



University of  
Massachusetts  
Amherst

## Detecting Candidate Preknowledge of Items Using A Predictive Checking Method

|               |   |
|---------------|---|
| Item Type     | Dissertation (Open Access)  |
| Authors       | Wang, Xi  |
| DOI           | <a href="https://doi.org/10.7275/8965479.0">10.7275/8965479.0</a>                                 |
| Download date | 2026-04-21 22:50:10   |
| Link to Item  | <a href="https://hdl.handle.net/20.500.14394/20067">https://hdl.handle.net/20.500.14394/20067</a> |

DETECTING CANDIDATE PREKNOWLEDGE OF ITEMS USING A PREDICTIVE  
CHECKING METHOD

A Dissertation Presented

by

XI WANG

Submitted to the Graduate School of the  
University of Massachusetts Amherst in partial fulfillment  
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Education

© Copyright by Xi Wang 2016

All Rights Reserved

DETECTING CANDIDATE PREKNOWLEDGE OF ITEMS USING A PREDICTIVE  
CHECKING METHOD

A Dissertation Presented

by

XI WANG

Approved as to style and content by:

---

Ronald K. Hambleton, Chair

---

Craig S. Wells, Member

---

Anna Liu, Member

---

Joseph B. Berger, Senior Associate Dean  
College of Education

## **DEDICATION**

To my dear grandma, who taught me to smile to all difficulties

## ACKNOWLEDGMENTS

Up until this moment, I still cannot believe five years have passed since I first came to the U.S. on August 21, 2011. Many people say one's early twenties are the best ages in one's life. If so, I am very lucky and happy to spend my best time in the graduate school to pursue my doctoral degree. The five years' study and life in the U.S. has been an exciting journey for me, also filled with bitter sweet moments sometimes. At the end of this journey, I would like to express the deepest gratitude to my dear professors, friends and family.

First of all, I would like to thank my advisor Prof. Ronald Hambleton, who is also the chair of my dissertation committee. As an advisor, Prof. Hambleton has always been very supportive for what I want to do, and he has been giving me helpful advice on my academic and career development. As an internationally renowned scholar, he is not only extremely knowledgeable of psychometrics, but also highly experienced with applying such knowledge to solve practical problems in the real world. Besides providing suggestions on solving specific problems in my study and research, he always encourages me to come up with research questions with practical significance. He is also very modest and always fully prepared for everything, which set a role model for me to never stop making things better.

I would also like to thank my other two committee members, Prof. Craig Wells and Prof. Anna Liu. Prof. Wells has also played a very important role in my academic development. I have been fortunate to have several opportunities to work with him during the past five years. He is smart, generous and has a very solid background in psychometrics and statistics. He is good at coming up with ideas in problem solving, and

pointing out new research directions, so our collaborations are always effective and productive. Without him, I will not be able to grow so fast in my academic development. My gratitude also goes to Prof. Anna Liu, who serves as the outside committee member for my dissertation. I would like to thank her for her time and suggestions on my dissertation.

This dissertation project is funded by Educational Testing Service (ETS) through the Harold Gulliksen Psychometric Research Fellowship. It is my great honor to be a recipient of this fellowship award. I would like to thank ETS for this funding opportunity. I am more than happy to have three excellent mentors from ETS: Drs Fred Robin, Hongwen Guo, and Neil Dorans. They have contributed significantly to the second study in this dissertation. My mentors and I had a lot of discussions on this project while I was in Princeton last summer, which was an invaluable experience for me. Although it was often hard to reach an agreement among us, it is through those discussions that we made things better and added more practical significance to this study.

My appreciation also goes to a special friend, Dr. Yang Liu, who has made significant methodological contributions to this dissertation. In fact, the idea of this dissertation project started from my conversations with Dr. Liu. Having had a solid training in both psychometric and statistical theories, he can always provide some effective suggestions whenever I got stuck on some problems. His attitude towards research - trying to provide rigorous justification to every detail – also sets a role model for me as a young scholar.

In addition to receiving academic guidance from different people, I am very lucky to also have support and love from my family and friends. REMP is a loving family, and

I am glad to be able to share my happiness and sorrow with my dearest friends Fen Fan, Joshua Marland, Hwanggyu Lim, HyunJoo Jung, Yooyoung Park, and Hongyu Diao. The time we spent together is the best time I had in graduate school. My gratitude also goes to all REMP professors as well as Peg Louraine and Emily Pichette. In addition, I appreciate the constant care and support from my ETS intern friends, Fei Chen, Huili Liu and Xin Luo, all of whom are so smart and humorous. Their words of comfort and encouragement always have the magic to cheer me up. I would also like to thank my parents and grandparents for raising me up, and for being understanding of my dreams. Unfortunately, my grandma was not able to see me finishing this journey, but I know I did not let her down and she would be proud of me as always.

Lastly, I would like to express my appreciation to my husband, Mengwei Li, who is also my best friend in life. Thank you for giving up other good opportunities to come here to be with me, and thank you for bringing so much happiness into my life.

Although my journey to the Ph.D. degree is almost completed, I know that learning will not come to an end as I walk out of graduate school. Instead, it has just started.

## **ABSTRACT**

# **DETECTING CANDIDATE PREKNOWLEDGE OF ITEMS USING A PREDICTIVE CHECKING METHOD**

SEPTEMBER 2016

XI WANG, B.S., BEIJING NORMAL UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

In on-demand high-stakes testing programs such as GRE and TOEFL, some items are repeatedly used across test administrations to reduce the cost of developing new items constantly. Item exposure provides an opportunity for examinees to have knowledge of particular test items in advance of their administration. It poses a threat to test security and ultimately will result in invalid test scores. Therefore, many testing programs conduct quality control to monitor test compromise at individual and/or group level. A predictive checking method is proposed in this study to detect examinee preknowledge on exposed items. We consider a scenario where a test can be divided into two subsets of items: one consisting of secure items with very low exposure rates and the other consisting of possibly compromised items (i.e. insecure items) which have been exposed for a while. An examinee's proficiency distribution is first obtained from secure items and then the predictive distribution for the examinee's test scores on the insecure items is constructed. The extent of test compromise is determined by comparing an individual's observed score on the insecure items with the predictive distribution. To evaluate the effectiveness of this approach, three studies are conducted: the first study investigates the statistical

properties (i.e. type-I error and power) of this method under four factors through Monte Carlo simulation; the second study applies this method to two simulated test compromise situations that are likely to happen in practice, and compares this method to three other detection approaches; the third study applies this method to a real dataset to demonstrate its practice use. Findings from the simulation studies suggest that the predictive checking method is effective in detecting examinees' preknowledge in the unsecure subset given a moderate to large test compromise rate, while maintaining its type-I error close to or lower than the nominal level. It also demonstrates similar or better performance than the other approaches under investigation. These results have implications for conducting quality control at individual examinee level in an on-demand testing program.

## TABLE OF CONTENTS

|   |      |
|---|------|
| ACKNOWLEDGMENTS .....   | v    |
| ABSTRACT .....  | viii |
| LIST OF TABLES .....  | xiii |
| LIST OF FIGURES .....   | xv   |
| CHAPTER   |      |
| 1. INTRODUCTION .....   | 1    |
| 1.1 Background .....  | 1    |
| 1.1.1 Introduction to the Problem of Test Security .....                  | 1    |
| 1.1.2 Test Security in Adaptive Tests .....                               | 3    |
| 1.2 Statement of the Problem .....  | 7    |
| 1.3 Research Purpose .....  | 11   |
| 1.4 Educational Significance .....  | 12   |
| 2. LITERATURE REVIEW .....  | 14   |
| 2.1 Overview of Person-fit Statistics .....                               | 15   |
| 2.2 Methods Specific to Detecting Item Preknowledge .....                 | 21   |
| 2.2.1 Methods not using information from secure items .....               | 21   |
| 2.2.2 Methods using information from secure items .....                   | 26   |
| 2.3 Summary of Existing Methods .....                                     | 39   |
| 3. METHODOLOGY .....  | 41   |
| 3.1 Predictive Checking Method .....                                      | 41   |
| 3.1.1 Mathematical Definition and Properties .....                        | 41   |
| 3.1.2 Implementation of Predictive Checking .....                         | 44   |
| 3.1.2.1 Estimation of $p(\boldsymbol{\theta} \mathbf{y}_1)$ from T1 ..... | 45   |
| 3.1.2.1.1 Bayesian Posterior Distribution .....                           | 45   |
| 3.1.2.1.2 Fiducial Distribution .....                                     | 46   |

|         |  |     |
|---------|--|-----|
| 3.1.2.2 | Sampling From $p(\theta y_1)$ .....                        | 49  |
| 3.1.2.3 | Test Statistics .....                                      | 50  |
| 3.1.2.4 | Item-set Level and Item Level Detection.....               | 51  |
| 3.2     | Likelihood Ratio Test .....                                | 53  |
| 3.3     | Adapted KL Divergence .....                                | 55  |
| 3.4     | Regression-based approach.....                             | 57  |
| 4.      | <b>STUDY 1: EVALUATION OF PREDICTIVE CHECKING</b> .....    | 60  |
| 4.1     | Study Design.....  | 60  |
| 4.2     | Data Simulation .....                                      | 62  |
| 4.3     | Evaluation Criteria.....                                   | 63  |
| 4.4     | Results.....   | 64  |
| 4.4.1   | Recovery of $\theta$ by Different Estimation Methods ..... | 64  |
| 4.4.2   | Type-I error at the item-set level .....                   | 65  |
| 4.4.3   | Power at the item-set level.....                           | 66  |
| 4.4.4   | Type-I Error at the item level.....                        | 68  |
| 4.4.5   | Power and False Positive Rate at the item level .....      | 69  |
| 4.5     | Discussion.....  | 72  |
| 5.      | <b>STUDY 2: COMPARISON OF METHODS</b> .....                | 74  |
| 5.1     | Background.....  | 74  |
| 5.2     | Shallow Pool Simulation.....                               | 75  |
| 5.3     | Key Exposure Simulation .....                              | 79  |
| 5.4     | Evaluation Criteria.....                                   | 83  |
| 5.5     | Results in Shallow Pool Simulation.....                    | 84  |
| 5.5.1   | Theta estimation error .....                               | 84  |
| 5.5.2   | Detection rate at the person-level .....                   | 88  |
| 5.5.3   | Detection rate at the group level .....                    | 95  |
| 5.6     | Results in Key Exposure Simulation .....                   | 99  |
| 5.6.1   | Person-level Detection Result.....                         | 99  |
| 5.6.2   | Group-level Detection Results.....                         | 102 |
| 5.7.    | Discussion.....  | 103 |
| 6.      | <b>REAL DATA APPLICATION</b> .....                         | 107 |
| 6.1     | Data Description .....                                     | 107 |
| 6.2     | Data Analysis .....  | 108 |

|   |     |
|---|-----|
| 6.3 Results.....  | 110 |
| 6.3.1 Detection rate.....                                 | 110 |
| 6.3.2 Classification Consistency.....                     | 112 |
| 6.3.3 Detection Characteristics by Different Methods..... | 113 |
| 7. DISCUSSION AND CONCLUSIONS .....                       | 116 |
| APPENDICES  |     |
| A. TABLES .....   | 124 |
| B. FIGURES .....  | 134 |
| BIBLIOGRAPHY.....   | 136 |

## LIST OF TABLES

| Table   | Page |
|---|------|
| 4.1: Bias and MSE from Different Estimation Methods.....  | 65   |
| 4.2: Empirical Type-I Error Using Fiducial and Jeffreys Prior .....                                 | 66   |
| 4.3: Power Rate Using Fiducial and Jeffreys Prior.....  | 68   |
| 4.4: Empirical Type-I Error at Item Level .....   | 69   |
| 4.5: Item-Level Power and False Positive Rate .....   | 71   |
| 5.1: True Item Parameters in Shallow Pool Situation in Study 2 .....                                | 77   |
| 5.2: Conditions in Shallow Pool Situation.....  | 78   |
| 5.3: Simulation Conditions in Key Exposure Situation .....  | 81   |
| 5.4: Summary of True Item Parameters in Key Exposure Generation .....                               | 82   |
| 5.5: BIAS and RMSE in Null Condition .....  | 85   |
| 5.6: Average $\theta$ Inflation ( $\theta - \theta$ ) in Test Compromise Conditions.....            | 86   |
| 5.7: Power with True Item Parameters .....  | 90   |
| 5.8: Power with Item Parameter Estimates from 3PLM.....   | 91   |
| 5.9: Power with Item Parameter Estimates from 2PLM.....   | 92   |
| 6.1: Detection Rate (EAP change) of Different Methods.....  | 110  |
| 6.2: Classification consistency among different methods.....  | 113  |
| A.1: True Item Parameters in Study 1 .....  | 124  |
| A.2: Empirical Type-I Error at Item-set Level Using Fiducial and Jeffreys Prior .....               | 126  |
| A.3: Empirical Type-I Error at Item-set Level Using Normal Prior.....                               | 127  |
| A.4: Power at Item-set Level Using Fiducial and Jeffreys Prior When<br>Compromise Rate is 100%..... | 128  |
| A.5: Power at Item-set Level Using Fiducial and Jeffreys Prior When<br>Compromise Rate is 60%.....  | 129  |
| A.6: Power at Item-set Level Using Normal Prior .....   | 130  |

|   |     |
|---|-----|
| A.7: Empirical Type-I Error at Item Level ..... | 131 |
| A.8: Empirical Power at Item Level .....        | 132 |
| A.9: False Positive Rate at Item Level.....     | 133 |

## LIST OF FIGURES

| Figure  | Page |
|---|------|
| 1.1: Example of a 1-3-3 MST design.....   | 6    |
| 5.1: Type-I error rate at person level .....  | 89   |
| 5.2: The left panel shows the $\theta$ posterior distributions on T1 and T2 in the condition with 40% compromised responses in both stages. The red, black and blue lines represent the use of true item parameters, 3PLM item parameter estimates and 2PLM item parameter estimates. The right panel shows the log posterior ratio between T1 and T2 when the three types of item parameters are used..... | 95   |
| 5.3: Type-I error rate of the three methods (PC=Predictive checking, KL=KL divergence, and LR=likelihood ratio) among different examinee ability groups.....  | 96   |
| 5.4: Detection power among examinees with ability distribution of $N(0,1)$ .....  | 97   |
| 5.5: power among examinees with ability distribution of $N(-1,1)$ .....   | 98   |
| 5.6: Person-level detection rate across different conditions .....  | 99   |
| 5.7: Distribution of standardized residuals at different $\theta$ levels. ....  | 101  |
| 5.8: Group-level detection rate across different conditions.....  | 102  |
| 6.1: Posterior distribution of $\theta$ for cases detected by different methods among modified examinees .....  | 114  |
| 6.2: Posterior distribution of $\theta$ for cases detected by different methods among unmodified examinees. ....  | 115  |
| 6.3: Scatterplot of summed score (left) and EAP (right) on T1 and T2.....   | 115  |
| B.1: Plots to check assumptions in simple linear regression in study 2. ....  | 134  |
| B.2: Plots to check assumptions in simple linear regression in real data analysis.....  | 135  |

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

#### 1.1.1 Introduction to the Problem of Test Security

With the rapid advance of information technology, computer-based testing (CBT) is gradually replacing the traditional paper-and-pencil tests and becoming the mainstream in large-scale assessments in the 21st century. A number of high-stake testing programs, such as the GRE, TOEFL and SAT, have been using computer-based testing for years. Currently, the two common-core assessment consortia-the Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced- are both administering their annual as well as their formative tests through computers. One of the advantages that CBT offers is the “on-demand” test scheduling, which means a test is offered on a large number of time slots within a testing window, and an examinee can choose to take the test at any available time slot. Although on-demand testing brings much convenience for test takers, it becomes more difficult for the testing agency to insure test security.

Due to the large number of testing administrations in an on-demand testing, it is practically impossible to use a different test form for each test administration. The cost for item development is high. For example, in a legal case about twenty years ago, Educational Testing Service (ETS) reported that it cost them about \$1000 to produce a quality test item; that figure is undoubtedly much higher today. Due to the high expenses of item development, test items are repeatedly used across different test administrations.

