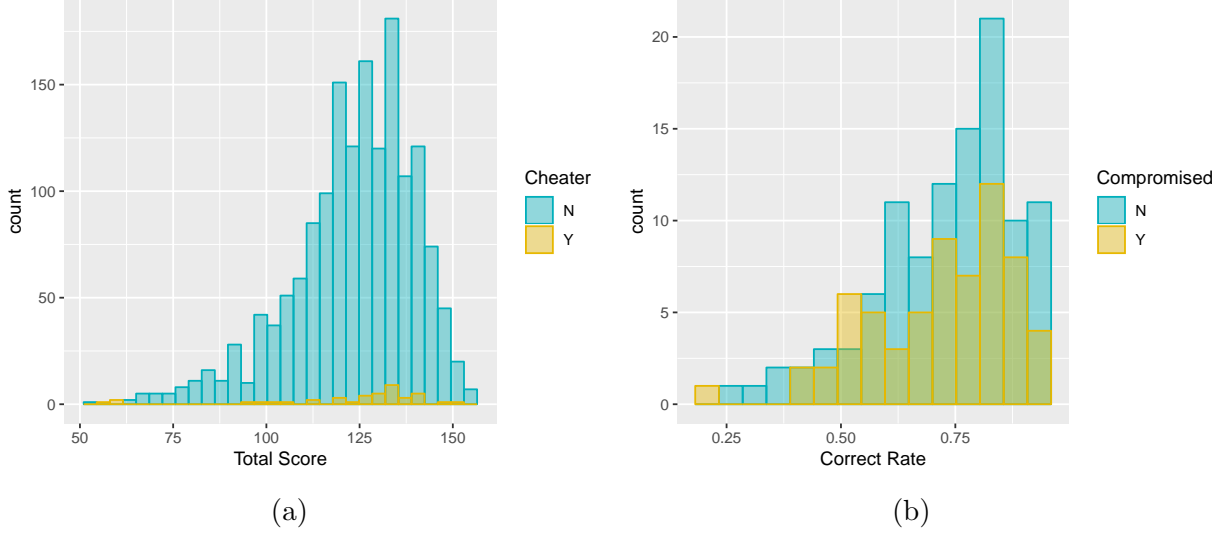


Figure 4: Descriptive analysis. Panel (a): Histogram of test takers’ total scores by the testing program’s cheating labels. Panel (b): Histogram of items’ correct rates by the testing program’s compromise labels.

of the mean response time in logarithm scale for test takers and items, respectively.

From these plots, it is not difficult to see that the corresponding summary statistics do not have much information about the labels on the test takers and items. In fact, the area under the curves (AUC) of the corresponding receiver operating characteristic (ROC) curves are 55.2% and 71.7% for the classification 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. As we will see in the sequel, the proposed method substantially improves upon these benchmarks.

5.2 Detection based on Item Responses

We start with analysing item responses using the reduced model (i.e. only analyse item responses and not time responses). Using the algorithm given in Appendix B, three MCMC chains were run with random starting points. Their convergence was assessed by trace plots and the Gelman and Rubin (GR) diagnostic statistic (Gelman and Rubin, 1992). The Gelman-Rubin R statistics applied to the parameters which are not person- or item-specific

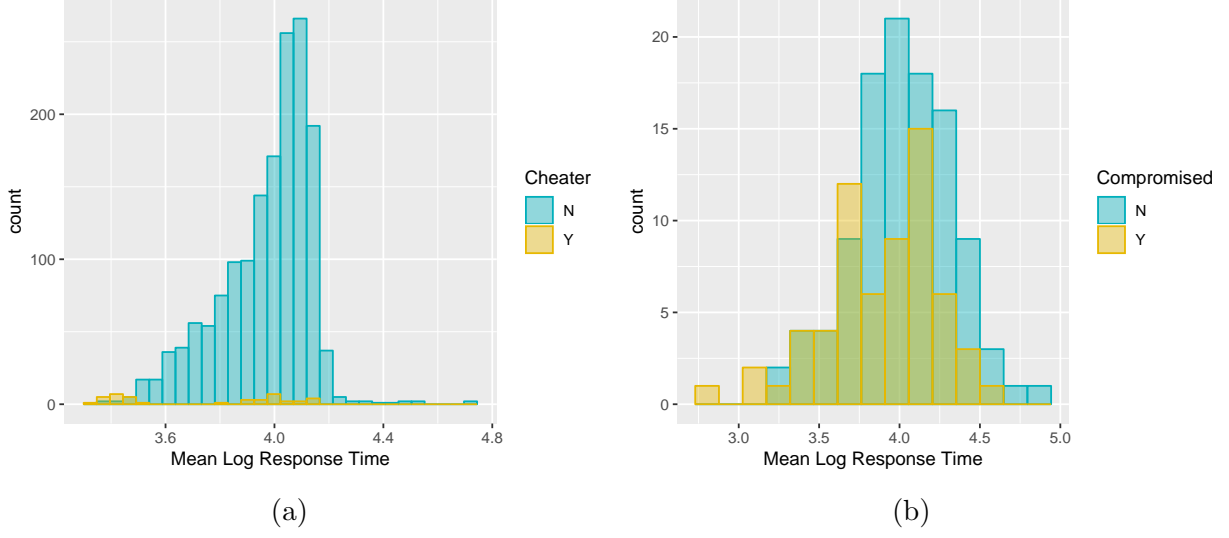


Figure 5: Descriptive analysis. Panel (a): Histogram of test takers’ mean log-time by the testing program’s cheating labels. Panel (b): Histogram of items’ mean log-time by the testing program’s compromise labels.

(see Table 3 for the list of these parameters) are below 1.20, suggesting that the chains converged to their equilibrium distributions after 10,000 iterations.

Inference is drawn based on 24,000 posterior samples from the three converged chains, where each contributes 8,000 samples. We first compare the fitted model with its null version using the DIC measure described in Section 4.2, to answer the question “does our item response data show evidence of cheating?”. Recall that $\pi_1 = \pi_2 = 0$ in the null model, meaning that there are no cheaters or compromised items. The DIC value for the null model is also based on 24,000 posterior samples from an MCMC algorithm. The DIC values for the proposed and the null models are 138,282.6 and 218,308.4, respectively. The smaller DIC for the proposed model suggests that item preknowledge is likely to exist among the test takers.