This provides an opportunity for examinees who take the test earlier to steal the items and then share them with future examinees. Of course, examinees are asked not to share items, and they are sometimes told that if caught, there can be seriously punished, but not all examinees follow the rules.

Examinees who take the test later can then use item preknowledge to gain score increases. Items could be stolen through some “spy” cameras that can be hidden in glasses, pens or watches (Wollack & Fremer, 2013, pp. 48), or could be stolen through earlier test takers’ memorization. The former type of item stealing could still be detected by well-trained proctors, while the later type can never be explicitly detected. The stolen items could be shared among friends, posted on the Internet, or distributed through some organized efforts. An example of organized item-theft efforts is the well-publicized 1994 ETS-Kaplan incident, where Kaplan Test Prep sent 20 of its employees to take the computerized version of the GRE to memorize as many items as possible and then reproduce the items later for some type of distribution to future examinees (Wollack & Fremer, 2013). Another example for item sharing occurred on the Graduate Management Admissions Test (GMAT). In 2008, the Graduate Management Admissions Council (GMAC) cancelled the scores of 84 students as they found those students had access to some stolen items that were posted on a website (Hechinger, 2008). GMAC also provided an example for an item posted on the website and the actual item used in the test, and showed that the memorized items had contained most of the information from the actual test item.

### 1.1.2 Test Security in Adaptive Tests

In this section, discussion is focused on test security problems specific to adaptive test designs, as adaptive tests have been widely applied in many high-stake testing programs in recent years. The goal of an adaptive test is to “tailor” test items to the examinee’s ability level. The idea of using an adaptive test can be traced back to the Binet–Simon (1905) intelligence test, where the test questions were adapted to the estimate of an examinee’s mental age based on the examinee’s responses to earlier test questions. It is the use of computer-based testing that makes it possible to widely implement adaptive test in large-scale, high-stakes assessments. Adaptive tests can be further categorized into two designs according to the level of adaption: item-level adaptive tests and module-level adaptive tests. The former is often known as a *computerized adaptive test* (CAT), while the latter is known as a *multistage adaptive test* (MST). CAT has been implemented in large-scale testing programs for decades (e.g. van der Linden & Glas, 2010). Examples include the Armed Services Vocational Aptitude Battery (ASVAB) by the U.S. Department of Defense, the nurse licensure and certification exam (NCLEX/CAT) by the National Council of State Boards of Nursing, and the early CAT-versioned Graduate Record Examination (GRE) by ETS. Currently, CAT is being applied in SMARTER Balanced common-core assessments as well. MST, as a compromise between CAT and fixed-form test, has received increasing interest in recent years, and it is now adopted by several operational testing programs including the revised GRE General test and the Certified Public Accountant Exam.

In CAT, an examinee’s proficiency is estimated after each item is administered and the selection of the next item is based on the current proficiency estimate and on

constrains for item content and exposure. In CAT, different examinees typically are administered different items. This provides an advantage for test security as examinees taking the test later may not get exactly the same items as earlier examinees, and thus the overall item exposure rate is reduced. However, since CAT aims to select items that can provide most statistical information for an examinee's proficiency, items with the better psychometric properties tend to be selected more often than others. This results in uneven exposure rates among different items. In addition, examinees at similar proficiency levels are likely to receive the same items, albeit in a different order. Therefore, when item-sharing is among friends, who are more likely to be of similar proficiency levels, even some items with low overall exposure rates (among the examinee population) can be compromised. Also, in organized item-theft efforts, thieves typically targeted at items with middle to high difficulty levels, so items that are more likely to be administered to high-proficiency examinees will have higher exposure rates (Stocking & Lewis, 1998).

The high exposure or conditional exposure rate of some items at a certain proficiency level leaves the item pool vulnerable to item-theft. The simulation study conducted by McLeod, Lewis, and Thissen(2003) illustrated how quickly an item pool could become compromised with an organized item-theft effort. They simulated the item memorization-sharing strategy in a 28-item CAT: a group of source examinees were simulated to take the test and memorize all the items administered to them, and then a list of memorized items was shared with beneficiary examinees who would later take the test with item preknowledge. Their results showed that when eight sources were used, approximately 125 items out of 494 items in the pool were compromised, and the beneficiary examinee could receive an average of about 18-19 memorized items in a later

test administration, which could result in an average score gain of 30 points out of 60 total points for low-proficiency examinees, and an average score gain of 15-20 points for medium-proficiency examinees.

Different from CAT, an MST administers a series of sets of items adaptively to examinees (e.g. Yan , von Davier, & Lewis, 2014). Within each item set, which is called a *module* (also called *item block* or *testlet*), items are fixed and administered linearly to examinees (perhaps in some applications the items within a module may even be administered randomly to examinees as an additional way to enhance test security). An MST design consists of two or more stages and each stage could consist of one or more modules. At each stage, the module whose difficulty level is the most approximate to an examinee's proficiency level is administered, subject to the routing rules implemented. Figure 1.1 below shows an example of a three-stage MST design. Stage 1 consists of a module of moderate difficulty, and all examinees receive this module. Stage 2 and 3 both consist of three modules of different difficulty levels. Based on an examinee's proficiency estimate from a previous stage, one of the modules in Stage 2 or 3 is administered adaptively to the examinee. In practice, numerous parallel forms are constructed for a given module to ensure the maximum exposure of each module does not exceed a particular rate (Luecht, 2003). In this way, the item exposure rate in an MST can be controlled prior to test administration simply by specifying the number of parallel modules.

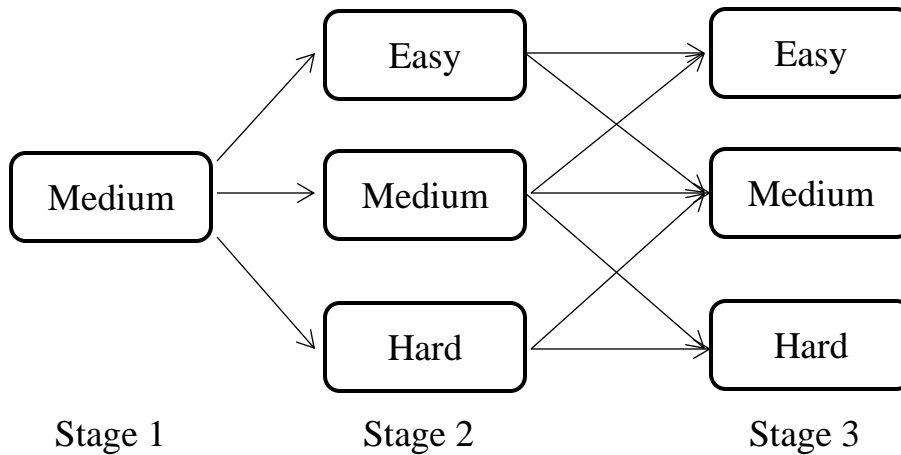


Figure 1.1: Example of a 1-3-3 MST design

Although it is easier to control item exposure rate in a MST design compared to a CAT design, the study by Wang, Zheng, and Chang (2014) suggests the use of MST may create a less secure condition than CAT in the circumstance of item-sharing and organized item theft, if an entire MST module is repeated used. To quantify test security, they used two types of statistics: the mean and standard deviation (SD) of test overlap rate among all possible pairs of examinees, while the test overlap rate is the proportion of common items shared by any two examinees. With both analytical and simulation results, they showed that the mean test overlap is approximately the same in both CAT and MST, but the SD of test overlap rate is always larger in MST. A large SD means certain groups of examinees share a larger number of common items than others, and thus the test overlap rate in MST tends to be more extreme in certain groups than that in CAT. A more intuitive understanding for this is that since modules are used repeated in different test administrations in MST, examinees who are administered the same module(s) will share the entire module(s) in common, while examinees who are administered different module(s) will share no items in common. Wang, Zheng, and Chang (2014) further demonstrated the adverse effect of large SD by simulating an organized item-theft

scenario. They found that on average, a future examinee could receive more compromised items that are memorized by earlier test takers in MST than in CAT, which ultimately led to larger misclassification rate in MST.

## **1.2 Statement of the Problem**

Item preknowledge forms a big threat to test security and ultimately could result in invalid test scores for many examinees. When there are a lot of compromised items or when a large number of examinees have item preknowledge, measurement accuracy and validity will be severely jeopardized and the invalid decisions or inferences made based on test scores will cause negative consequences for both the testing program and the stake-holders.

Due to the concern for the potential destructive consequences that item preknowledge could have on test score validity, many testing programs have devoted a lot of effort to reduce the likelihood of examinees gaining prior knowledge on test items. Some preventive procedures include controlling the item exposure rate using some exposure-control algorithms in item-selection (e.g., Georgiadou, Triantafillou, & Economides, 2007), increasing the size of the item bank, and reducing the testing window size. In addition to using preventive procedures, post-hoc analyses are often conducted as a quality control tool to monitor test compromise at the individual or the group level. Post-hoc analyses are often based on statistical methods and they can be implemented after or even during the test administration. By detecting item preknowledge at the individual or the group level, on one hand, one can evaluate the severity of test compromise in the entire examinee population or in a specific subpopulation, so that some actions, such as altering test designs or adjusting testing windows, can be taken in

time to enhance test security in a certain examinee population. On the other hand, if there is strong statistical evidence showing an examinee has used preknowledge on a large number of items in a test administration, a testing agency can conduct further investigations to make a decision on score cancellation, so as to ensure the accuracy and validity of test scores.

Numerous statistical procedures can be used to detect examinee preknowledge on test items. A comprehensive review of different methodologies is provided in Chapter 2. One type of method is to conduct person-fit analysis for an individual's response vector. Various person-fit statistics (e.g., Meijer & Sijtsma, 2001; Karabatsos, 2003), have been designed to detect response patterns that are inconsistent with the measurement model (called *aberrant responses*). Since item preknowledge typically results in a type of aberrant responses where examinees make correct responses on items that they would not have answered correctly based on their proficiency alone, person-fit analysis can be applied to detect item preknowledge. However, person-fit statistics share some problematic features that, to date, have limited their effectiveness. First of all, the calculation of many person-fit statistics, especially item response theory (IRT)-based statistics, typically relies on estimates of examinees' proficiency, which is usually biased by the involvement of aberrant responses in determining the proficiency estimates. When there are a large proportion of aberrant responses, the bias in the examinee's proficiency estimate may affect the power of the person-fit statistic to a large extent. Second, users of person-fit statistics often want to conduct hypothesis testing to see if there is a significant statistical difference between a person's response vector and the expected response vector under the null (i.e., model-fit) hypothesis. This requires the knowledge of the sampling

distribution for a person-fit statistic. Some statistics have known asymptotic or exact null distributions, but most do not. In addition, research has shown that the empirical null distribution of some statistics deviate from their theoretical asymptotic distributions when the number of items is relatively small (e.g. Li & Olejnik, 1997; Reise, 1995).

The first problem could be addressed if one draws information about an examinee's proficiency from a known subset of items on which the examinee most likely does not have preknowledge. These items could come from items that have never been exposed before (i.e. secure items), such as the pretest items on an operational test, or could come from pretested items that have rarely been exposed in a certain examinee population. The information from the secure subset of items can be used to infer the extent to which an examinee uses item preknowledge on a subset of possibly compromised items (i.e. unsecure items). The choice of unsecure item subset may depend on one's prior information. For example, a testing agency may be able to find the operational items that are posted online, so those items could form the unsecure item-set. The choice may also depend on likely item exposure scenarios in specific test designs. For example, in MST, if an entire module is reused in different test administrations, two examinees may share the same module if they have similar proficiency levels. Therefore, exposure will occur at the module level, and thus one module could form the unsecure subset. In another circumstance, if modules are re-assembled in different administrations but the same items are used for module re-construction, exposure will occur at item level. Since it is uncertain which items are compromised, the entire operational section can be used to form the unsecure subset. In CAT, item preknowledge is more likely to happen on items with a high exposure/conditional exposure rate. So the unsecure subset could

consist of high-exposure items, while the low-exposure items can be added to the set of secure items.

The second problem could be addressed by constructing the empirical distribution for a certain statistic through simulation, instead of relying on the exact or asymptotic distribution (e.g., Meijer & Nering, 1997; Nering, 1997; Reise, 1995). The empirical distribution is usually constructed by simulating response data according to an IRT model, and then computing the statistic based on each simulated response dataset. Since the true item or person parameters in an IRT context are unknown in practice, response data are often simulated based on the point estimate of item or person parameters. However, using point estimates does not take into account the uncertainty in the estimated item or person parameters, especially when the sample size on which item calibration or proficiency estimation is carried out is small. A better way to account for estimation error is to use the distribution of the estimated parameters.

To address the two limitations above, a *predictive checking* method (Geisser, 1993) is proposed in this study. The predictive checking method first draws inferences about an individual's proficiency parameter (i.e. person parameter) from responses on the secure subset of items. This will provide a valid baseline of the examinee's performance. Then predictions are made for the individual's responses on the unsecure subset of items based on the distribution of the estimated person parameter. An individual's observed response vector on the unsecure subset is then compared to the predicted responses through a test statistic. There are several advantages to conduct predictive checking to detect misfitting responses. First of all, there is no need to know the asymptotic distribution of a test statistic, as the sampling distribution of the statistic will be

constructed empirically. Second, the construction of the sampling distribution takes the uncertainty about estimated parameters into account. Third, predictive checking is a general method to evaluate model fit. In this study, it is applied to the detection of item preknowledge specifically, but it can be used to detect other types of aberrant responses, such as comparing examinees' performance on the last few items and on items in the earlier stage of the test to detect test speededness. Predictive checking can be conducted not only on item responses, but also on item response times, as long as the model for generating item responses/ response times is known. It can be implemented to detect misfit on a set of items and also on an individual item. As will be seen in Chapter 3, this method is very flexible and easy to implement.

### **1.3 Research Purpose**

The fundamental research question in this study is how effective the predictive checking method is to detect item preknowledge on exposed items in terms of its type-I error and power. Three studies were conducted in sequence to answer this research question. First of all, a simulation study was conducted to understand the statistical properties of the predictive checking method. The type-I error and power of this method were systematically investigated by manipulating four factors – the number of items in the secure and insecure subset, the proportion of truly compromised items in the insecure subset, and the estimation method to obtain the distribution of an individual's proficiency parameter. This method was applied to detect item preknowledge on a set of known exposed items, and on each individual item, so its effectiveness at both the item-set level and the item-level was evaluated. Secondly, this method was compared with three other methods to detect test compromise in two simulated situations that are likely

to happen in reality: one in an MST design, and the other in a fixed form test design. Lastly, a real data analysis was conducted to demonstrate the practical use of the predictive checking method and to investigate its detection consistency with the three methods considered in the second simulation study.

#### **1.4 Educational Significance**

The educational significance of this study can be seen from two perspectives. From the practical perspective, as there is an increasing use of on-demand testing programs in large-scale high-stake assessments, test security is of primary concern. The method proposed in this study can contribute to building a forensic monitoring system in on-demand testing programs. This method can be used as a quality-control post-administration analysis to identify potential test security problems in specific examinee subgroups. It could also be implemented during test administration if the test is administered on computer. As McLeod and Lewis (1999) and McLeod, Lewis and Thissen (2003) suggested, one way to rescue the test administration after an examinee is suspected of using item preknowledge is to administer some highly secure items. This could ensure measure accuracy to a certain extent so as to create a fairer testing environment. In addition, by detecting compromised item-sets/items, one can also expect to get more accurate proficiency estimate from uncompromised items.

From a methodology perspective, on one hand, the evaluation of the predictive checking method and its comparison with other existing methods contribute to the literature of detecting test fraud using statistical methods, and it has methodological implications for choosing the appropriate method in different situations. On the other hand, the predictive checking method can be used as a general method to detect other

types of aberrant responses, so it is hoped that this study can contribute to the person-fit literature and provide new insight for more effective detection of person misfit.

The rest of the study is organized as follows: Chapter 2 provides a literature review of existing methodologies to detect person-level aberrant responses both due to general misfit and specifically due to item preknowledge, with an emphasis on the latter. Chapter 3 describes the technical details for the implementation of the predictive checking method, and introduces three other methods to compare to the predictive checking. Chapter 4 summarizes the simulation study conducted to evaluate the statistical properties of the predictive checking method, including the simulation design, results and discussions for the results. Similarly, Chapter 5 summarizes the simulation study on the comparisons of different methods. Chapter 6 provides information about the nature of real data and summarizes the detection results by applying different methods to the real dataset. Lastly, Chapter 7 provides a general discussion on the implications of the findings, the limitations of the current study and possible directions of future study.

## **CHAPTER 2**

### **LITERATURE REVIEW**

In this chapter, methods that can be applied to detect cheating due to item preknowledge are reviewed. Specifically, this chapter is organized into the following three sections:

- (1) In the first section, a brief overview of person-fit statistics is provided. Person-fit statistics are reviewed as they are typically used to detect response patterns that are inconsistent with the measurement model, and cheating responses are a specific type of inconsistent responses. Therefore, in theory, person-fit statistics can be directly applied to detect item preknowledge. Some problematic features with person-fit statistics as well as potential problems of using person-fit in detecting item preknowledge is discussed in the review.
- (2) In the second section, methods that have been proposed to specifically focus on detecting examinee aberrant responses due to item preknowledge are reviewed in details. The rationale and technical details of each method are described, and the studies conducted to evaluate the effectiveness of each method are summarized. The advantages and disadvantages of each method are discussed.
- (3) In the third section, a summary based on the literature review is provided. The characteristics of existing methods are summarized, providing the justification for the development of the new method in this study.

## 2.1 Overview of Person-fit Statistics

Person-fit methods refer to a set of statistical methods for evaluating the fit of a person's response vector on a set of items to a measurement model or to other response patterns in a sample of people (Meijer & Sijtsma, 2001). Misfitting response vector usually occurs when an individual's responses are affected by some construct irrelevant factors, which are often called aberrant response behaviors, such as careless responding, test speededness, warm-up behavior (i.e. incorrect responses on the items at the beginning of a test due to the problem of getting started), etc. There exist over thirty statistics in the person-fit literature. Depending on whether an item response theory (IRT; see, Hambleton, Swaminathan, & Rogers, 1991) model is assumed to fit an individual's item responses, person-fit statistics can be classified as nonparametric and parametric.

Most non-parametric statistics measure the deviation of an individual's response pattern to the "Guttman perfect pattern". A Guttman pattern does not permit a correct response on a relatively difficult item with an incorrect response on a relatively easier item. Therefore, for a person with summed score  $r$  out of a total of  $n$  items (consider dichotomously scored items only), a "Guttman perfect pattern" should only consist of correct responses on the  $r$  easiest items. Examples of non-parametric statistics include the  $G$  statistic (Guttman, 1944,1950) and normed  $G$  (van der Flier, 1977), which count the number of item response pairs that do not conform to Guttman pattern; person point-biserial correlation (Donlan & Fischer, 1968), which is the correlation between an individual's response vector and a vector of proportion correct across persons on each item; caution index  $C$  (Sato,1975) and modified caution index (Harnisch & Linn, 1981), which are based on the ratio of two covariances- one between an individual's response

vector and a vector of proportion correct, and the other between the Guttman perfect pattern and a vector of proportion correct; agreement ( $A$ ), disagreement ( $D$ ), and dependability ( $E$ ) indices (Kane & Brennan, 1980), which are based on the sum of item scores weighted by the proportion correct on each item and the maximum sum which is achieved when the response pattern is the Guttman perfect pattern);  $U3$  and standardized  $U3$  (van der Flier, 1980) which are based on the sum of item score weighted by the log-ratio of the proportion correct on each item, and the sum of log-ratio of proportion correct over  $r$  easiest items as well as the sum of log-ratio of proportion correct over  $r$  hardest items;  $H^T$  (Sijtsma, 1986), which measures the similarity between an individual's response vector to the response vectors of the remaining sample.

Among all non-parametric statistics, only  $U3$  and standardized  $U3$  have known asymptotic sampling distributions, which are asymptotically normal, so critical values from a normal distribution can be used to classify misfitting response patterns when using  $U3$  or standardized  $U3$ . For the rest of the non-parametric statistics, their sampling distributions are unknown, so the significance probability for an observed value of a given statistic cannot be determined. This may not be a serious problem for using these statistics as descriptive measures, but it limits the usefulness of these statistics in hypothesis testing.

In contrast to non-parametric statistics, parametric statistics compare an individual's response pattern to the expected pattern under an IRT model. An IRT model specifies the probability of an individual with proficiency  $\theta$  correctly responding to a dichotomous item  $i$  (i.e.,  $P_i(\theta)$ ) by

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))} \quad (1)$$

where  $a_i$  is the item discrimination parameter,  $b_i$  is the item difficulty parameter,  $c_i$  is the pseudo-guessing parameter. By setting  $a_i = 1, c_i = 0$  for all items, one-parameter logistic model (1PLM) or the Rasch model (1960) is obtained. By setting  $c_i = 0$  for all items and allowing  $a_i$  and  $b_i$  to vary across items, the two-parameter logistic model (2PLM) is obtained. By further removing the constraints for  $c_i$ , the three-parameter logistic model (3PLM) is obtained.

IRT-based parametric statistics can be further categorized as residual-based statistics, likelihood-based statistics, caution indices, and optimal statistics. Residual-based statistics are based on the mean squared residuals across a set of items. For example, the statistic  $U$  is the average squared residuals, and  $W$  is the sum of squared residuals weighted by the sum of variances on a set of items (Wright & Stone, 1979). The standardized version of  $U$  and  $W$  (i.e.  $ZU$ , and  $ZW$ ) were also developed (Wright & Masters, 1982) to remove the dependency of their distribution on  $\theta$  levels, and both standardized statistics were claimed to have an asymptotically standard normal distribution. However, both  $ZU$  and  $ZW$  were found to be poorly standardized (Drasgow, Levine, McLaughlin, 1987; Noonan, Boss, & Gessaroli, 1992). Poor standardization means the same value of a statistic can be classified as good fit for some  $\theta$  levels but as poor fit for other  $\theta$  levels if a single critical value is used for different  $\theta$  levels.

Likelihood-based statistics include  $l_0$  (Levine & Rubin, 1979) which is simply the log-likelihood function, and its standardized version  $l_z$  (Drasgow, Levine, & Williams, 1985), which follows an asymptotic standard normal distribution;  $M$  statistic (Molenaar

& Hoijsink, 1990, p.96), which is the term in  $l_0$  that depends on the response pattern under the Rasch model; normalized jackknife variance estimate ( $JK$ ) and the ratio of observed and expected information ( $O/E$ ; Drasgow, et al., 1987) which measure the flatness of the likelihood function.  $l_z$  is most widely used in the person-fit literature, and it has been demonstrated to perform at least as well as or better than many other person-fit statistics (e.g. Drasgow, et al., 1987; Li & Olejnik, 1997; Nering & Meijer, 1998). However, several studies (e.g. Li & Olejnik, 1997; Reise, 1995; van Krimpen-Stoop & Meijer, 1999) have shown that the standard deviation of the empirical distribution of  $l_z$  is less than 1 when  $\hat{\theta}$  is used at short to moderate test lengths (i.e. less than 60 items), and the empirical distribution of  $l_z$  differed across different  $\hat{\theta}$  values. In addition, a larger difference between the empirical and asymptotic distribution is observed in adaptive test designs. Snijders (2001) proposed  $l_z^*$  to correct the decreased variance of  $l_z$ , and van Krimpen-Stoop and Meijer (1999) showed  $l_z^*$  could make a difference in correcting reduced variance in a short fixed form test, but the empirical distribution of  $l_z^*$  still deviated from the standard normal distribution in CAT.  $JK$  and  $O/E$  are well standardized but they are insensitive to misfitting responses (Drasgow et al., 1987).

Caution indices (Tatsuoka & Linn, 1983) under IRT modeling are extensions of the caution index in the non-parametric framework. However, instead of comparing an individual's response vector to the proportion correct across persons on a set of items, IRT-based caution indices compare a response vector to the IRT model-implied probability. Caution indices of  $ECI2$ ,  $ECI3$  compare an individual's response vector to the mean probability across persons on a set of items, while indices of  $ECI4$ ,  $ECI5$ ,  $ECI6$  compare an individual's response vector to his/her probability on a set of items. Tatsuoka

(1984) also derived the standardized form for  $ECI1$ ,  $ECI2$ ,  $ECI4$ ,  $ECI5$ , but their theoretical sampling distributions are unknown.

All statistics above test the fit of a response vector to a model in a general sense, without assuming a particular misfitting behavior for the misfitting responses (e.g. cheating responses, violation of local independence). To test the null model against an alternative model for a particular type of aberrant responses, several optimal statistics are proposed. They are called optimal in the sense that they can achieve the highest detection power at the same type-I error rate among all methods. By specifying a model for the misfitting behavior in advance, Levine and Drasgow (1988) used a likelihood ratio statistic to compute the ratio between the likelihood of a response vector under a misfitting model, and the likelihood under the IRT model. Klauer (1991) tested the invariance of an individual's proficiency over subtests under a Rasch model by using a two-parameter exponential family to model misfitting responses with an extra person parameter  $\eta$  that represents the difference between  $\theta$ 's on two subtests ( $\eta = \theta_1 - \theta_2$ ) and testing  $H_0: \eta = 0$  against  $H_1: \eta \neq 0$ .

Although the idea of optimal detection rate sounds appealing, the use of these optimal methods in the cheating problem considered in this study may be limited. For instance, to use the likelihood ratio statistic, specifying the right model for the misfitting responses is necessary for obtaining the optimal power, but the right cheating model is hardly known in practice. For the invariance test by Klauer (1991), although the invariance problem seems similar to the test compromise problem considered here, the two problems may not be simply regarded as equivalent to each other. For misfit due to  $\theta$  invariance, there is a systematic difference in  $\theta$  between the two subtests, which means  $\theta$

is changed by the same amount on all items in one subtest. However, for misfit due to the cheating problem considered in this study, first of all, depending on the exposure scenario and how the secure and insecure sections are formed, it's possible that some items in the insecure section are not compromised. Therefore, assuming  $\theta$  is changed on every item in the insecure section is unreasonable. Second, even when all the items in the insecure section are compromised, the amount of change in  $\theta$  on a particular item depends on a person's memorization of that item, so the assumption that  $\theta$  is changed by the same amount on all items seems too strong to be realistic. Based on the two arguments above, the optimal detection property may not hold when the invariance test is applied to the item-preknowledge detection here.

The overview of person-fit statistics above suggests that although a lot of person-fit statistics have been proposed in the literature, the effectiveness of many statistics may be limited by poor standardization, lack of a theoretical sampling distribution, as well as discrepancy between the empirical and asymptotic distribution. In addition, most statistics do not assume a particular type of aberrant responses, and for the optimal statistics that assume a specific aberrant responding behavior, the optimal detection rate may not be achieved in the problem in the present study due to the difficulty of specifying the right model for responses under item preknowledge. Therefore, the usefulness of person-fit statistics may be limited in detecting item preknowledge in particular. Other than using person-fit statistics, several other methods have been proposed to specifically focus on detecting item preknowledge.

## 2.2 Methods Specific to Detecting Item Preknowledge

A detailed review of methods specific to detecting item preknowledge is provided in this section. First of all, two methods that do not draw information from secure items are reviewed, and then followed by methods that utilize information from secure items.

### 2.2.1 Methods not using information from secure items

$Z_C$ . McLeod and Lewis (1999) proposed using a residual-based statistic,  $Z_C$ , to detect item preknowledge.  $Z_C$  is based on the standardized residual between an observed response (0/1) on each item and the probability of a correct response based on an IRT model. Instead of averaging the residual across all items in a test,  $Z_C$  divides the items into three categories according to their difficulty levels: easy, medium and difficult, and computes the residual difference between easy and difficult items. The formula for  $Z_C$  is

$$Z_C = \frac{\overline{Easy}[P_i(\hat{\theta}) - u_i] - \overline{Difficult}[P_i(\hat{\theta}) - u_i]}{\sqrt{\left\{ \sum_{Easy} \frac{\{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\}}{n_{Easy}^2} \right\} + \left\{ \sum_{Difficult} \frac{\{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\}}{n_{Difficult}^2} \right\}}} \quad (2)$$

where  $u_i$  is the response on item  $i$ ,  $P_i(\hat{\theta})$  is the probability of a correct response on item  $i$  under an IRT model,  $n_{Easy}$  is the number of easy items, and  $n_{Difficult}$  is the number of difficult items. Large positive values of  $Z_C$  indicate the examinee does not answer easy items correctly but answers hard items correctly, implying a misfit response pattern.

The expected value of the numerator of  $Z_C$  is 0 in model-fit condition (since  $E(u_i) = P_i(\theta)$  in model-fit condition), and the two summation terms in the denominator each correspond to the residual variance on one type of item, so  $Z_C$  is a standardized

statistic. Analytically,  $Z_C$  has an asymptotic standard normal distribution when the Lindeberg condition (e.g. Billingsley, 1986) is satisfied, so critical values from standard normal distributions were used to flag misfitting responses.

McLeod and Lewis compared  $Z_C$  to two other statistics-  $l_z$  and  $ECl4_z$  in detecting item preknowledge in CAT. They simulated item preknowledge on 50 relatively difficult items – the items that are most frequently exposed to the top 5% examinees – out of a bank consisting of 348 items, and evaluated the effectiveness of the statistics at two test lengths- 10 items and 28 items. They compared the conditional mean of each statistic (conditional on  $\theta$ ) between the null condition and item-preknowledge condition, and the distributional differences of each statistic between the null group and the cheating group. Their findings suggested that none of the three statistics were well standardized when  $\hat{\theta}$  was used in the calculation – the mean of each statistic was less than 0 and the standard deviation was less than 1 in the null condition, indicating using a normal approximation for each statistic is inappropriate in short tests. They found that  $Z_C$  demonstrated larger distributional differences between the null and cheating group than the other two statistics, but the marginal power analysis showed that  $Z_C$  only had slightly larger power than the other two statistics when the false alarm rate was smaller than 0.025, and all three statistics generally had low power- for example, their power was lower than 0.2 at the false alarm rate of 0.05. In addition, McLeod and Lewis (1999) pointed out using  $Z_C$  may be problematic in CAT since some examinees were not administered any easy or difficult items, which made it impossible to compute  $Z_C$  for those examinees. For instance, in their simulation,  $Z_C$  could not be computed for 481 out of 1,650 examinees

when the calculation was based on 28 items, and the number increased to 565 when only 10 items were used for calculation.

**A Bayesian Method.** McLeod et al. (2003) proposed a Bayesian posterior log-odds ratio approach to detect item preknowledge in CAT. The posterior log-odds ratio is defined as

$$\log\left[\frac{p(s = 1|u_1, \dots, u_n)/[1 - p(s = 1|u_1, \dots, u_n)]}{p(s = 1)/[1 - p(s = 1)]}\right] \quad (3)$$

where  $s$  denotes an examinee's item preknowledge status,  $s = 1$  means the examinee has had preknowledge on items in a certain bank, and  $s = 0$  means the examinee does not have item preknowledge;  $u_i$  is the response on item  $I$ ; so  $p(s = 1|u_1, \dots, u_n)$  is the posterior probability that an examinee has item preknowledge, and  $p(s = 1)$  is the prior probability. To calculate the posterior probability, the likelihood of item responses given an examinee has item preknowledge needs to be known. McLeod et al. (2003) defined the probability of a correct response given an examinee is using item preknowledge as

$$p(u_i = 1|s = 1, \theta) = 1 * p(m_i) + p(u_i = 1|\bar{m}_i, \theta) * p(\bar{m}_i) \quad (4)$$

where  $m_i$  denotes item  $i$  has been memorized and  $\bar{m}_i$  denotes item  $i$  has not been memorized. Equation (4) breaks the probability of a correct response into two components: one is the probability of responding correctly due to item memorization (i.e. if the examinee is administered an item that s/he has memorized) and the other is the probability of responding correctly due to the examinee's real proficiency (i.e. if the examinee is administered an item that s/he has not memorized before).  $p(u_i = 1|\bar{m}_i, \theta)$  in equation (4) is simply the probability defined by the standard IRT model specified in equation (1).  $p(m_i)$  is the probability that item  $i$  has been memorized. McLeod et al.

defined  $p(m_i)$  with three classes of models: the first class assumes  $p(m_i)$  to be a constant, the second class assumes  $p(m_i)$  to be a function of item difficulty, and the third class defines  $p(m_i)$  empirically using the exposure rate of item  $i$  among a certain number of examinees who are trying to steal the items obtained from a simulation study.  $p(\bar{m}_i)$  is simply  $1-p(m_i)$ .