We then examine the classification results. Panel (a) of Figure 6 gives the box plots of the posterior means of ξ_i for the cheating and non-cheating groups (defined by the testing program), respectively. As we can see, the posterior means of ξ_i for the cheating group tend to be close to 1 and those for the non-cheating group tend to be close to 0, with some exceptions. Panel (b) of Figure 6 gives box plots of the posterior means of η_j for the compromised and non-compromised items (defined by the testing program), respectively.

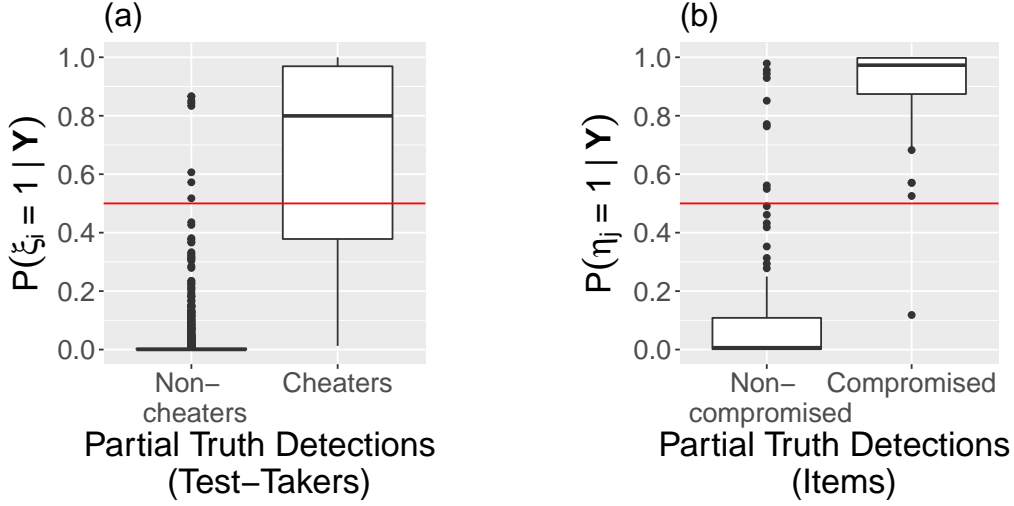


Figure 6: Applying the reduced model for item responses. Panel (a): Box plots of the posterior means of ξ_i for the cheating and non-cheating groups (defined by the testing program). Panel (b): Box plots of the posterior means of η_j for the compromised and non-compromised items (defined by the testing program).

Similarly, the posterior means of η_j for the compromised items tend to be close to 1 and those for the non-compromised items tend to be close to 0. The corresponding ROC curves for the classification of the cheating and compromise labels are presented in Figure 7. The AUCs for these two ROC curves are 0.868 and 0.836, respectively. They are substantially larger than the ones given by the summary statistics discussed in Section 5.1. We remark that these results on the detection accuracy should be interpreted with caution, due to possible labelling errors.

Moreover, Panel (a) of Figure 8 shows the local FDR and the local FNR as functions of the number of detections, respectively, when applying the proposed compound decision rule to test takers. As we can see, as the number of detections increases, the local FDR increases and the local FNR decreases. The same plot for items is given in Panel (b) of Figure 8. Specifically, the numbers of detections under different thresholds are given in Table 2, where we control local FDR for test takers and control local FNR for items. Again, we remark that the validity of the detection results depends on the extent to which our model assumptions hold. Therefore, we suggest treating such detection results as initial screening results, rather

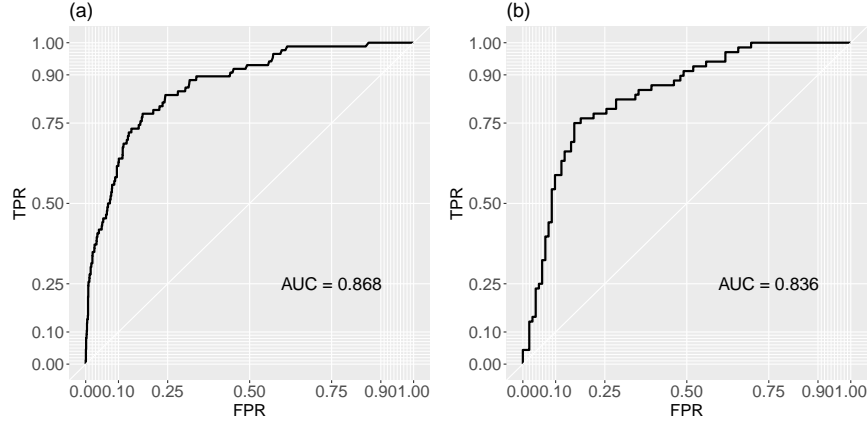


Figure 7: Applying the reduced model for item responses. Panel (a): ROC curve for the classification of cheaters (labelled by the testing program) by the posterior means of ξ_i . Panel (b): ROC curve for the classification of compromised items (labelled by the testing program) by the posterior means of η_j . The x- and y-axes of an ROC curve give the true positive rate (TPR) and false positive rate (FPR) for classification, respectively.

than as the final decisions.

Finally, posterior means and 95% credible intervals for the global parameters are presented in Table 3, where the global parameters refer to the parameters that are not person-specific or item-specific. In particular, the posterior mean of the proportion of cheaters is 2.8%, with 95% credible interval (2.0%, 3.6%). This estimate is close to the proportion of 2.5% calculated based on the cheating labels from the testing program. The posterior mean of the proportion of compromised items is 40.1%, with 95% credible interval (38.7%, 43.3%).

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

Table 2: Applying the reduced model for item responses. 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.