The posterior probability that an examinee uses item preknowledge is

$$\begin{aligned}
 p(s = 1|u_1, \dots, u_n) = & \\
 & \int p(u_n|s = 1, \theta)p(s = 1, \theta|u_1, \dots, u_{n-1})d\theta \times \\
 & [\int p(u_n|s = 1, \theta)p(s = 1, \theta|u_1, \dots, u_{n-1})d\theta + \\
 & \int p(u_n|s = 0, \theta)p(s = 0, \theta|u_1, \dots, u_{n-1})d\theta]^{-1}
 \end{aligned} \tag{5}$$

where  $p(u_n|s = 1, \theta)$  is specified in equation (4),  $p(u_n|s = 0, \theta)$  is specified in equation (1);  $p(s = 1, \theta|u_1, \dots, u_{n-1})$  or  $p(s = 0, \theta|u_1, \dots, u_{n-1})$  can be calculated in an iterative procedure by knowing that

$$\begin{aligned}
 p(s, \theta|u_1, \dots, u_{n-1}) = p(u_{n-1}|s, \theta)p(s, \theta|u_1, \dots, u_{n-2}) \times \\
 [\int p(u_{n-1}|s = 1, \theta)p(s = 1, \theta|u_1, \dots, u_{n-2})d\theta \quad (6) \\
 \int p(u_{n-1}|s = 0, \theta)p(s = 0, \theta|u_1, \dots, u_{n-2})d\theta]
 \end{aligned}$$

and

$$\begin{aligned}
 p(s, \theta|u_1) \\
 = \frac{p(u_1|s, \theta)p(s)p(\theta)}{\int p(u_1|s = 1, \theta)p(s = 1)p(\theta)d\theta + \int p(u_1|s = 0, \theta)p(s = 0)p(\theta)d\theta}
 \end{aligned} \tag{7}$$

To use the final log-odds ratio to identify cheating examinees, a positive ratio indicates there is more suspicion an examinee is using item preknowledge given his/her responses on the test items than there was before test administration, and a negative ratio

means the opposite. A ratio of 0 means the probability that an examinee is cheating is the same before and after test administration. Since there is not a sampling distribution for the log-odds ratio, a subjective decision needs to be made for the choice of the critical value.

McLeod et al. evaluated the effectiveness of this procedure by simulating an organized item-theft scenario in a 28-item CAT. They first simulated source examinees who are taking the test in order to memorize all items administered to them, and then compiled a list of compromised items by combining each source's memorized items. For examinees in the memorizing group, a correct response due to item preknowledge was introduced when one is administered an item from the compromised list. They investigated the difference between the empirical distribution of the log-odds ratio in the null and cheating group, and the marginal ROC curves, and their results suggested this statistic demonstrated a noticeable distributional differences between the null and the cheating group, especially when the cheating examinee had a low proficiency level and when there was a larger amount of compromised items. The ROC curves showed that this procedure could effectively detect item preknowledge when the probability of an item being memorized, i.e.,  $p(m_i)$ , was defined empirically through simulation or defined in relation to item difficulty.

As compared to most person-fit statistics, the Bayesian procedure defined a model for the cheating responses, and it demonstrated some promise for use as a test security control procedure. However, the choice of the detection criterion is relatively subjective, and the results implied that the null distribution of the log-odds ratio depended on how the cheating model was specified and varied among different  $\theta$  levels, so the practical

usefulness of this procedure might depend to some extent on the choice of the detection criterion and the specification of the cheating model.

### 2.2.2 Methods using information from secure items

The methods reviewed in this section distinguish between secure and insecure items according to their exposure rates, and use information from responses to secure items as a baseline for an examinee's performance. Segall (2002) and Shu (2010, 2013) used expanded IRT models to decompose an examinee's observed performance due to his/her real proficiency and the use of item-preknowledge. Belov (2013, 2014) compared the posterior distributions for  $\theta$  obtained from the two types of items via Kullback-Leibler divergence index (Kullback & Leiber, 1951). Li, Gu and Manna (2014) applied a regression-based method proposed by Haberman (2008) to identify group-level cheating due to item preknowledge. Wang, Li, and Gu (2014) calculated person-fit statistics using  $\hat{\theta}$  from the secure items to eliminate the effect of systematic error in  $\hat{\theta}$  due to the presence of cheating responses on the effectiveness of person-fit statistics.

*Expanded IRT models.* Segall (2002) specified a model for characterizing test compromise by incorporating a latent variable representing an examinee's item-preview propensity in the standard IRT model. The conditional probability of a correct response is defined as

$$P(u_{ij} = 1 | \theta_j, k_{ij}) = 1 + (1 - k_{ij})(p_{ij}^{(c)} - 1) \quad (8)$$

$$k_{ij} \sim \text{Bernoulli}(p_{ij}^{(\omega)}) \quad (9)$$

$$p_{ij}^{(\omega)} = \phi_i * \Phi[\alpha_i + \beta_i \omega_j] \quad (10)$$

$$\omega_j \sim N(0,1) \quad (11)$$

$\omega_j$  is the item-preview propensity parameter for examinee  $j$ ,  $\alpha_i$  and  $\beta_i$  are the item parameters on the item-preview dimension, and  $\Phi[\alpha_i + \beta_i\omega_j]$  is the normal ogive function to model the probability that examinee  $j$  has preknowledge on item  $i$ .  $\phi_i$  is the indicator for item type:  $\phi_i = 1$  means an item is an unsecure item, and thus  $p_{ij}^{(\omega)} = \Phi[\alpha_i + \beta_i\omega_j]$ , while  $\phi_i = 0$  means an item is a secure item, and thus  $p_{ij}^{(\omega)} = 0$ , indicating that there is no possibility that an examinee has preknowledge on this item.  $k_{ij}$  is the indicator for item-preknowledge status:  $k_{ij} = 1$  indicates examinee  $j$  has preknowledge on item  $i$ , and  $k_{ij} = 0$  indicates no preknowledge, and the probability of  $k_{ij} = 1$  is  $p_{ij}^{(\omega)}$ . Equation (8) indicates that when  $k_{ij} = 1$ , the probability of a correct response is 1, and when  $k_{ij} = 0$ , the probability of a correct response is  $p_{ij}^{(c)}$ , which is defined by the standard IRT model in equation (1).

To characterize test compromise, a variable representing item-level score gain,  $g_{ij}$ , is also defined.  $g_{ij} = 1$  only if an examinee cannot answer the item correctly based on real proficiency and guessing, and a correct answer is obtained through the use of item preknowledge, so

$$g_{ij} \sim \text{Bernoulli}(p_{ij}^{(g)}) \quad (12)$$

$$p_{ij}^{(g)} = k_{ij}(1 - c_i) \left(1 - \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}\right) \quad (13)$$

Markov Chain Monte Carlo (MCMC; e.g. Gelman et al., 2013) technique is used for model estimation. To quantify the extent of test compromise, posterior draws for model parameters are used to calculate four summary statistics. The first statistic is the

expected item preview frequency on each item across all examinees, i.e.,  $\sum_j p_{ij}^{(\omega)}$ . The second statistics is the expected score gain frequency on each item across all examinees, i.e.,  $\sum_j p_{ij}^{(g)}$ . The third statistic is the total number of previewed items for each examinee on the entire test, i.e.,  $\sum_j k_{ij}$ , and the fourth statistic is the total test score gain for each examinee, i.e.,  $\sum_j g_{ij}$ . Segall conducted a simulation on 60 items in a linear test where 40 items were possibly compromised and 20 items were secure. Zero, moderate to severe test compromise conditions were simulated by manipulating the proportion of compromised items (0%, 30%, 100%) in the unsecure item set, and compared the distribution of the four summary statistics to their true distributions. A small sample of 100 examinees were used for model calibration. The findings suggested the distribution of the estimated statistics are close to their true values in all compromise conditions, indicating the power of this method is sufficiently large to detect preknowledge at both item and examinee level, and the type-I error is also under control.

Different from methods based on person-fit measures, Segall's model not only evaluates the test compromise at person level, but also provides diagnostic measures for item-level compromise, and the effect of item preknowledge on test score gain can be directly estimated. This provides more information than person-fit measures.

Additionally, due to the distinction between the two types of items, an examinee's true proficiency can be estimated without being affected by the presence of cheating responses, which greatly improves the power of the method too.

Segall's model makes an assumption that a correct response is made with 100% certainty when an examinee has preknowledge on an exposed item. Shu (2010, 2013) argued that assuming an examinee memorized every exposed item correctly and

successfully retrieved the correct answer during the test was too unrealistic, and thus it may limit the flexibility of the model. Shu (2010, 2013) proposed a deterministic gated IRT model to characterize item preknowledge. Similar as Segall's model, Shu's model uses two latent variables- one representing an examinee's real proficiency (i.e.,  $\theta_{tj}$ ), and the other representing an examinee's cheating ability (i.e.,  $\theta_{cj}$ ). The conditional probability of a correct response is defined as

$$P(u_{ij} = 1 | \theta_{tj}, \theta_{cj}, T_j, \phi_i, b_i) = \begin{cases} P^t(u_{ij} = 1 | \theta_{tj}, b_i), & \text{when } \phi_i = 0, T_j = 0 \\ P^t(u_{ij} = 1 | \theta_{tj}, b_i), & \text{when } \phi_i = 0, T_j = 1 \\ P^c(u_{ij} = 1 | \theta_{tj}, b_i), & \text{when } \phi_i = 1, T_j = 0 \\ P^c(u_{ij} = 1 | \theta_{cj}, b_i), & \text{when } \phi_i = 1, T_j = 1 \end{cases} \quad (14)$$

and

$$P^t(u_{ij} = 1 | \theta_{tj}, b_i) = \frac{\exp(\theta_{tj} - b_i)}{1 + \exp(\theta_{tj} - b_i)} \quad (15)$$

$$P^c(u_{ij} = 1 | \theta_{cj}, b_i) = \frac{\exp(\theta_{cj} - b_i)}{1 + \exp(\theta_{cj} - b_i)} \quad (16)$$

$$T_j = 1, \text{ when } \theta_{tj} < \theta_{cj} \quad (17)$$

Equations (15) and (16) define the probability of a correct response due to one's true proficiency and cheating ability respectively. They both take the form of the standard IRT models.  $\phi_i$  denotes the type of an item.  $\phi_i = 1$  represents an unsecure item, and  $\phi_i = 0$  represents a secure item.  $T_j$  is the indicator for cheater.  $T_j = 1$  represents examinee  $j$  has item preknowledge, and  $T_j = 0$  indicates the examinee does not. As equation (14) suggests, the probability only depends on an examinee's cheating ability

when the examinee is a cheater and the item is unsecure. To solve the identification problem,  $\sum b_i = 0$  is specified as a constraint.

MCMC is used for model estimation. The posterior probability that an examinee has item preknowledge is used as the summary statistic, and it is compared to a cut-off (i.e.  $P_C$ ) defined by the user to classify an examinee as being a cheater or not. A simulation study was conducted to evaluate the effectiveness of this method in conditions with different proportions of compromised items in the entire test (0%, 30%, 50%, 70%), proportions of cheaters (5%, 35% and 70%) and different levels of score gains (high-, medium-, and low-effective). Responses to a 40-item linear test was simulated and a sample size of 2000 was used. The cut-off point of 0.9 was used, indicating only if the posterior mean of  $T_j$  is greater than 0.9, an examinee was classified as cheater. Specificity analyses (i.e. the classification accuracy among non-cheaters) suggest that the classification accuracy among non-cheaters is above 0.96 in all conditions, indicating that the false positive rate of this method is small. The sensitivity analyses (i.e. the classification accuracy among cheaters) suggest that the sensitivity of this method is quite high when there is only a smaller proportion of cheaters. The sensitivity decreased when the proportion of cheaters increased, as a result of the negative influence of cheating responses on item parameter estimation. This method had the higher sensitivity when the amount of each type of item is about the same (i.e. 50% secure items and 50% unsecure items) compared to the unbalanced proportion between the two types of items (i.e. 30% secure vs 70% unsecure, or 70% secure vs 30% unsecure), since the estimation for  $\theta_{tj}$  and  $\theta_{cj}$  was of equal precision when two types of items are of the same length.

Shu (2013) also applied this method to a real dataset, which consisted of 14 unsecure items and 21 secure items with more than 15,000 examinees. Findings from the real dataset supported the applicability and validity of the model. As for the applicability of the model in practice, the MCMC estimation for the model parameters converged well in 8,000 iterations, indicating that stable estimations can be achieved in practice. As for the validity of the model, the real proficiency,  $\theta_t$ , estimated from the deterministic gated model had high correlation for the estimates obtained from the standard IRT models solely based on the unexposed items. The model consistently flagged the same proportion of examinees with item preknowledge in different random samples with the same sample size. Shu (2013) also analyzed the characteristics of the examinees that are identified as using preknowledge, and found that this model tended to identify examinees of low/medium true proficiency with significant score gains (i.e. large values of  $\theta_c - \theta_t$ ).

Both the simulation study and the real data analysis suggested some success and promise of applying Shu's model in detecting item preknowledge. However, there are two major limitations with this approach. First of all, the deterministic gated model was only developed based on the 1PL-IRT model, which often fails to demonstrate a reasonable fit to educational data. The model could be extended to the 2PL- or 3PL- IRT model, but that increases the estimation complexity and its practical applicability remains unknown. Second, Shu assumed that the cheating model takes the same form as the standard IRT model, and the only difference in the cheating and non-cheating condition lies in the difference in  $\theta$ . This is equivalent to assuming lack of theta invariance between the two types of items. However, as discussed before, item preknowledge is not as simple as lack of theta invariance. By assuming the cheating model to be  $P^c(u_{ij} = 1 | \theta_{cj}, b_i) =$

$\frac{\exp(\theta_{cj}-b_i)}{1+\exp(\theta_{cj}-b_i)}$ , we are assuming that as an item becomes easier,  $P^c(u_{ij} = 1|\theta_{cj}, b_i)$

increases, but the increase in  $P^c(u_{ij} = 1|\theta_{cj}, b_i)$  probability is more relevant to the success of an examinee's item memorization or the exposure rate of an item, instead of the item difficulty. Therefore, the assumption for the cheating model may be questioned. This is the similar problem with the method proposed by Segall (2002) and McLeod et al. (2003), where a cheating model needs to be specified. Since both Shu (2010, 2013) and Segall (2002) simulated cheating with the model they proposed, their simulation studies are free of the problem for mis-specifying the cheating model. But the flexibility of their models is limited by the extent to which a cheating mechanism can be adequately represented by a particular cheating model.

***Regression-based method.*** Regression is commonly used in practice to identify outliers which have large score difference on two subtests (Haberman, 2008; Lewis, Lee, & von Davier, 2012). For the detection of preknowledge, a simple linear regression model is built to predict an examinee's score on the unsecure section using his/her score on the secure section, or the other way around, and examinees with large standardized residuals are flagged for further investigation.

Li, Gu and Manna (2014) conducted a simulation study to evaluate the effectiveness of this method in detecting preknowledge in a state assessment. They built the regression model to predict an examinee's score on the secure items using the score on the unsecure items. An outlier was identified if the residual was greater than two RMSE. Responses were simulated on 60 items in a linear test to mimic a typical state assessment. As the focus of their study was to evaluate the detection rate in different schools, schools with different sample sizes were simulated, and two variables were

manipulated – proportion of cheaters in each school (25%, 50%, 75%) and proportion of exposed items (5%, 17%, 33%). The regression model was built based on all the simulated schools, and the analysis was conducted for each school. The power in their study was very small- lower than 0.05 in all conditions, which implies that the practical usefulness of the method may be very limited.

***Person-fit Analysis with Purified  $\hat{\theta}$ .*** Wang et al. (2015) applied two person-fit statistics to compare an examinee’s performance on the two types of items. Different from standard person-fit analysis, instead of obtaining  $\hat{\theta}$  from the entire response vector,  $\hat{\theta}$  was estimated based on secure items only to prevent  $\hat{\theta}$  from being contaminated by the cheating responses on the possibly compromised items. The person-fit statistics were then calculated with the “purified”  $\hat{\theta}$ .

Two person-fit statistics were investigated: one is an adapted version of  $Z_c$  (denoted as  $Z_c^*$ ) and the other is  $l_z$ . For the calculation of  $Z_c$ , instead of comparing the residuals between easy and hard items, Wang et al. compared the residuals between secure and unsecure items, so  $Z_c^*$  takes the following form

$$Z_c^* = \frac{\overline{Exposed[P_i(\hat{\theta}) - u_i]} - \overline{Secure[P_i(\hat{\theta}) - u_i]}}{\sqrt{\left\{ \sum_{Exposed} \frac{\{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\}}{n_{Exposed}^2} \right\} + \left\{ \sum_{Secure} \frac{\{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\}}{n_{Secure}^2} \right\}}} \quad (18)$$

To increase the power of  $l_z$ ,  $l_z$  was calculated based on the exposed items only.  $l_z$  takes the following form

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}} \quad (19)$$

$$\text{where } l_0 = \sum_{g=1}^k \{U_g \ln P_g(\theta) + (1 - U_g) \ln [1 - P_g(\theta)]\}$$

$$E(l_0) = \sum_{g=1}^k \{P_g(\theta) \ln P_g(\theta) + (1 - P_g(\theta)) \ln [1 - P_g(\theta)]\},$$

$$\text{Var}(l_0) = \sum_{g=1}^k P_g(\theta) (1 - P_g(\theta)) \left[ \ln \frac{P_g(\theta)}{1 - P_g(\theta)} \right]^2$$

Wang et al. investigated the effectiveness of the two statistics in two scenarios. In the first scenario, item parameters were known, so the cheating responses did not have any impact on item parameters. In the second scenario, item parameters were unknown and item calibration was conducted with the presence of cheating responses, so the impact of bias in item parameter estimates introduced by the cheating responses on the effectiveness of both person-fit statistics was evaluated. The empirical distributions of both statistics were also investigated and compared to their asymptotic distribution (i.e.  $N(0,1)$ ).

Wang et al. simulated dichotomous responses to 60 items in a linear test for both scenarios. In the first scenario, proportion of exposed items was manipulated (25%, 50%, 75%). The empirical null distributions for both statistics showed a deviation from their theoretical distribution when  $\hat{\theta}$  was obtained from a small number of items (e.g. 15 items), but empirical null distribution of  $Z_c^*$  was much closer to its theoretical distribution than  $l_z$ . High detection power for both statistics were observed, and  $Z_c^*$  was also found to be more powerful to detect item preknowledge among high-proficiency examinees than  $l_z$ . In the second scenario, proportion of cheaters (10%, 30%) and proportion of exposed items (25%, 50%) were manipulated, and the results suggested large bias in item parameters could significantly reduce the power of both statistics, and greatly inflate the false positive rate, especially for  $l_z$ , which is similar as found by Shu (2013).

These findings suggest that the power of person-fit statistics could be largely improved by removing bias in  $\hat{\theta}$  caused by aberrant responses. However, the problem of discrepancy between the empirical and the theoretical null distribution still exists and may limit the effectiveness of person-fit analysis. Wang et al. conducted power analysis using the critical values from the empirical null distribution, which was simulated using the true  $\theta$  for each examinee. True  $\theta$  is never known in practice, and the empirical null distribution can only be simulated based on  $\hat{\theta}$ , which contains estimation error, especially when  $\hat{\theta}$  is obtained from a small number of items.

***Kullback-Leibler (KL) Divergence.*** Belov, Pashley, Lewis and Armstrong (2007) first proposed the idea of using the KL divergence to detect item preknowledge. Belov (2013, 2014) extended the work by considering situations where the set of exposed items is unknown, and by detecting test compromise at both the group and person level. KL divergence measures the dissimilarity between two distributions, defined as

$$D(P_1||P_2) = E \left( \ln \frac{P_1(x)}{P_2(x)} \right) \quad (20)$$

where  $P_1$  and  $P_2$  represent two distributions for random variable  $x$ . For discrete distributions,  $D(P_1||P_2) = \sum_{i=1}^n P_1(x_i) \ln \frac{P_1(x_i)}{P_2(x_i)}$ , and for continuous distributions,  $D(P_1||P_2) = \int P_1(x_i) \ln \frac{P_1(x_i)}{P_2(x_i)} dx$ . Large values of  $D(G||H)$  indicate large difference between  $G$  and  $H$ .

Belov et al. (2007) defined the person-level cheating index ( $h$ ) as the KL divergence between the posterior distributions of  $\theta$  obtained from the two types of items, i.e.,

$$h = D(P(\theta_j|\mathbf{u}_1)||P(\theta_j|\mathbf{u}_2)), \quad (21)$$

where  $\mathbf{u}_1$  and  $\mathbf{u}_2$  are response vectors on secure and unsecure items, respectively. To identify the aberrant groups, Belov (2013, 2014) defined the group-level index ( $g_c$ ) as the extended form of KL divergence between the distributions of  $h$  in one group and its distribution in every other group, i.e.,

$$g_c = \sum_{x \in C} (D(H_c||H_x) + D(H_x||H_c)) \quad (22)$$

where  $c$  is the index for a group, and  $C$  denotes the set of groups,  $H_c$  and  $H_x$  denote the distribution of  $h$  in group  $c$  and  $x$ . The sum of two KL divergence in equation (22) is to balance the asymmetry between  $D(H_c||H_x)$  and  $D(H_x||H_c)$ .

To detect examinees using item preknowledge, Belov (2013) proposed a two-stage detection method. In stage 1,  $g_c$  was computed for each group, and the empirical distribution of  $g_c$  was constructed using all the data. Given a significance level, groups with large values of  $g_c$  at the tail of the empirical distribution were identified as the aberrant group. In stage 2, the empirical distribution for  $h$  was constructed using data from the groups not identified in stage-1, and the critical value for  $h$  was found from the empirical distribution and used to identify aberrant examinees in the aberrant group identified in stage-1. Belov (2013) pointed out that the groups could be formed according to various relations between examinees. For instance, a group can consist of examinees taking the test in the same geographic location, or examinees from the same high school, same test-preparation center or same social networks. By conducting a two-stage detection, the aberrant groups were removed from building the empirical distribution of  $h$ , so that the empirical distribution is closer to that in the null condition. Furthermore, the

person-level detection is only limited to the individuals in the aberrant group, which can help reduce the type-I error rate. Simulations studies conducted by Belov (2013, 2014) demonstrated that this two-stage approach was more effective in reducing the type-I error rate than only conducting stage 2 without removing aberrant groups.

Belov (2013) considered three detection situations. In the first situation, all groups have preknowledge on the same set of exposed items, and the compromised item-set is known. An example of this situation is that a test agency can find some exposed items on the Internet. In the second situation, each group has preknowledge on a different subset of exposed items. The compromised subset unique to each group is unknown, but each subset belongs to a known collection of compromised subsets  $\Omega = (W_1, \dots, W_m)$ . An example is that various subsets of exposed items are found on different Internet sources. In the third situation, each group has preknowledge on a unique subset of exposed items, and each subset is unknown, which is the most realistic scenario.

In the first situation, the two-stage procedure was implemented directly. In the second and third situation, a “3D” algorithm was applied to first *detect* the aberrant group, and then *detect* the compromised item-set corresponding to each aberrant group, and lastly *detect* the aberrant persons given the aberrant group and compromised subset. The basic idea for finding the compromised subset unique to each group is to calculate  $g_c$  based on different subsets of items for a given group  $c$ , and the subset that gives the largest value for  $g_c$  is the compromised subset unique to group  $c$ . Therefore, for detection in the second situation, all the subsets in  $\Omega$  were compared to each other, while for detection in the third situation, since the number of all possible subsets is enormous,

Belov (2014) used a heuristic combinatorial optimization approach, called simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) to find the compromised subset.

Belov (2013, 2014) conducted a series of simulation studies to evaluate the effectiveness of the method in all three situations. A 50-item CAT was simulated using the item bank for the Law School Admission Test. Different proportion of aberrant groups (5%-20%) and different proportion of cheaters in an aberrant group (5%-20%) were manipulated in the simulation. In the first situation, item preknowledge was introduced to 4 and 12 items respectively, and the two-stage approach had low type-I error rate and high power even when the aberrant subset only consisted of 4 items. In the second and third situation, each aberrant group was assigned a unique compromised subset. Item preknowledge was introduced to around 40% and 20% of the test in second and third situation, respectively. In both situations, the 3D algorithm was found to greatly increase the power compared to the algorithm not identifying the compromised subset unique to each group.

The method based on KL divergence proved to be effective in detecting the difference in an examinee's performance between two subsets of items, and the use of combinatorial optimization is helpful for finding the compromised subset of items. However, the effectiveness of the method may be limited by one potential problem. An assumption underlying detection in different situations is that all examinees in one aberrant group have preknowledge on the same set of compromised items. Belov (2014) argued that "considering how small a group can be (e.g., class) or how specific corresponding relation can be (e.g. same group in a social network), this assumption is realistic". Nevertheless, this poses a requirement for finding the correct relation among

examinees, so that each group only has one unique subset of compromised items. If the assumption is not satisfied, the power for detecting aberrant groups in stage 1 would be reduced, which would then affect the effectiveness at person-level detection. Below (2014) proposes a method to take multiple relations among examinees into account, but multiple comparisons need to be conducted when multiple relations are considered and procedures used to control the familywise error rate would reduce the power to some extent. In addition, one purpose of using the two-stage detection algorithm is to make the empirical distribution of person-level index approximate to its empirical null distribution. Although the two-stage algorithm proved to be effective in this study, it only provides a less optimal solution to finding the null distribution for the statistic, and it relies on the effective detection of aberrant groups. As will be seen from the following section, the predictive checking method constructs the exact empirical null distribution in an easier way than the method based on KL divergence.

### **2.3 Summary of Existing Methods**

To summarize existing methods in the literature, first of all, methods based on person-fit analysis are mostly aimed at detecting the general misfit of an individual's response vector. With regular person-fit analysis, the cause of an aberrant response vector is not assumed, and thus a response vector flagged by a person-fit statistic could be due to other aberrant behaviors than item preknowledge. In applying person-fit statistics in detecting item preknowledge, one could obtain an individual's proficiency estimate from the secure items and calculate the person-fit statistic on the unsecure items, as in Wang et al. (2014), but the point estimation of an individual's proficiency does not take estimation

error into account, and Wang et al. (2015) showed the empirical null distribution of a statistic did not always approximate its asymptotic distribution, which is a common problematic feature with many person-fit statistics in short or medium-length tests.

Second, for approaches that are developed to detect item preknowledge in particular, methods proposed by McLeod et al. (2003), Segall (2002) and Shu (2010, 2013) all require the specification of a cheating model to characterize one's performance given item preknowledge. It is hard to know the goodness-of-fit of a particular cheating model in practice, and the misspecification of cheating model may affect the effectiveness of the method to some extent. The regression-based approach used by Li et al. (2014) had very little power to detect preknowledge on exposed items, so its practical usefulness may be limited. The KL divergence measure used by Belov (2011, 2013, 2014) demonstrated some promises to detect item preknowledge both at the group and at the person level, but the person-level detection relies on a group of individuals, and the sampling distribution used to flag person-level misfit is not strictly the null distribution of the person-level statistic. As will be seen in the next section, the predictive checking method can overcome some of the problems with the existing methods, and thus the present study has the potential for methodological contributions to the literature.

## **CHAPTER 3**

### **METHODOLOGY**

In Chapter 3, the mathematical definition and statistical properties of the predictive checking method are first introduced, followed by an explanation of the technical details for implementing the method. Specifically, four aspects related to the method are discussed in detail, including the estimation approach, sampling procedure, choice of test statistics and detection at different levels. For comparison with the predictive checking method, three other approaches, including the likelihood ratio test, adapted KL divergence and the regression-based method, are used and discussed here. The studies on the evaluation of the predictive checking method and on the comparison of the predictive checking method with the other methods are described in Chapter 4 to Chapter 6.

For notation simplicity, the secure subset is denoted as T1, and the unsecure subset is denoted as T2 henceforth. Let  $\mathbf{Y}=(Y_1, Y_2, \dots, Y_n)$  denote a random vector of responses on a set of items, and let  $\mathbf{y}=(y_1, y_2, \dots, y_n)$  be a realization of  $\mathbf{Y}$ .

### **3.1 Predictive Checking Method**

#### **3.1.1 Mathematical Definition and Properties**

Predictive checking is a general method to evaluate the goodness-of-fit of a given model based on predictive inference (Geisser, 1993). The basic idea under predictive checking is to predict the possible outcomes of future observations based on current observations, and when future observations become available, they are compared to the predictions to check the appropriateness of the model used for prediction. As the first step

in predictive checking, predictive inferences about unobserved data are made by constructing their distribution conditional on data that have been observed. Then the model-fit can be checked by comparing the observed data to the predictive distribution. To be specific, consider a test is divided into two subsets, and let  $\mathbf{y}_1, \mathbf{y}_2$  be an examinee's observed responses on subset I and II respectively. Let  $\boldsymbol{\omega}$  denote the unknown parameter(s) in the model, and  $p(\boldsymbol{\omega} | \mathbf{y}_1)$  be the posterior distribution of  $\boldsymbol{\omega}$  conditional on responses on subset I. Let  $\tilde{\mathbf{Y}}_2$  be the response data on subset II that would be observed (i.e., predictive data) if the responses on subset II come from the same model as  $\mathbf{y}_1$ , and let  $p(\tilde{\mathbf{Y}}_2 | \boldsymbol{\omega})$  be the likelihood distribution for predictive response vector  $\tilde{\mathbf{Y}}_2$  given parameter(s)  $\boldsymbol{\omega}$ . By “averaging” over all possible values of  $\boldsymbol{\omega}$ , the distribution of  $\tilde{\mathbf{Y}}_2$  conditional on  $\mathbf{y}_1$  is

$$p(\tilde{\mathbf{Y}}_2 | \mathbf{y}_1) = \int p(\tilde{\mathbf{Y}}_2 | \boldsymbol{\omega}) p(\boldsymbol{\omega} | \mathbf{y}_1) d\boldsymbol{\omega}. \quad (22)$$

Predictive checking evaluates the model fit by comparing the observed responses  $\mathbf{y}_2$  to the distribution of predictive data  $\tilde{\mathbf{Y}}_2$ . Typically, a test statistic  $T(\mathbf{y})$  is defined to measure the discrepancy between the observed and predictive data, so  $T(\mathbf{y}_2)$  is compared to the predictive distribution of  $T(\tilde{\mathbf{Y}}_2)$ , and the predictive  $p$ -value is used to summarize the comparison. The predictive  $p$ -value in a one-tailed test is

$$\Pr(T(\tilde{\mathbf{Y}}_2) \geq T(\mathbf{y}_2) | \mathbf{y}_1) = \int_{T(\tilde{\mathbf{Y}}_2) \geq T(\mathbf{y}_2)} p(\tilde{\mathbf{Y}}_2 | \mathbf{y}_1) d\tilde{\mathbf{Y}}_2 \quad (23)$$

for right-tailed test or

$$\Pr(T(\tilde{\mathbf{Y}}_2) \leq T(\mathbf{y}_2) | \mathbf{y}_1) = \int_{T(\tilde{\mathbf{Y}}_2) \leq T(\mathbf{y}_2)} p(\tilde{\mathbf{Y}}_2 | \mathbf{y}_1) d\tilde{\mathbf{Y}}_2 \quad (24)$$

for left-tailed test, and the  $p$ -value in a two-tailed test is  $2\min(\Pr(T(\tilde{\mathbf{Y}}_2) \geq T(\mathbf{y}_2) | \mathbf{y}_1), \Pr(T(\tilde{\mathbf{Y}}_2) \leq T(\mathbf{y}_2) | \mathbf{y}_1))$ . A  $p$ -value close to zero indicates the observed

response pattern is unlikely to be produced by the null model, and thus it indicates model misfit.