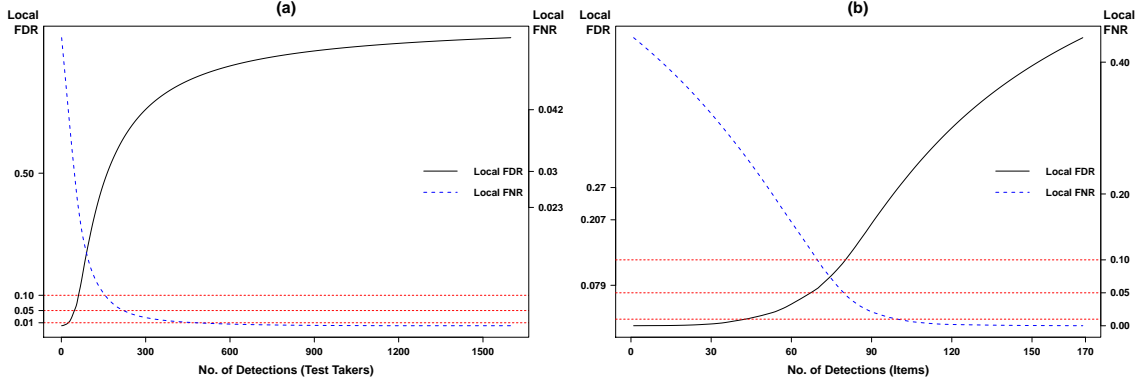


Figure 8: Applying the reduced model for item responses: The local FDR (represented by black solid curves) and the local FNR (represented by blue dashed curves) as functions of the number of detections.

	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)	(0.669, 0.854)	(-1.237, -0.912)	(0.758, 0.959)

Table 3: Applying the reduced model for item responses. The row labelled “EAP” shows the posterior means of the global parameters, where EAP represents the Expected A Posteriori, and the row labelled “95% CI” provides the corresponding 95% credible intervals.

This estimate is close to, but slightly higher than, the proportion of 37.6% given by the testing program. It may be the case that the testing program missed several compromised items during its labelling process. Furthermore, the posterior mean of δ is 0.895. That is, 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. Again, we point out that these interpretations depend on the extent to which our model holds and thus should be taken with caution.

5.3 Detection based on Item Responses and Response Times

We continue the modelling process by incorporating information from response times. The full model is applied to the dataset consisting of both item responses and response times. Three MCMC chains were used to fit the model. According to the GR statistics applied to the global parameters, the chains converged to their equilibrium distributions after 18,000

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

Table 4: Applying the full model for item responses and response times. 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.

iterations.

Similar as above, inference is drawn based on 24,000 posterior samples from the three chains after convergence. We compare this model with its null version by DIC, to answer the question “does item response time information help detect cheating?” Recall that these two models are the same, except that the response-time drift parameter $\gamma = 0$ in the null model. The DIC values for the proposed full model and its null version are 176,935.2 and 214,201.3, respectively. The smaller DIC value for the proposed full model suggests that response times contain substantial information about the cheating indicators.

The classification results are similar to those based only on item responses, and thus some plots shown above are omitted here. In particular, the ROC curves based on the posterior means of ξ_i and η_j have AUCs of 0.892 and 0.867, respectively, where these AUC values are slightly higher than those from the reduced model. In addition, the numbers of detections for test takers and items are shown in Table 4, where we still control local FDR for test takers and control local FNR for items. Comparing the results in Tables 2 and 4, generally more detections tend to be made under the full model. This is likely due to the fact that the posterior distributions tend to be more concentrated under the full model as it utilises more information.

Posterior means and 95% credible intervals for the global parameters are given in Table 5. Comparing Tables 3 and 5, we find that the estimates of the common parameters shared by the two models are close to each other. In particular, the 95% credit intervals overlap for each parameter. In addition, based on the posterior mean of Σ , the correlation between

	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)	(0.986, 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)

Table 5: Applying the full model for item responses and response times. The row labelled “EAP” shows the posterior means of the global parameters and the row labelled “95% CI” provides the corresponding 95% credible intervals.

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. Such a high correlation between the two factors is not uncommon for high-stake tests. For example, Wang et al. (2013) report a similar level of correlation between the ability and speed factors in a high-stake computerised adaptive test, under a similar Bayesian hierarchical model but without a cheating component. The estimated 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.

6 Simulation Study

We now present two simulation studies for evaluating the finite-sample performance of the proposed method. The first study focuses on settings where our model is correctly specified, and the second study investigates the sensitivity of the proposed method under various forms of model misspecification.

6.1 Study I

6.1.1 Settings

We consider simulation settings that mimic real educational tests. Specifically, we consider two settings for the sample size N and item size J , (1) $N = 2,000, J = 200$, and (2) $N = 4,000, J = 400$. This leads to two different settings, where the detection is expected to be more accurate under the second setting given its larger sample and item sizes. In what follows, these two settings are referred to as S1 and S2, respectively.

For each setting, we generate 50 independent datasets under the full model, with the global parameters fixed across the datasets. The proportion parameters π_1 and π_2 are set to 10% and 40%, respectively, the drift parameters δ and γ are both set to be 1.2, and the rest of the global parameters are set to be the same as the posterior means in Table 5 from the real data analysis above. For each dataset, we apply both the reduced model for item responses and the full model for item responses and response times. An additional simulation study is presented in the appendix that shares a similar setting with the current study, except that the item size J is set to mimic educational tests with a smaller number of items. More specifically, this additional study considers two settings for N and J : (1) $N = 2,000, J = 50$, and (2) $N = 4,000, J = 100$, and the rest of the settings remain the same. Similar results are observed in this additional study as those below from Study I.

6.1.2 Results

The analysis is conducted using our parallel tempering MCMC algorithm. For each dataset, we run 10,000 iterations, with the first 3,000 iterations as the burn-in. The results are based on the posterior samples from the last 7,000 iterations.

We first examine the classification results. For each model and each simulated dataset, we classify the test takers based on the posterior means of ξ_i and evaluate the classification based on the AUC value of the corresponding ROC curve. A larger AUC value implies higher

Test taker					Item			
S1			S2		S1		S2	
AUC	Reduced	Full	Reduced	Full	Reduced	Full	Reduced	Full
25%	0.953	0.954	0.951	0.959	0.951	0.959	0.951	0.959
50%	0.981	0.983	0.984	0.987	0.976	0.980	0.979	0.981
75%	0.990	0.994	0.993	0.993	0.992	0.994	0.997	0.996

Table 6: Study I: Overall classification performance based on the posterior means of ξ_i and η_j . For each model, each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 50 independent datasets.