In the context of this study, assuming item parameters are available, which is often the case in a continuous testing program, the unknown parameter in an IRT model is simply the person proficiency parameter  $\theta$ . Assuming the true value of an examinee's proficiency parameter is  $\theta_0$ , as the number of items in T1 increases to infinity,  $p(\theta|\mathbf{y}_1)$  will converge to the point mass distribution at  $\theta_0$  by posterior consistency (van der Vaart, 1998), and thus  $p(T(\tilde{\mathbf{Y}}_2)|\mathbf{y}_1)$  will converge to its true distribution  $p(T(\tilde{\mathbf{Y}}_2)|\theta_0)$ . Due to such consistency, the predictive  $p$ -value is an asymptotically frequentist  $p$ -value which has a uniform distribution over  $[0, 1]$  under the null model.

The predictive checking procedure proposed here is similar to the *posterior predictive checks* (PPC) in a Bayesian framework, which is a common technique to check model fit in Bayesian analysis (e.g., Rubin, 1984; Gelman, Meng, & Stern, 1996; Gelman et al., 2013). PPC has been applied in assessing model fit or person fit in the IRT literature (e.g. Glas & Meijer, 2003; Sinharay, 2005; Sinharay, 2015; Sinharay & Johnson, 2003; Sinharay, Johnson, & Stern, 2006). However, in PPC, the construction of the posterior distribution of unknown parameters and the calculation of test statistics are based on the same dataset. The double use of the data causes the  $p$ -value in PPC to be conservative and thus it is less likely to reject the null hypothesis and to detect model misfit (e.g., Bayarri & Berger, 2000; Robins, van der Vaart, & Ventura, 2000). Different from PPC, the construction of the posterior distribution and the computation of the observed test statistic in the proposed method are performed on different sets of items, so the  $p$ -values in this method need not be conservative as in PPC.

This method provides several advantages over existing methods in detecting item preknowledge. Compared to methods that use person-fit statistics (e.g., Wang, Li, & Gu, 2015) or KL Divergence (Belov, 2013, 2014), the sampling distribution of the test statistic constructed in predictive checking takes into account the uncertainty in  $\theta$  estimation. In addition, the sampling distribution is not derived based on asymptotic theories, so it is expected to work well for small sample sizes. Compared to using expanded IRT models to characterize item preknowledge propensity (Segall, 2002; Shu, 2010), no model for the examinee's performance under preknowledge condition needs to be assumed with predictive checking, and it is much easier to implement and can be used either during or after test administration.

### 3.1.2 Implementation of Predictive Checking

The implementation of this method consists of three steps. The first step is to estimate the posterior distribution  $p(\theta|\mathbf{y}_1)$  from T1. The second step is to construct the predictive distribution  $p(\tilde{\mathbf{Y}}_2|\mathbf{y}_1)$  on T2. The analytic form for  $p(\tilde{\mathbf{Y}}_2|\mathbf{y}_1)$  is hard to derive in the IRT framework, so simulation is used to construct  $p(\tilde{\mathbf{Y}}_2|\mathbf{y}_1)$ . Specifically,  $K$  samples of  $\theta$ , denoted  $(\theta^1, \dots, \theta^K)$ , are first drawn from  $p(\theta|\mathbf{y}_1)$ . Then for each  $\theta^k$  ( $k = 1, \dots, K$ ), predictive response vectors on T2,  $\tilde{\mathbf{y}}_2^k$  ( $k = 1, \dots, K$ ) are generated in the null condition assuming there is no item preknowledge. This will result in  $K$  sets of predictive response data. In the final step, a test statistic is chosen and the test statistic is calculated for each predictive dataset and thus the predictive distribution for the test statistic,  $p(T(\tilde{\mathbf{Y}}_2)|\mathbf{y}_1)$ , is constructed. The test statistic computed from observed responses on T2

can then be compared to its predictive distribution. The following four sections discuss the details for implementing each step in this study.

### **3.1.2.1 Estimation of $p(\theta|\mathbf{y}_1)$ from T1**

The  $\theta$  distribution can be constructed in three ways, including the normal approximation of the maximum likelihood estimator (MLE) of  $\theta$ , the posterior distribution in Bayesian framework, and the fiducial distribution from generalized fiducial inference (Hannig, 2009; 2013). Normal approximation of the MLE of  $\theta$  is not considered as the normal approximation may not work well when the number of secure items is small, and in practice, it is often hard to have a large number of secure items. Bayesian posterior distribution and fiducial distribution are considered here. Bayesian posterior distribution is used due to both its popular use in IRT estimation and its ease of implementation. Fiducial distribution can be interpreted as a posterior calculated from a data-dependent non-informative prior, and it is considered here because its application in IRT parameter estimation suggested it can lead to better item parameter recovery than Bayesian approach with a non-informative prior when sample size is small (Liu, 2015).

#### **3.1.2.1.1 Bayesian Posterior Distribution**

According to Bayes' rule,  $p(\theta|\mathbf{y}_1) \propto p(\theta)p(\mathbf{y}_1|\theta)$ , where  $p(\theta)$  is the prior density and  $p(\mathbf{y}_1|\theta)$  is the likelihood for response pattern on T1. As for the choice of prior distribution, standard normal distribution is often used as the prior distribution in IRT scoring estimation, but preliminary analyses suggested using  $N(0,1)$  as prior would lead to large inflation of type-I error at extreme theta levels when the number of secure items is small. Therefore, less-informative prior distributions are employed so that less

shrinkage is introduced to the resulting posterior distribution. Two less informative priors are explored: one is a normal distribution with mean of 0 and standard deviation of 2 - i.e.,  $N(0,2^2)$  and the other is Jeffreys' prior. Jeffreys' prior is considered here since it has been shown to result in good coverage-efficiency balance for the binomial proportion (e.g., Brown, Cai, & DasGupta, 2001). If item parameters are the same for all binary items, the problem for  $\theta$  estimation is then isomorphic to the problem of binomial proportion estimation. This is because when all items have the same parameters, the item responses are independent and identically distributed (i.i.d.) Bernoulli trials, and the probability of a correct response can be estimated as a binomial proportion. Jeffreys' prior is proportional to the square root of Fisher information for  $\theta$ :

$$p(\theta) = I(\theta)^{1/2} = \left[ \sum_{i=1}^I \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \right]^{1/2} \quad (25)$$

where  $i$  is the index for item, and  $I$  is the total number of items,  $P_i(\theta)$  is the probability of a correct response on item  $i$  and  $Q_i(\theta) = 1 - P_i(\theta)$ .  $P_i'(\theta)$  is the first derivative of  $P_i(\theta)$  with respect to  $\theta$ .

### 3.1.2.1.2 Fiducial Distribution

The idea of fiducial distribution was first introduced by Fisher (1930) as an attempt to make probability statements about unknown parameters without assuming a prior distribution. Basically, fiducial inference is based on the idea of switching the role between the parameters and the data. The logic of fiducial inference can be illustrated by a normal-location example. Suppose  $X_1, \dots, X_n$  are i.i.d. random variables from  $N(\mu, \sigma^2)$ , with known  $\sigma^2$  but unknown  $\mu$ . To make an inference about  $\mu$ , as  $\bar{X} \sim N(\mu, \sigma^2/n)$ , where

$\bar{X} = \sum_{i=1}^n X_i / n$ ,  $\bar{X}$  can be expressed as  $\bar{X} = \mu + U \cdot \sigma / \sqrt{n}$ , where  $U$  is a random variable from  $N(0,1)$ . This is equivalent to  $\mu = \bar{X} - U \cdot \sigma / \sqrt{n}$ . After observing  $\bar{X} = \bar{x}$ , the fiducial distribution for  $\mu$  is  $N(\bar{x}, \sigma^2/n)$ .

In generalized fiducial inference (Hannig, 2009; 2013), the definition of a fiducial distribution starts with defining the *data generating equation*, which is an expression representing the association among data ( $\mathbf{X}$ ), parameters in the model ( $\boldsymbol{\omega}$ ) and randomness ( $\mathbf{U}$ ) whose distribution does not depend on  $\boldsymbol{\omega}$ , i.e.,  $\mathbf{X} = G(\boldsymbol{\omega}, \mathbf{U})$ . For instance, in the normal location example above, the data generation equation is  $\bar{X} = \mu + U \cdot \sigma / \sqrt{n}$ . Then the solution set for  $\boldsymbol{\omega}$  is found from the data generating equation, denoted as  $Q(\mathbf{X}, \mathbf{U}) = \{\boldsymbol{\omega}: \mathbf{X} = G(\boldsymbol{\omega}, \mathbf{U})\}$ . In the normal location example, the solution set for  $\mu$  is  $\mu = \bar{X} - U \cdot \sigma / \sqrt{n}$ . The solution set for  $\mu$  is a singleton set, but sometimes the solution set may contain no solution or more than one solution. The empty solution case is avoided by conditioning the solution set on  $Q(\mathbf{X}, \mathbf{U}) \neq \emptyset$ . When there are multiple solutions, one needs to select one according to some possibly random rules, denoted  $V(Q(\mathbf{x}, \mathbf{U}^*))$ . After observing  $\mathbf{X} = \mathbf{x}$ , the generalized fiducial quantity is defined as

$$V(Q(\mathbf{x}, \mathbf{U}^*)) | \{Q(\mathbf{x}, \mathbf{U}^*) \neq \emptyset\} \quad (26)$$

where  $\mathbf{U}^*$  is an independent copy of  $\mathbf{U}$ . More details about generalized fiducial inference can be found in Liu and Hannig (2016), Hannig (2009; 2013), and Hannig, Iyer, Lai, and Lee (2016).

In applications to IRT models, take the 2PL model as an example, the 2PL model takes the form of

$$P(Y_i = 1 | a_i, b_i, \theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))} \quad (27)$$

The data generating equation for a person's response to an item  $i$ ,  $Y_i$ , is

$$Y_i = \begin{cases} 1 & \text{if } U_i \leq P(Y_i = 1 | a_i, b_i, \theta) \\ 0 & \text{if } U_i > P(Y_i = 1 | a_i, b_i, \theta) \end{cases} \quad (28)$$

where  $U_i$  represents the randomness and  $U_i \sim \text{Uniform}(0,1)$ . Equation 28 is equivalent to

$$Y_i = \begin{cases} 1 & \text{if } A_i \leq a_i\theta - a_i b_i \\ 0 & \text{if } A_i > a_i\theta - a_i b_i \end{cases} \quad (29)$$

where  $A_i = \log \frac{U_i}{1-U_i}$ , and  $A_i \sim \text{Logistic}(0,1)$ . Assuming item parameters ( $a_i$  and  $b_i$ ) are

known, when  $a_i > 0$ , the solution set for  $\theta$  from one single response is

$$\theta \in \begin{cases} \left[ \frac{A_i + a_i b_i}{a_i}, +\infty \right) & \text{if } Y_i = 1 \\ \left( -\infty, \frac{A_i + a_i b_i}{a_i} \right) & \text{if } Y_i = 0 \end{cases} \quad (30)$$

Given a vector of responses on  $n$  items ( $Y_1, Y_2, \dots, Y_n$ ), let  $I_0$  be the index sets for incorrect

responses, i.e.,  $I_0 = \{i: Y_i = 0, i = 1, 2, \dots, n\}$ , and  $I_1$  be the index sets for correct

responses, i.e.,  $I_1 = \{i: Y_i = 1, i = 1, 2, \dots, n\}$ . Let  $s = \sum_{i=1}^n Y_i$  be the observed total score,

and let  $m_0 = \min_{i \in I_0} \frac{A_i + a_i b_i}{a_i}$  and  $m_1 = \max_{i \in I_1} \frac{A_i + a_i b_i}{a_i}$ . The solution set for  $\theta$  based on

( $Y_1, Y_2, \dots, Y_n$ ) is

$$\theta \in \begin{cases} [m_1, +\infty), & \text{if } s = n \\ (-\infty, m_0), & \text{if } s = 0 \\ (m_1, m_0), & \text{if } 1 \leq s \leq n - 1 \\ \emptyset, & \text{otherwise} \end{cases} \quad (31)$$

If the solution set is non-empty, it is an interval instead of a single value. So the

following selection rule is applied: if  $s = n$ ,  $\theta = m_1$ ; if  $s = 0$ ,  $\theta = m_0$ ; if  $1 \leq s \leq n -$

$1$ ,  $\theta = m_0$  with probability of 0.5 and  $\theta = m_1$  with probability of 0.5.

Note that equation 31 combined with the selection rule gives a single point of  $\theta$  corresponding to a fixed vector of ( $A_1, A_2, \dots, A_n$ ). The fiducial distribution of  $\theta$  is

derived based on the joint distribution of  $(A_1, A_2, \dots, A_n)$ . In particular, if  $s = n$  or  $0$ ,  $A_i^* \sim \text{Logistic}(0,1)$  and  $A_i^*$ 's are mutually independent, so  $f(A_1^*, A_2^*, \dots, A_n^*) = \prod_{i=1}^n f(A_i^*)$ . If  $1 \leq s \leq n - 1$ , in order for the solution to be non-empty,  $A_i^*$ 's should subject to  $m_1 < m_0$ , and each  $A_i^* \sim \text{Logistic}(0,1)$ , which means  $A_i^*$ 's should be chosen such that the value of  $\frac{A_i^* + a_i b_i}{a_i}$  corresponding to any correct response is smaller than that corresponding to an incorrect response. The details of sampling for  $(A_1^*, A_2^*, \dots, A_n^*)$  from their joint distribution is discussed in the next section.

### 3.1.2.2 Sampling From $p(\theta|\mathbf{y}_1)$

With Bayesian approach, to sample from the  $\theta$  distribution, a discrete approximation to the posterior distribution is used by evaluating the posterior probability over a grid of  $\theta$  values from -5 to 5 with an equal increment of 0.001. The probability at each  $\theta_g$  ( $g = 1, \dots, 10001$ ) is computed as  $p(\theta_g|\mathbf{y}_1) / \sum_{g=1}^{10001} p(\theta_g|\mathbf{y}_1)$ , and then the *sample* command in R (Version 3.1.1; R Core Team, 2014) is used to get a sample of  $\theta$  values according to their probability.

With fiducial approach, 1000 samples of vector  $\mathbf{A}^* = (A_1^*, A_2^*, \dots, A_n^*)$  are first drawn from their joint distribution, denoted as  $\mathbf{A}^{*(k)} = (A_1^*, A_2^*, \dots, A_n^*)^{(k)}$  ( $k=1, 2, \dots, 1000$ ). Then for a given sample  $\mathbf{A}^{*(k)}$ ,  $\theta^{(k)}$  is obtained based on equation (31) and the selection rule. To draw  $(A_1^*, A_2^*, \dots, A_n^*)^{(k)}$  from their joint distribution, if  $s = n$  or  $0$ ,  $A_i^*$  is simulated from  $\text{Logistic}(0,1)$  for all  $i$  ( $i=1, 2, \dots, n$ ), and  $\theta^{(k)}$  simply takes  $m_1$  or  $m_0$ . If  $1 \leq s \leq n - 1$ , Gibbs sampling (e.g. Gelman et al., 2013) technique is implemented. Gibbs sampler draws each  $A_i^*$  from its conditional distribution on all other parameters, so it decomposes the problem of drawing from a  $n$ -dimensional multivariate distribution into

drawing from a series of one-dimensional distributions. The algorithm starts with arbitrarily selected starting values of  $(A_1^{*(0)}, A_2^{*(0)}, \dots, A_n^{*(0)})$  which satisfies  $m_1 < m_0$ . The algorithm then proceeds to update each component in  $\mathbf{A}^*$  in turn in one sample.

Specifically, at the  $t$ th sample ( $t=1, 2, \dots, 2000$ ),

$$A_1^{*(t)} \text{ is drawn from } p(A_1^* | A_2^{*(t-1)}, A_3^{*(t-1)}, \dots, A_n^{*(t-1)})$$

$$A_2^{*(t)} \text{ is drawn from } p(A_2^* | A_1^{*(t)}, A_3^{*(t-1)}, \dots, A_n^{*(t-1)})$$

⋮

$$A_n^{*(t)} \text{ is drawn from } p(A_n^* | A_1^{*(t)}, A_2^{*(t)}, A_3^{*(t)}, \dots, A_{n-1}^{*(t)}).$$

Let  $\mathbf{A}^*_{(-i)}$  denote the vector  $\mathbf{A}^*$  excluding component  $A_i$ . If the  $i$ th response is correct,  $p(A_i^* | \mathbf{A}^*_{(-i)})$  is the density of Logistic(0,1) truncated from above at  $a_i m_0 - a_i b_i$ , and if the  $i$ th response is incorrect,  $p(A_i^* | \mathbf{A}^*_{(-i)})$  is the density of Logistic(0,1) truncated from below at  $a_i m_1 - a_i b_i$ .

The convergence of the algorithm was assessed visually via trace plots, and a preliminary analysis suggested convergence was reached quickly. The fiducial distribution was constructed using the last 1000 samples.

### 3.1.2.3 Test Statistics

Three test statistics are considered to use in predictive checking. The first statistic is the summed score on T2, the second statistic is the point estimate of  $\theta$  (denoted as  $\hat{\theta}$ ) from T2, and the third statistic is the variance of the posterior distribution of  $\theta$  from T2 (i.e.,  $\text{var}(\theta | \mathbf{y}_2)$ ). The first two statistics are supposed to be responsive to item preknowledge, as the direct statistical effect caused by item preknowledge is test score

increase. The summed score is easy to calculate, but it only provides partial information from the response pattern. For example, given a total of  $n$  binary items, there are  $2^n$  possible response patterns, but there are only  $(n+1)$  possible summed scores (i.e. scoring from 0 to  $n$ ). Also, when the number of items is small, the predictive distribution of the summed score only consists of a few categories, and the discreteness of the predictive distribution may limit the detection power. In contrast,  $\hat{\theta}$  provides more information than the summed score. The total number of possible  $\hat{\theta}$  can be as many as the total number of possible response patterns in a two-parameter or three-parameter IRT model. In this study,  $\hat{\theta}$  is computed via expected a posteriori (EAP), which is a commonly-used estimator for  $\theta$  in IRT. The choice of the third statistic is similar to the two person-fit statistics proposed by Drasgow et al. (1987) to evaluate the flatness of the likelihood function. A large variance of the  $\theta$  posterior may suggest that the likelihood function is flat, which implies that the responses provide less information for  $\theta$  estimation. When the summed score or the EAP is used, a one-tailed test is conducted and an unusually large score is identified as being aberrant, since score increase is of primary concern in the case of item preknowledge. When variance of  $p(\theta|\mathbf{y}_2)$  is used, a two-tailed test is conducted, as the variance could be either too large or too small in the case of item preknowledge.

#### **3.1.2.4 Item-set Level and Item Level Detection**

Discussions above have been focused on applying predictive checking to the entire T2. In practice, the choice of unsecure items could be difficult. In some situations, one may know exactly which items have high risks of being compromised, such as the items found being posted on the Internet. In that situation, predictive checking can be

applied on those items altogether. However, a more realistic scenario is that different examinees could have preknowledge on different subsets of items, and which items and how many items are compromised are unknown to us. In addition, the power of predictive checking could also be affected if a set of items only consists of a small proportion of truly compromised items. Therefore, a method for item-level detection is proposed here.

The idea for item-level detection is to evaluate the change of  $\hat{\theta}$  from including a particular item to excluding that item. If a correct response on an item is due to item preknowledge,  $\hat{\theta}$  is expected to show a decrease by excluding the item from estimation. In contrast, if an item response (regardless of correct or incorrect response) fits the model, the change in  $\hat{\theta}$  by excluding the item from estimation is expected to be caused by estimation error or by the increased proportion of aberrant responses in the remaining response vector. Specifically, to conduct predictive checking for item  $i$ , one still obtains  $p(\theta|\mathbf{y}_1)$  from secure items, draws  $N$  samples from  $p(\theta|\mathbf{y}_1)$ , and simulates  $N$  sets of predictive response data  $\tilde{\mathbf{y}}_2$  on a set of items including item  $i$ . Then  $\hat{\theta}$  is computed on each predictive dataset first with item  $i$  included (denoted as  $\hat{\theta}(\tilde{\mathbf{y}}_2)$ ), and then with item  $i$  excluded (denoted as  $\hat{\theta}(\tilde{\mathbf{y}}_2^{-i})$ ). Correspondingly,  $\hat{\theta}$  is computed on the observed responses on the same set of items first with item  $i$  included (denoted as  $\hat{\theta}(\mathbf{y}_2)$ ), and then with item  $i$  excluded (denoted as  $\hat{\theta}(\mathbf{y}_2^{-i})$ ). The difference between the two  $\hat{\theta}$ 's is used as the test statistic. The observed difference,  $\Delta\hat{\theta}_i = \hat{\theta}(\mathbf{y}_2^{-i}) - \hat{\theta}(\mathbf{y}_2)$ , is then compared to the predictive distribution of  $\tilde{\Delta}\hat{\theta}_i = \hat{\theta}(\tilde{\mathbf{y}}_2^{-i}) - \hat{\theta}(\tilde{\mathbf{y}}_2)$ . Although it is expected to see a negative change in  $\hat{\theta}$  after deleting a compromised item and thus a left-tailed test should be used, preliminary analyses suggest that using a one-tailed test could result in an

excessive inflation of type-I error rates at medium  $\theta$  levels. Therefore, a two-tailed test was used to flag unusually large  $|\Delta\hat{\theta}_i|$  in this study.

Sections 3.2 to 3.4 below describe the three methods used to compare to the predictive checking method. These three methods are the likelihood ratio test, the adapted KL divergence and the regression-based method. The reason to choose each method for comparison is also justified in each of the sections below.

### **3.2 Likelihood Ratio Test**

Likelihood ratio test is considered here as it is commonly used to compare the goodness of fit between two nested models in statistics and it has an asymptotic sampling distribution. For the purpose of this study, the two models compared here are a) the person's ability remains constant throughout T1 and T2, and b) the person's ability on T1 is lower than that on T2. This is equivalent to testing the null hypothesis

$$H_0: \theta_1 = \theta_2$$

against the alternative hypothesis

$$H_1: \theta_1 < \theta_2$$

where  $\theta_1$  and  $\theta_2$  represent a person's ability on T1 and T2 respectively. The idea of the likelihood ratio test used here is similar to the invariance test used by Klauer (1991), as discussed in Chapter 2. It is also similar to the likelihood ratio statistic proposed by Levine and Drasgow (1988), but the test used here does not specify a particular misfitting model, so the optimal detection property does not hold here. However, as discussed in Chapter 2, unless we can find the right model to characterize an examinee's performance

under the preknowledge situation, the optimal detection will not be achieved by any means.

The likelihood ratio test is very easy to implement. It compares the maximum likelihood of a response pattern under  $H_0$  to that under  $H_1$ . Specifically, the maximum likelihood of the response pattern on two subtests under  $H_0$  is

$$L(\hat{\theta}) = \prod_{i=1}^n P_i(\hat{\theta})^{Y_i} (1 - P_i(\hat{\theta}))^{1-Y_i} \quad (32)$$

where  $n$  is the total number of items on two subtests,  $Y_i$  is the item response on item  $i$ , and  $\hat{\theta}$  is the maximum likelihood estimator (MLE) based on the responses throughout T1 and T2. The maximum likelihood under  $H_1$  is

$$L(\hat{\theta}_1, \hat{\theta}_2) = \prod_{i=1}^{n_1} P_i(\hat{\theta}_1)^{Y_i} (1 - P_i(\hat{\theta}_1))^{1-Y_i} * \prod_{i=1}^{n_2} P_i(\hat{\theta}_2)^{Y_i} (1 - P_i(\hat{\theta}_2))^{1-Y_i} \quad (33)$$

where  $n_1$  and  $n_2$  are the number of items on T1 and T2,  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are the MLE obtained from the responses on T1 and T2 respectively.  $\frac{L(\hat{\theta}_1, \hat{\theta}_2)}{L(\hat{\theta})}$  is the likelihood ratio, and small values of the likelihood ratio gives evidence to  $H_0$ . By taking the logarithm of the likelihood ratio, i.e.,  $\Delta = \log L(\hat{\theta}_1, \hat{\theta}_2) - \log L(\hat{\theta})$ ,  $2\Delta$  is used as the test statistic in a two-sided test (i.e.  $H_1: \theta_1 \neq \theta_2$ ), and it follows asymptotically a chi-square distribution with  $df = 1$  when  $H_0$  is true (Lehmann, 1999, p. 526-527). With  $H_1$  being directional in our study, the following statistic is used in a one-sided test:

$$R = \begin{cases} \sqrt{2\Delta}, & \text{if } \hat{\theta}_1 < \hat{\theta}_2 \\ -\sqrt{2\Delta}, & \text{if } \hat{\theta}_1 \geq \hat{\theta}_2 \end{cases} \quad (34)$$

Given  $H_0$  is true, the likelihood ratio test is asymptotically equivalent to a Wald test (Casella & Berger, 2002, p. 493) and by the Wald test,  $\hat{\theta}_1 - \hat{\theta}_2$  follows asymptotically a

normal distribution with mean of 0. Therefore,  $P(\hat{\theta}_1 < \hat{\theta}_2) = P(\hat{\theta}_1 > \hat{\theta}_2)$  under  $H_0$ , and as a result,  $R$  follows asymptotically a standard normal distribution. Large values of  $R$  leads to the rejection of  $H_0$ .

### 3.3 Adapted KL Divergence

The KL divergence is considered here as it is similar to the predictive checking method and existing research (e.g. Belov, 2013; 2014) has shown promising results regarding its performance. The KL divergence aims at comparing the posterior distributions of  $\theta$  between T1 and T2, while the predictive checking in this study aims at detecting the shift of  $\hat{\theta}$ . A key difference between the two methods is how the sampling distribution of the test statistic is constructed. Predictive checking constructs the sampling distribution using the predictive inference, while Belov used information from the examinee group and attempted to construct the sampling distribution through a purification process. Using group information to construct the sampling distribution has a potential problematic feature in that even after a purification process, the group could still consist of some compromised responses, and thus the sampling distribution is not strictly the null distribution of the statistic. Using group information to construct the sampling distribution could have another problem with an adaptive test design, as examinees within a group may take different test items, and thus the sampling distribution of the KL divergence may depend on the item parameters used to estimate the  $\theta$  posterior distributions. Therefore, as one adaptation from Belov's approach, this study employs a simulation approach to construct the sampling distribution instead of using group information.

The KL divergence used by Belov takes the form of

$$h = D(P(\theta|\mathbf{y}_1)||P(\theta|\mathbf{y}_2)) = \sum_{q=1}^Q P(\theta_q|\mathbf{y}_1) \ln \frac{P(\theta_q|\mathbf{y}_1)}{P(\theta_q|\mathbf{y}_2)} \quad (35)$$

where  $q$  is the index for quadrature points, and  $Q$  is the total number of quadrature points,  $P(\theta_q|\mathbf{y}_1)$  and  $P(\theta_q|\mathbf{y}_2)$  are the posterior probability at  $\theta_q$  from T1 and T2 respectively.

The value of  $h$  measures how the two posterior distributions distinguish from each other, but does not reflect the direction to which the posterior distribution shifts. For the purpose of detecting test compromise, we are more concerned with the change in the direction of  $\theta_1 < \theta_2$ , so as a second adaptation, a signed KL divergence statistic is developed to use in a one-sided test. The signed statistic takes the form of

$$s = \begin{cases} h, & \text{if } \hat{\theta}_1 < \hat{\theta}_2 \\ -h, & \text{if } \hat{\theta}_1 \geq \hat{\theta}_2 \end{cases} \quad (36)$$

Large values of  $s$  supports the alternative hypothesis of  $\theta_1 < \theta_2$ .

To implement the adapted KL divergence approach, the KL statistic is first calculated for each person according to equation 35 and 36. In this study, 41 quadrature ( $Q=41$ ) points from -4 to 4 with an equal increment of 0.20 are used.  $P(\theta_q|\mathbf{y}_1)$  and  $P(\theta_q|\mathbf{y}_2)$  are both obtained with the Bayesian approach using the Jeffreys prior. Then the sampling distribution of  $s$  in the null condition is constructed based on the definition of the test size in a composite null hypothesis:  $\sup_{\theta \in \theta_0} P(\text{reject } H_0 | \theta)$  (Casella & Berger, 2002, p. 385). It is obtained by using the most conservative sampling distribution at a number of  $\theta$  values that cover the typical range of possible values of  $\theta$  in practice. The use of the most conservative sampling distribution ensures the largest type-I error rate for  $\theta$  in the null space does not exceed the nominal level.

Specifically, in this study, the construction of the sampling distribution of  $s$  involves the following two steps:

1. Choose a total of 61  $\theta$  values from -3 to 3 with an equal increment of 0.1. For each  $\theta_k^*$  ( $k = 1, \dots, 61$ ), simulate responses on T1 and T2 in the null condition given  $\theta_1 = \theta_2 = \theta_k^*$ , where  $\theta_1$  and  $\theta_2$  represent a person's ability on T1 and T2, and then calculate  $s$  based on the simulated responses. Repeat the simulation process for  $N$  times ( $N=1000$  in this study) at each  $\theta_k^*$ , and this results in the sampling distribution of  $s$  at  $\theta_k^*$ , denoted as  $p(s|\theta_1 = \theta_2 = \theta_k^*)$ .
2. Identify the cut-off value,  $c_k$ , corresponding to the nominal right-tailed  $\alpha=0.05$  for each  $p(s|\theta_1 = \theta_2 = \theta_k^*)$ . The sampling distribution with the largest cut-off value is used as final sampling distribution of  $s$  to identify unusually large outliers, i.e.  $p(s|\theta_1 = \theta_2) = p(s|\theta_1 = \theta_2 = \theta_m^*)$ , where  $m = \arg \max_k c_k$ .