S1							S2					
FDP	Reduced			Full			Reduced			Full		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.038	0.088	0.007	0.037	0.086	0.004	0.026	0.067	0.004	0.025	0.072
50%	0.012	0.048	0.091	0.011	0.048	0.092	0.007	0.031	0.079	0.006	0.029	0.083
75%	0.016	0.052	0.099	0.015	0.056	0.096	0.009	0.039	0.092	0.007	0.033	0.088

Table 7: Study I: Local FDR control for test takers. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 50 independent datasets.

classification accuracy. Similarly, the classification of the items is based on the posterior means of η_j and the accuracy is measured by the corresponding AUC value. These results are shown in Table 6. It can be observed that the classification is slightly more accurate under setting S2, due to the increased sample and item sizes. Moreover, the AUC values given by the full model tend to be slightly larger than those from the reduced model, thanks to the additional information from response times.

We further evaluate the proposed compound decision rules. For each dataset, we control local FDR and local FNR at levels 1%, 5%, and 10% for test takers and items, respectively. We evaluate each decision rule by examining the resulting FDP and FNP; see Section 3.2 for the definitions of FDP and FNP. The results are given in Tables 7 and 8 for the classifications of test takers and items, respectively. According to these tables, the FDP is well-controlled for test takers and so is the FNP for items.

Finally, we show the results on the estimation of the global parameters, as these param-

S1							S2					
FNP	Reduced			Full			Reduced			Full		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.049	0.091	0.010	0.045	0.089	0.006	0.024	0.063	0.007	0.025	0.065
50%	0.011	0.051	0.098	0.012	0.048	0.092	0.010	0.031	0.079	0.011	0.033	0.077
75%	0.013	0.059	0.104	0.012	0.057	0.096	0.015	0.039	0.091	0.012	0.037	0.089

Table 8: Study I: Local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 50 independent datasets.

S1							S2						
Reduced model							Reduced model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ		π_1	π_2	σ_{11}	μ_1	ω_{11}	δ
Bias	0.13	0.09	-0.15	-0.19	-0.11	0.13	Bias	0.09	0.05	0.11	-0.08	-0.08	0.09
Variance	0.14	0.12	0.37	0.23	0.27	0.31	Variance	0.11	0.15	0.29	0.25	0.21	0.27
S1							Full model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ	
Bias	0.11	-0.08	0.07	-0.24	-0.08	0.08	-0.12	-0.04	0.07	0.14	0.09	-0.16	
Variance	0.16	0.11	0.34	0.19	0.32	0.35	0.09	0.07	0.12	0.13	0.08	0.77	
S2							Full model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ	
Bias	0.07	-0.03	0.11	-0.18	-0.05	0.11	-0.08	-0.06	0.09	0.15	0.05	-0.11	
Variance	0.12	0.08	0.33	0.21	0.34	0.31	0.05	0.09	0.08	0.10	0.12	0.63	

Table 9: Study I: Accuracy of the posterior mean estimator of the global parameters. The bias and variance for the posterior mean estimator are calculated based on the 50 replications.

eters have substantive interpretations in cheating detection. Specifically, we focus on the posterior mean estimator, for which bias and variance are estimated based on the results from 50 independent replications. These results are presented in Table 9. The bias, in general, tends to be close to zero for all the global parameters from both models and both settings. In addition, the estimation tends to be more accurate under setting S2, due to the increased sample and item sizes.

Setting	Misspecification	N	J
S3	(1)	2,000	200
S4	(2)	2,000	200
S5	(3)	2,000	200
S6	(1)	4,000	400
S7	(2)	4,000	400
S8	(3)	4,000	400

Table 10: Study II: Six simulation settings, where (1)-(3) correspond to three forms of model misspecification, including the misspecification of (1) the baseline model, (2) the relationship between ξ_i and θ_i , and (3) the common drift parameter.

6.2 Study II

6.2.1 Settings

In this study, we investigate the sensitivity of the proposed method under several forms of model misspecification. For the clarity of simulation settings, we focus on the misspecification of the item-response submodel. That is, we generate item-response data from a misspecified model and then apply our reduced model to classify the test takers and items. The overall classification performance, as well as the performance of the proposed compound decision rules, is evaluated. We focus on three forms of model misspecification whose details are discussed below, including the misspecification of (1) the baseline model, (2) the relationship between ξ_i and θ_i , and (3) the common drift parameter. The three forms of model misspecification, together with two settings for N and J as in Study I, lead to six different settings as summarised in Table 10. For each setting, except for the misspecified part, the global parameters are set the same as those in Study I. For each setting, 50 independent datasets are generated.

We now discuss the three forms of model misspecification in detail. For the baseline model, we replace the Rasch model by the two-parameter logistic model, an IRT model that is widely used in educational testing. That is, the following item-response submodel is

assumed

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

where α_j is known as the discrimination parameter. Note that the proposed model can be viewed as a special case where $\alpha_j = 1$ for all j . In the misspecified model, we generate the discrimination parameters α_j from a uniform distribution $U[1, 1.5]$.

In the proposed model, θ_i and ξ_i are assumed to be independent, meaning that whether a test taker cheats or not is independent of his/her ability. This assumption may not hold and it is likely that these two variables are negatively associated, i.e., test takers with lower ability are more likely to cheat in a test. To mimic this situation, we generate (θ_i, ξ_i) jointly from a Gaussian copula. That is, we first generate (θ_i, ξ_i^*) from a bi-variate normal distribution, with mean vector $(-0.867, 0)^\top$ and covariance matrix $((0.289, -0.134)^\top, (-0.134, 1)^\top)$. Under this bivariate normal distribution, the correlation between θ_i and ξ_i^* is -0.25 . We then let $\xi_i = 1_{\{\xi_i^* \geq z_{0.9}\}}$, which is obtained by truncating ξ_i^* at $z_{0.9}$, the 90% quantile of the standard normal distribution, so that $P(\xi_i = 1) = 0.1$. Under this Gaussian copula model, the marginal distributions of θ_i and ξ_i remain the same as those in Study I, while a negative association is introduced between the two variables.

For model parsimony, it is also assumed in the proposed model that the drift parameter δ is common across all the test takers and items. This assumption may not hold in practice. Therefore, in this misspecified model, instead of using a constant drift, we assume the drift parameter to be both item- and person-specific. That is, we assume

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

where the drift parameters δ_{ij} are generated i.i.d. from a uniform distribution $U[1, 1.5]$.

AUC	Test takers						Item					
	S_3	S_4	S_5	S_6	S_7	S_8	S_3	S_4	S_5	S_6	S_7	S_8
25%	0.904	0.889	0.945	0.912	0.922	0.947	0.921	0.903	0.949	0.947	0.937	0.945
50%	0.965	0.927	0.975	0.969	0.950	0.981	0.965	0.936	0.964	0.962	0.951	0.976
75%	0.983	0.989	0.994	0.980	0.984	0.993	0.981	0.971	0.988	0.985	0.984	0.991

Table 11: Study II: Overall classification performance based on the posterior means of ξ_i and η_j . For each model, each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 50 independent datasets.

6.2.2 Results

We evaluate the proposed method under the six settings above. Similar to Study I, we evaluate the overall classification performance by AUC and the performance of the compound decision rules by the corresponding FDP and FNP. The results are given in Tables 11 through 13. Specifically, the AUC values in Table 11 are comparable to those from the correctly specified model in Table 6, though the AUCs from settings S_3 , S_4 , S_6 , and S_7 are slightly smaller. As further shown in Tables 12 and 13, the compound decision rules tend to control the corresponding FDP and FNP under the targeted levels, except when the target level is 1%. That is, when controlling the local FDR and local FNR to be below 1% for test takers and items, respectively, the resulting FDP and FNP tend to exceed the targeted level under all six settings. This is likely due to the fact that the posterior probabilities cannot be accurately obtained when they are close to 0 or 1, under model misspecification.

Overall, the proposed method is reasonably robust against several forms of model specification, though the performance may be slightly affected. However, under potential model misspecification, the method should be used with caution if we aim to control local FDR or local FNR to be below a very small threshold (e.g., 1%).

7 Concluding Remarks

In this paper, we propose a Bayesian hierarchical model for the detection of latent dif-

FDP	S3			S4			S5		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.013	0.031	0.061	0.018	0.042	0.059	0.012	0.028	0.067
50%	0.015	0.039	0.072	0.023	0.048	0.065	0.026	0.039	0.072
75%	0.019	0.046	0.074	0.026	0.059	0.070	0.029	0.043	0.079
FDP	S6			S7			S8		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.004	0.029	0.075	0.014	0.036	0.059	0.009	0.025	0.061
50%	0.007	0.041	0.089	0.017	0.043	0.066	0.014	0.041	0.064
75%	0.012	0.045	0.093	0.024	0.047	0.072	0.023	0.052	0.075

Table 12: Study II: Local FDR control for test takers. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 50 independent datasets.

FNP	S3			S4			S5		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.007	0.021	0.061	0.019	0.032	0.078	0.015	0.039	0.072
50%	0.014	0.032	0.074	0.022	0.037	0.073	0.028	0.043	0.081
75%	0.017	0.038	0.076	0.024	0.041	0.085	0.030	0.047	0.087
FNP	S6			S7			S8		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.024	0.045	0.012	0.031	0.072	0.014	0.036	0.064
50%	0.011	0.025	0.059	0.015	0.036	0.079	0.026	0.046	0.073
75%	0.014	0.032	0.082	0.025	0.041	0.086	0.029	0.054	0.083

Table 13: Study II: local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 50 independent datasets.

ferential item functioning in item-response-type multivariate data. The proposed method simultaneously detects outlying individuals and items that deviate from a given baseline model. Furthermore, a compound decision theory is proposed for the detection of two-way outliers under a Bayesian decision framework. Statistical inference is carried out under a fully Bayesian framework for which a parallel tempering MCMC algorithm is developed. The proposed model is largely motivated by, and applied to, the detection of test takers who benefit from item preknowledge and compromised items in educational tests. The proposed method requires little prior knowledge about either the test takers with item preknowledge or the compromised items, and thus is directly applicable to operational tests as a monitoring tool or more generally about outlying cases and items in other applications.

The proposed method is successfully applied to data from a licensure test which is known to suffer from item preknowledge. In this study, two models are applied, including the reduced model for item responses and the full model for item responses and response times. Both models accurately detect the potential cheaters and compromised items identified by the testing program, suggesting their usefulness in practice. In addition, the full model performs slightly better than the reduced one, suggesting that response-time information may help detect cheating. However, it should be noted that the labels provided by the testing program in this example are not the ground truth and thus the accuracy measures may be compromised. The validity of the proposed method remains to be checked through applications to other educational tests. We note that a simple model, such as the one applied in the case study, may be more preferable for the detection of cheating in educational tests, even though more general models are available as discussed in Section 2. This is because the numbers of test takers with preknowledge and compromised items are usually small in an educational test, which makes the effective sample sizes small for estimating parameters related to the outlier classes. In that case, a more complex model may lead to a high variance in the estimation, which further yields inaccurate classifications.

Limitations of the proposed method have been discussed in Section 2.1.4. as well as its

robustness against model misspecification in Section 6.2. Another limitation is that it only models a specific type of cheating, i.e., preknowledge due to item leakage. It does not handle other types of cheating behaviours, such as copying others’ answers, electronic transmission of data, hiring stand-ins, and bribing test administrators to correct one’s answers. To investigate different types of cheating behaviours, different sources of information are needed and suitable statistical methods remain to be developed. For example, to detect copying behaviour, a statistical model is needed to characterise the similarity between the item responses from two test takers, possibly taking into account their response process information (e.g., response time), seat locations in a test centre, etc. We leave these problems for future investigation.

Missing data are widely encountered in educational tests that may be informative for the detection of cheating, though not observed in our real data example. For an educational test with cheating test takers, the missingness of a response likely depends on whether the test taker is cheating and whether the item is compromised. If many missing responses are observed, then the current framework should be extended by modelling the probabilities of responding. This problem is left for future development, for which ideas may be borrowed from latent variable models for non-ignorable missingness (e.g., Kuha et al., 2018; O’Muircheartaigh and Moustaki, 1999).

Besides the applications to cheating detection in educational tests, future research will be conducted to investigate the computation, model evaluation and comparison in other areas of application, such as voting behaviours and psychological measurement. Specifically, MCMC algorithms for the inference of the proposed model will be further explored. Although our parallel tempering algorithm works well for the current analysis, its performance will be evaluated under more settings, especially large-scale settings (larger numbers of individuals and items). In addition, other tempering methods, such as simulated tempering, can be explored. Moreover, goodness-of-fit issues and model selection will be further studied. In particular, the use of Bayes factors and BIC for comparing the proposed model with several

relevant models will be investigated.

Appendix

A Proof of Proposition 1

Proof. We let $p_{(1)} < \dots < p_{(n)} \in (0, 1)$ be all the distinct values for $P(\xi_i = 1|\mathbf{Y})$, $i = 1, \dots, N$, where n is less than or equal to N as there might be ties. We further let $p_{(0)} = 0$ and $p_{(n+1)} = 1$. Then by the form of local FDR in (6), it is easy to verify that $\text{fdr}_\zeta(Y)$ is a step function of ζ , where $\text{fdr}_\zeta(Y)$ is a constant in interval $[p_{(t-1)}, p_{(t)})$, for any $t = 1, \dots, n+1$. Therefore, $\text{fdr}_\zeta(Y)$ is left continuous in ζ .

We further show that $\text{fdr}_\zeta(Y) > \text{fdr}_{\zeta'}(Y)$, when $\zeta \in [p_{(t-1)}, p_{(t)})$ and $\zeta' \in [p_{(t)}, p_{(t+1)})$, for any $t = 1, \dots, n$. When $t < n$, we have

$$\text{fdr}_{\zeta'}(Y) = \frac{\sum_{i=1}^N (1 - P(\xi_i = 1|\mathbf{Y})) 1_{\{P(\xi_i=1|\mathbf{Y}) \geq p_{(t+1)}\}}}{\sum_{i=1}^N 1_{\{P(\xi_i=1|\mathbf{Y}) \geq p_{(t+1)}\}}},$$

and

$$\text{fdr}_\zeta(Y) = \frac{(\sum_{i=1}^N (1 - P(\xi_i = 1|\mathbf{Y})) 1_{\{P(\xi_i=1|\mathbf{Y}) \geq p_{(t+1)}\}}) + (\sum_{i=1}^N (1 - p_{(t)}) 1_{\{P(\xi_i=1|\mathbf{Y}) = p_{(t)}\}})}{\sum_{i=1}^N 1_{\{P(\xi_i=1|\mathbf{Y}) \geq p_{(t+1)}\}} + \sum_{i=1}^N 1_{\{P(\xi_i=1|\mathbf{Y}) = p_{(t)}\}}}.$$

As $1 - p_{(t)} > 1 - P(\xi_i = 1|\mathbf{Y})$ when $P(\xi_i = 1|\mathbf{Y}) \geq p_{(t+1)}$, $\text{fdr}_\zeta(Y) > \text{fdr}_{\zeta'}(Y)$. When $t = n$, it is easy to see that $\text{fdr}_\zeta(\mathbf{Y}) > \text{fdr}_{\zeta'}(\mathbf{Y})$ as $\text{fdr}_{\zeta'}(\mathbf{Y}) = 0$. This completes the proof for the properties of $\text{fdr}_\zeta(Y)$. The proof for the non-decreasing property of $\text{fdr}_\zeta(\mathbf{Y})$ is similar and thus is omitted here.

By the left-continuity of $\text{fdr}_\zeta(\mathbf{Y})$, we have

$$\text{fdr}_{\zeta^*(\mathbf{Y}; \rho)}(\mathbf{Y}) \leq \rho.$$

In addition, by the construction of $\zeta^*(\mathbf{Y}; \rho)$, $\zeta' > \zeta^*(\mathbf{Y}; \rho)$ for any $\zeta' \neq \zeta^*(\mathbf{Y}; \rho)$ also satisfying

$\text{fdr}_{\zeta'}(\mathbf{Y}) \leq \rho$. Then by the non-decreasing property of $\text{fnr}_{\zeta}(\mathbf{Y})$, $\text{fnr}_{\zeta'}(\mathbf{Y}) \geq \text{fnr}_{\zeta^*(\mathbf{Y};\rho)}(\mathbf{Y})$. Therefore, $\zeta^*(\mathbf{Y};\rho)$ solves the optimisation problem (8). \square

B A Parallel Tempering Algorithm

As mentioned in Section 4.2, the standard MCMC algorithms, such as the Metropolis-Hastings algorithm, suffer from slow-mixing for our problem, due to the presence of many discrete variables and the interactions between these discrete variables in the current problem.

Let Ξ be a generic notation for the parameters and hyperparameters to be sampled. Note that $\Xi = \{\theta_i, \xi_i, \beta_j, \eta_j, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta : i = 1, \dots, N, j = 1, \dots, J\}$ for the reduced model, and $\Xi = \{\theta_i, \xi_i, \tau_i, \beta_j, \eta_j, \alpha_j, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta, \gamma, \kappa : i = 1, \dots, N, j = 1, \dots, J\}$ for the full model, respectively. Recall that \mathbf{Z} is used as the generic notation for data, where $\mathbf{Z} = \mathbf{Y}$ and (\mathbf{Y}, \mathbf{T}) for the reduced model and the full model, respectively. We use $f(\Xi|\mathbf{Z})$ as a generic notation for the posterior distribution of interest. The goal is to sample Ξ from the target posterior distribution $f(\Xi|\mathbf{Z})$.

The algorithm involves sampling K MCMC chains with tempered target distributions. More specifically, let $0 < \psi_1 < \psi_2 < \dots < \psi_K = 1$ be a pre-specified sequence of temperature levels. Then the k th chain has a target distribution $f_k(\Xi|\mathbf{Z}) \propto (f(\Xi|\mathbf{Z}))^{(1/\psi_k)}$, where \propto means that the two sides differ by a normalising constant which does not depend on Ξ . The target distribution of the K th chain is our target posterior distribution. Let t be the current iteration number and $\Xi^{k,t}$ be the current samples from the k th chain. The parallel tempering algorithm performs the following steps in the $t + 1$ th iteration.

1. For each of the chains, sample $\Xi^{k,t+1}$ given $\Xi^{k,t}$ using a Metropolis-Hastings within Gibbs sampler, which will be further discussed below.
2. Randomly sample a pair of adjacent chains, k and $k + 1$, and use a Metropolis-Hastings update to decide whether to swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$. That is, a

Bernoulli random variable with success probability

$$\min \left\{ 1, \frac{f_k(\Xi^{k+1,t+1}|Z)f_{k+1}(\Xi^{k,t+1}|Z)}{f_k(\Xi^{k,t+1}|Z)f_{k+1}(\Xi^{k+1,t+1}|Z)} \right\}$$

is generated to decide whether to swap or not. If the Bernoulli random variable takes value 1, then we swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$ and otherwise, we reject the swap and keep their statuses unchanged.

For simplicity, the MCMC sampling within each chain is conducted by using a Metropolis-Hastings within Gibbs sampler. That is, Ξ is split into multiple blocks. Each block is sampled given all the others, using a random-walk Metropolis-Hastings sampler, for which the step size is tuned following Roberts and Rosenthal (2001) that is based on the Metropolis-Hastings acceptance rate. For the reduced model, Ξ is split into 10 blocks, including (1) $\theta_i, i = 1, \dots, N$, (2) $\xi_i, i = 1, \dots, N$, (3) $\beta_j, j = 1, \dots, J$, (4) $\eta_j, j = 1, \dots, J$, (5) δ , (6) π_1 , (7) σ_{11} , (8) π_2 , (9) μ_1 , and (10) ω_{11} . For the full model, Ξ is split into 14 blocks, including (1) $\theta_i, i = 1, \dots, N$, (2) $\tau_i, i = 1, \dots, N$, (3) $\xi_i, i = 1, \dots, N$, (4) $\beta_j, j = 1, \dots, J$, (5) $\alpha_j, j = 1, \dots, J$, (6) $\eta_j, j = 1, \dots, J$, (7) δ , (8) γ , (9) κ , (10) π_1 , (11) Σ , (12) π_2 , (13) $\boldsymbol{\mu}$, and (14) Ω .

The specification of the number and levels of the temperatures also needs some tuning. A fine-tuned system tends to have faster mixing. We suggest choosing the number and levels of the temperatures by following the theoretical guidance given in Atchadé et al. (2011) that is based on the Metropolis-Hastings acceptance rate.

C An Additional Simulation Study

We provide an additional simulation study under settings similar to Study I in Section 6, but with smaller values of J to mimic educational tests with relatively smaller item sizes.

Test taker					Item			
AUC	C.S1		C.S2		C.S1		C.S2	
	Reduced	Full	Reduced	Full	Reduced	Full	Reduced	Full
25%	0.934	0.936	0.956	0.962	0.937	0.944	0.954	0.961
50%	0.955	0.954	0.960	0.971	0.964	0.963	0.972	0.976
75%	0.969	0.967	0.971	0.978	0.973	0.975	0.978	0.981

Table 14: Overall classification performance based on the posterior means of ξ_i and η_j . For each model, each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 100 independent datasets.

C.1 Settings

We consider two settings, with (1) $N = 2,000, J = 50$, and (2) $N = 4,000, J = 100$. The rest of the simulation setting is exactly the same as that of Study I. These two settings are referred to as settings C.S1 and C.S2, respectively. For each setting, 100 independent data sets are simulated.

C.2 Results

The analysis is conducted using the parallel tempering MCMC algorithm described above. For each dataset, we run 10,000 iterations, with the first 3,000 iterations as the burn-in. The results are based on the posterior samples from the last 7,000 iterations.

Our results are given in Tables 14 through 17. The results are similar to those from Study I. Table 14 gives the AUC values for the classification of test takers and items. Tables 15 and 16 evaluate the performance of the local FDR control and local FNR control procedures for the classification of test takers and items, respectively. More specifically, Table 15 gives the corresponding FDP values when controlling the local FDR at 1%, 5%, and 10% levels for the test takers. Table 16 gives the FNP values when controlling the local FNR at 1%, 5%, and 10% levels for the items. Finally, the bias and variance for the posterior mean estimator are given in Table 17.

C.S1							C.S2					
FDP	Reduced			Full			Reduced			Full		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.008	0.032	0.089	0.008	0.035	0.085	0.007	0.028	0.072	0.008	0.029	0.069
50%	0.012	0.047	0.092	0.011	0.046	0.091	0.011	0.043	0.078	0.009	0.037	0.073
75%	0.013	0.051	0.098	0.013	0.053	0.099	0.013	0.047	0.084	0.012	0.044	0.087

Table 15: Local FDR control for test takers. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 100 independent datasets.

S1							S2					
FNP	Reduced			Full			Reduced			Full		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.033	0.063	0.007	0.032	0.061	0.007	0.029	0.059	0.006	0.031	0.062
50%	0.012	0.037	0.068	0.009	0.036	0.067	0.009	0.038	0.067	0.007	0.036	0.065
75%	0.013	0.046	0.071	0.010	0.043	0.069	0.012	0.045	0.072	0.012	0.041	0.071

Table 16: Local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 100 independent datasets.

C.S1							C.S2						
Reduced model							Reduced model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ		π_1	π_2	σ_{11}	μ_1	ω_{11}	δ
Bias	0.11	-0.02	-0.05	-0.11	0.17	0.22	Bias	0.05	0.15	-0.04	-0.13	0.20	-0.07
Variance	0.16	0.09	0.32	0.43	0.17	0.30	Variance	0.21	0.11	0.39	0.25	0.22	0.22
C.S1							Full model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ	
Bias	-0.11	-0.08	0.07	0.24	-0.08	0.11	-0.12	-0.04	-0.12	0.14	-0.08	-0.16	
Variance	0.17	0.19	0.32	0.32	0.21	0.30	0.11	0.17	0.15	0.23	0.00	0.37	
C.S2							Full model						
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ	
Bias	-0.04	0.06	0.08	-0.15	-0.17	-0.15	-0.04	-0.11	-0.15	0.11	-0.04	-0.11	
Variance	0.21	0.18	0.23	0.29	0.31	0.35	0.07	0.19	0.12	0.12	0.18	0.46	

Table 17: Accuracy of the posterior mean estimator of the global parameters. The bias and variance for the posterior mean estimator are calculated based on the 100 replications.

References

- Agresti, A. and Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52:119–126.
- Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37:3099–3132.
- Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011). Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, 21:555–568.
- Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13:171–187.
- Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, West Sussex, UK.
- Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, 50:141–163.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57:289–300.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 397 – 472. Addison-Wesley, Oxford, England.
- Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39:331–348.
- Boughton, K. A. and Yamamoto, K. (2007). A hybrid model for test speededness. In von Davier, M. and H. Carstensen, C., editors, *Multivariate and mixture distribution Rasch models*, pages 147–156. Springer, New York, NY.
- Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58:1–37.

- Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall, Boca Raton, FL.
- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39:83–87.
- Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970.
- Chen, Y., Lee, Y.-H., and Li, X. (2020). Item quality control in educational testing: Change point model, compound risk, and sequential detection. *arXiv preprint arXiv:2008.10104*.
- Chen, Y. and Li, X. (2019). Compound sequential change point detection in multiple data streams. *arXiv preprint arXiv:1909.05903*.
- Cho, S.-J., Suh, Y., and Lee, W.-y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, 35:48–61.
- Cizek, G. J. and Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge, New York, NY.
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62:7–28.
- Duncan, K. A. and MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, 8:41–66.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104.
- Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23:1–22.
- Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, Cambridge, UK.

- Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, 29:285–301.
- Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160.
- Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–534.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.
- Geyer, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors, *Handbook of Markov chain Monte Carlo*, pages 295–311. Chapman & Hall/CRC, Boca Raton, FL.
- Goegebeur, Y., De Boeck, P., Wollack, J. A., and Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73:65–87.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.
- Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54:761–771.
- Holland, P. and Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates, New York, NY.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.
- Katzgraber, H. G., Trebst, S., Huse, D. A., and Troyer, M. (2006). Feedback-optimized par-

- allel tempering monte carlo. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P03018.
- Kingston, N. and Clark, A. (2014). *Test fraud: Statistical detection and methodology*. Routledge, New York, NY.
- Kuha, J., Katsikatsou, M., and Moustaki, I. (2018). Latent variable modelling with non-ignorable item nonresponse: multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 181:1169–1192.
- Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin Co., New York, NY.
- Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press, Boca Raton, FL.
- Mavridis, D. and Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate behavioral research*, 43:453–475.
- Mavridis, D. and Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, 18:1016–1034.
- McLeod, L., Lewis, C., and Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27:121–137.
- Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge, New York, NY.
- Moustaki, I. and Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, 101:644–653.
- O’Leary, L. S. and Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In Cizek, G. J. and Wollack,

- J. A., editors, *Handbook of quantitative methods for detecting cheating on tests*, pages 151–163. Routledge, New York, NY.
- O’Muircheartaigh, C. and Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162:177–194.
- Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172.
- Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7:887–902.
- Poole, K. T. and Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*, 35:228–278.
- Poole, K. T., Rosenthal, H., and Koford, K. (1991). On dimensionalizing roll call votes in the US Congress. *The American Political Science Review*, 85:955–976.
- Quintero, A. and Lesaffre, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine*, 37:2440–2454.
- Ramsay, J. and Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, 56:365–379.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Nielsen and Lydiche, Copenhagen, Denmark.
- Reckase, M. (2009). *Multidimensional item response theory*. Springer, New York, NY.
- Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, 61:509–528.
- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:731–792.
- Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In Neyman, J., editor, *Proceedings of the second Berkeley symposium on*

- mathematical statistics and probability*, Berkeley, CA. University of California Press.
- Robbins, H. (1956). An empirical Bayes approach to statistics. In Neyman, J., editor, *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, pages 157–163. University of California Press, Berkeley, CA.
- Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16:351–367.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.
- Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, 27:163–179.
- Shao, J. (2003). *Mathematical statistics*. Springer, New York, NY.
- Shu, Z., Henson, R., and Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, 78:481–497.
- Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, 42:46–68.
- Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement*, 41:403–421.
- Skorupski, W. P. and Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In Cizek, G. J. and Wollack, J. A., editors, *Handbook of quantitative methods for detecting cheating on tests*, pages 214–231. Routledge, New York, NY.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–639.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 76:485–493.
- Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false

- discovery rate control. *Journal of the American Statistical Association*, 102:901–912.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72:287–308.
- Veerkamp, W. J. and Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25:373–389.
- Wang, C., Chang, H.-H., and Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66:144–168.
- Wang, C. and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68:456–477.
- Wang, C., Xu, G., and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, 83:223–254.
- Wang, X. and Liu, Y. (2020). Detecting compromised items using information from secure items. *Journal of Educational and Behavioral Statistics*, 45:667–689.
- Wirth, R. and Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, 12(1):58.
- Wollack, J. A. and Fremer, J. J., editors (2013). *Handbook of test security*. Routledge, New York, NY.
- Yuan, K.-H. and Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, 51:63–88.
- Yuan, K.-H. and Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, 54:161–175.
- Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, 31:379–390.
- Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool

- of a CAT system. *Applied Psychological Measurement*, 38:87–104.
- Zhou, Z., Li, X., Wright, J., Candès, E., and Ma, Y. (2010). Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1518–1522. IEEE.