### 3.4 Regression-based approach

The regression-based approach is chosen here due to its common use in practice to model the inconsistency between test section scores (Haberman, 2008), and also because it does not need to be based on an IRT model. To implement this approach, a linear regression model is built based on a group of examinees to predict the summed score on T2 using the summed score on T1, i.e.,

$$s_{j2} = \alpha + \beta s_{j1} + \varepsilon_j \tag{37}$$

$$\varepsilon_j \sim N(0, \sigma^2) \text{ and } \sigma(\varepsilon_j, \varepsilon_i) = 0$$

where  $s_{j1}$  and  $s_{j2}$  are summed scores on T1 and T2 respectively for person  $j$ . Summed score is used because it is easy to calculate and more importantly, there is no need to

assume a parametric model to get summed score. Summed score on T2 is predicted by that on T1 instead of in the opposite direction since analytical result shows if there is a score increase due to preknowledge, the standardized residual tends to be larger this way when the assumptions for standard linear regression model hold (i.e. homoscedasticity and normal residual) and thus it is more powerful to detect preknowledge. The standardized residual for an observation  $j$  is calculated as

$$Zr_j = \frac{s_{j2} - \hat{\alpha} + \hat{\beta}s_{j1}}{\hat{\sigma}} \quad (38)$$

where  $\hat{\sigma}^2$  is the unbiased estimate of error variance (i.e.,  $\sigma^2$ ) and  $\hat{\sigma} =$

$$\sqrt{\frac{1}{N-2} \sum_{j=1}^N (s_{j2} - \hat{\alpha} - \hat{\beta}s_{j1})^2}. \text{ If } Zr_j > 1.65, \text{ person } j \text{ is flagged as an outlier}^1.$$

Regression-based method is different from the other three methods in that the regression model needs to be constructed based on a group of examinees, while the analysis in the other methods is simply based on the response vector of one examinee. When evaluating the regression-based method, responses in the null condition are generated for 5000 examinees with  $\theta \sim N(0,1)$ . In the cheating condition, compromised responses are simulated on 5% examinees. A relatively small number of cheaters is chosen here because cheating is still a low-probability event in a high-stakes operational testing program. Furthermore, considering the estimation of the regression line will be

---

<sup>1</sup> In regression analysis, residuals are typically standardized using studentized residual:

$t_i = \frac{s_{j2} - \hat{\alpha} + \hat{\beta}s_{j1}}{\hat{\sigma}\sqrt{(1-h_{ii})}}$ , where  $h_{ii}$  is the  $i$ th diagonal element in the hat matrix. In practice, standardized residual is calculated with  $\hat{\sigma}$  only in the denominator, since as the sample size is large, the two types of standardized residuals give very similar values, and using  $\hat{\sigma}$  simplifies the calculation. The critical value of 1.65 is the 95<sup>th</sup> percentile in a  $t$ -distribution with degree of freedom  $(n-2)$ , which is the sampling distribution of the studentized residual.

affected to different extent if cheaters follow different proficiency distributions, two distributions are assumed for the proficiency of cheaters:  $\theta \sim N(0,1)$  and  $N(-1,1)$ <sup>2</sup>. These two distributions are chosen based on the assumption that examinees with low to moderate proficiency levels have higher motivation to cheat.

---

<sup>2</sup> The cheating group with  $\theta \sim N(0,1)$  was chosen by randomly sampling 250 examinees from the 5000 examinees with  $\theta \sim N(0,1)$ . For the cheating group with  $\theta \sim N(-1,1)$ , a group of 250 was simulated from  $\theta \sim N(-1,1)$  first, and then another group of 4750 was simulated from  $\theta \sim N(250/4750,1)$ , and these two groups were combined to mimic a group of 5000 examinees with  $\theta \sim N(0,1)$ . The mixture distribution does not follow exactly a standard normal distribution. The mean of the mixture distribution is 0, and the standard deviation is slightly larger than 1, but the difference is small enough to have any practical consequences for the purpose of this study.

## CHAPTER 4

### STUDY 1: EVALUATION OF PREDICTIVE CHECKING

#### 4.1 Study Design

To evaluate the statistical properties of the predictive checking method, a simulation study (labeled as *Study 1*) was conducted to evaluate the influence of different factors on the effectiveness of the predictive checking method in detecting item preknowledge at both the item-set level and the item level. The effectiveness of different test statistics was also compared. Four factors were systematically manipulated: the number of items in T1, the number of items in T2, the proportion of truly compromised items in T2, and the estimation method to obtain  $p(\theta|T1)$ .

*Number of items in T1.* The first factor influences the dispersion of  $p(\theta|y_1)$  and then further to affect the predictive distribution  $p(T(\tilde{Y}_2)|y_1)$ . As the number of items in T1 increases,  $p(\theta|y_1)$  will be more concentrated around the true value of  $\theta$ , and accordingly  $p(T(\tilde{Y}_2)|y_1)$  will be closer to its true distribution. Theoretically, the longer T1 is, the better this method will work. However, one may not have many items in this set in practice, due to the cost of producing new items to be used as secure items or the limited number of pretest items an examinee receives in a test. Therefore, this study investigated three levels of relatively short test length for T1: 5, 10, and 20.

*Number of items in T2.* The second factor determines the discreteness of the predictive distribution of the test statistic, and further influences the power of the method. With fewer items in T2, the test statistic will have fewer categories and thus the power may be limited by the discreteness of the predictive distribution. The sample size of T2 is

likely to vary in practice. Depending on different exposure scenarios, T2 may consist of only a few items (such as items found posted on the Internet), or consist of a larger set of items (such as all items that have been repeatedly used before). Three test lengths of T2 were explored: 5, 10 and 20.

*Proportion of truly compromised items in T2.* The third factor may have opposite effects on power when detecting at the item-set level and at the item level. At the item-set level, fewer compromised items will result in smaller effect on score increase and thus the test statistic is less likely to lie at the tail of the predictive distribution. On the contrary, deleting one compromised item may result in a larger change in  $\hat{\theta}$  when the remaining items are all uncompromised than when the remaining items are all compromised. Therefore, smaller proportion of compromised items is likely to increase the power at item level. For the item-set level, two proportions were examined-60% and 100% - to see how power drops when some noise (i.e. uncompromised items) was introduced to T2. Compromised items were randomly selected from T2. For the item-level detection, two smaller proportions were examined- 20% and 40% to get an understanding of the effect of increased noise (i.e. compromised items in this case) on power.

*Estimation methods.* As discussed in sections 3.1.1.1 and 3.1.1.2,  $p(\theta|y_1)$  can be obtained using three approaches: the bayesian approach with two less-informative priors -  $N(0, 2^2)$  and Jeffreys prior, and the fiducial distribution.

## 4.2 Data Simulation

The probability of correctly answering an item in the null condition was specified by the 2PL IRT model. 2PLM was considered here since it typically demonstrates much better fit than the 1PLM to empirical data and it does not have the problem with the estimation for the pseudo-guessing parameter in the 3PLM. In addition, due to the de-emphasis on the use of multiple-choice questions in educational tests, multiple-choice items are partly replaced by short constructed-response items, so less guessing is involved in item response, and 2PLM could demonstrate a reasonable fit in that case.

Dichotomous responses were simulated at five theta levels, i.e.,  $\theta = -2, -1, 0, 1, 2$  to investigate the detection effectiveness at low to high theta levels. The level of examinee proficiency is related to the ease of preknowledge detection. It can be expected that it is easier to detect preknowledge among low-performance examinees than that among high-performance examinees. Responses were first simulated in the null condition according to the 2PLM in each condition. For responses in the item-preknowledge condition, the probability of a correct response in the null condition was increased by 0.5 on corresponding compromised items (the probability was set to 1 if it exceeds 1 after manipulation). The “0.5” probability increment was chosen to represent a medium effect size on score increase due to item preknowledge. Data generation at each  $\theta$  level was replicated for 1000 times. The item discrimination parameters for both subsets in all conditions were randomly sampled from a truncated lognormal distribution with mean around 1.1 and standard deviation around 0.5 (i.e.  $\log N(0,0.2)$ ) between 0.75 and 2, which represents a realistic range adopted in most person-fit studies (Rupp, 2013). Item difficulty parameters for both subsets were randomly sampled from a truncated  $N(0,1)$

with lower and upper bound of -2 and 2. Item parameter values were summarized in Table A.1 in Appendix A. The true item parameters were used in person-level analysis in study 1.

### 4.3 Evaluation Criteria

To evaluate different estimation methods for obtaining  $p(\theta|y_1)$ , bias and mean squared error (MSE) were computed to evaluate the recovery of  $\theta$  using different estimation methods. Bias was computed as  $\frac{\sum_{r=1}^R(\hat{\theta}_r - \theta)}{R}$ , and MSE was computed as  $\frac{\sum_{r=1}^R(\hat{\theta}_r - \theta)^2}{R}$ , where  $R$  is the total number of replications.  $\hat{\theta}$  is the point estimate for  $\theta$ , which is EAP in the Bayesian approach, and the median in the fiducial approach.

To evaluate the detection effectiveness at item-set level, empirical type-I error and power of this method in different conditions using different test statistics were evaluated at a nominal level of  $\alpha=0.05$ . The type-I error rate at the item-set level was computed as the proportion of times each T2 was flagged as being compromised in the null condition, and power was computed as the proportion of times each compromised T2 was flagged in the preknowledge condition.

For the item-level detection, the empirical type-I error rate was computed for a randomly-selected item in the null condition. As for power calculation, due to computational burden, the detection rate for only one compromised item was computed as the power rate in each preknowledge condition. In addition, the detection rate for one non-compromised item in the preknowledge condition was also computed to obtain the false discovery rate. The item parameters for each of the two selected items were set to be the same across different preknowledge conditions, so that the comparisons among

different conditions were not confounded by the difference in item parameters. Specifically, the item parameters for the compromised item for which the power was calculated for were fixed to be  $a=1.0$ , and  $b=1.0$ , while the item parameters for the non-compromised item for which the false discovery rate was calculated for were fixed to be  $a=1.0$  and  $b=0.0$ . The compromised item was chosen to be a relatively difficult item so as to create a favorable situation for evaluating power. If the power is low for a hard item, one would not expect item-level detection to be useful in practice.

## **4.4 Results**

### **4.4.1 Recovery of $\theta$ by Different Estimation Methods**

The bias results in Table 4.1 show that using the fiducial distribution results in larger bias than the two Bayesian priors when T1 only consists of 5 items, and when T1 consists of 10 or 20 items, compared to using the two Bayesian priors, using the fiducial distribution results in larger bias at the lowest  $\theta$  level but smaller bias at the highest  $\theta$  level. However, the MSE results show that using the fiducial distribution results in the lowest MSE in general while using the Jeffreys prior tends to lead to the highest MSE for all  $\theta$  levels. As MSE accounts for the balance between the bias and variance, from this perspective, the fiducial distribution leads to a better point estimate for  $\theta$  than the two Bayesian priors.

Table 4.1: Bias and MSE from Different Estimation Methods

| Approach       | $\theta$ | BIAS   |        |        | MSE   |       |       |
|----------------|----------|--------|--------|--------|-------|-------|-------|
|                |          | T1=5   | T1=10  | T1=20  | T1=5  | T1=10 | T1=20 |
| Fiducial       | -2       | 0.258  | 0.188  | -0.126 | 0.453 | 0.224 | 0.372 |
|                | -1       | 0.022  | -0.079 | -0.011 | 0.749 | 0.373 | 0.137 |
|                | 0        | -0.018 | -0.024 | 0.015  | 0.786 | 0.289 | 0.168 |
|                | 1        | -0.118 | 0.029  | 0.119  | 0.498 | 0.338 | 0.391 |
|                | 2        | -0.617 | -0.046 | 0.046  | 0.575 | 0.218 | 0.417 |
| Normal Prior   | -2       | 0.212  | -0.044 | -0.100 | 0.660 | 0.398 | 0.315 |
|                | -1       | 0.080  | -0.152 | -0.003 | 0.810 | 0.475 | 0.142 |
|                | 0        | 0.051  | -0.046 | 0.050  | 0.799 | 0.305 | 0.177 |
|                | 1        | 0.038  | 0.033  | 0.155  | 0.744 | 0.372 | 0.378 |
|                | 2        | -0.271 | 0.084  | 0.044  | 0.544 | 0.403 | 0.413 |
| Jeffreys Prior | -2       | -0.009 | -0.128 | -0.147 | 0.778 | 0.532 | 0.381 |
|                | -1       | -0.044 | -0.151 | -0.017 | 0.976 | 0.537 | 0.140 |
|                | 0        | 0.016  | -0.024 | 0.021  | 0.954 | 0.300 | 0.171 |
|                | 1        | 0.108  | 0.054  | 0.147  | 0.967 | 0.388 | 0.437 |
|                | 2        | -0.099 | 0.148  | 0.134  | 0.651 | 0.485 | 0.570 |

#### 4.4.2 Type-I error at the item-set level

Table 4.2 summarizes the type-I error rates when fiducial distribution and the Jeffreys prior are used. Results for the rest of the conditions are summarized in Table A.2 to A.3 in Appendix A. The nominal level for the type-I error is 0.05, and a 95% normal-approximation confidence interval for a type-I error rate of 0.05 out of 1000 replications is (0.036, 0.063). Results shows that when T2 is short, using the summed score or the EAP often leads to conservative type-I error rates, especially at extreme  $\theta$  levels, which is due to the fact that the predictive distribution concentrates on very few values when T2 is short. When T2 is long, using the posterior variance leads to inflated type-I errors at  $\theta=0$  when T1 is short, but when T1 is long, the type-I error rates all fall below the upper bound of the 95% confidence interval. Three estimation methods lead to similar results in most conditions, except that using  $N(0,2^2)$  leads to slightly inflated type-I error for the EAP when T1 is short and T2 consists of 20 items.

Table 4.2: Empirical Type-I Error Using Fiducial and Jeffreys Prior

| Stat <sup>1</sup> | $\theta$ | T1=5, T2=5       |                  | T1=5, T2=20 |        | T1=20, T2=5 |       | T1=20, T2=20 |       |
|-------------------|----------|------------------|------------------|-------------|--------|-------------|-------|--------------|-------|
|                   |          | JEF <sup>1</sup> | FID <sup>1</sup> | JEF         | FID    | JEF         | FID   | JEF          | FID   |
| Sum               | -2       | 0.002            | 0.004            | 0.006       | 0.000  | 0.021       | 0.021 | 0.045        | 0.041 |
|                   | -1       | 0.014            | 0.020            | 0.039       | 0.012  | 0.018       | 0.019 | 0.033        | 0.029 |
|                   | 0        | 0.033            | 0.033            | 0.047       | 0.031  | 0.018       | 0.019 | 0.025        | 0.020 |
|                   | 1        | 0.012            | 0.012            | 0.033       | 0.029  | 0.001       | 0.002 | 0.036        | 0.030 |
|                   | 2        | 0.001            | 0.001            | 0.039       | 0.026  | 0.000       | 0.000 | 0.037        | 0.031 |
| EAP               | -2       | 0.004            | 0.010            | 0.009       | 0.001  | 0.036       | 0.037 | 0.057        | 0.056 |
|                   | -1       | 0.033            | 0.042            | 0.051       | 0.014  | 0.039       | 0.038 | 0.046        | 0.046 |
|                   | 0        | 0.037            | 0.040            | 0.053       | 0.041  | 0.024       | 0.026 | 0.040        | 0.034 |
|                   | 1        | 0.014            | 0.014            | 0.051       | 0.035  | 0.001       | 0.002 | 0.044        | 0.043 |
|                   | 2        | 0.001            | 0.001            | 0.061       | 0.038  | 0.000       | 0.000 | 0.053        | 0.039 |
| VAR               | -2       | 0.004            | 0.004            | 0.025       | 0.014  | 0.016       | 0.017 | 0.050        | 0.048 |
|                   | -1       | 0.038            | 0.042            | 0.034       | 0.017  | 0.024       | 0.023 | 0.047        | 0.043 |
|                   | 0        | 0.037            | 0.047            | 0.099*      | 0.090* | 0.026       | 0.023 | 0.040        | 0.040 |
|                   | 1        | 0.035            | 0.034            | 0.017       | 0.009  | 0.033       | 0.034 | 0.057        | 0.049 |
|                   | 2        | 0.002            | 0.002            | 0.021       | 0.017  | 0.005       | 0.006 | 0.041        | 0.029 |

Note. <sup>1</sup>Stat=test statistics, JEF=Jeffreys prior, FID=fiducial distribution,

\* represents  $p$ -value exceeds the upper bound of the 95% normal-approximation confidence interval for the type-I error of 0.05

#### 4.4.3 Power at the item-set level

Table 4.3 summarizes the detection power when using the fiducial distribution and the Jeffreys prior. Results for the remaining conditions are summarized in Table A.4 to A.6 in Appendix A. The power increases as the lengths of T1 and T2 increases. When T2 is short, it is hard to detect preknowledge among high proficiency examinees: the power of all test statistics is less than 0.1 for  $\theta \geq 1$  when T2 only consists of 5 items; when T2 is increased, using the summed score or the EAP leads to power above 0.5 for  $\theta = 1$  when the entire T2 is compromised. The power for the summed score and the EAP both decreases to a large extent as the compromise rate in T2 drops from 100% to 60%, but the power for the posterior variance shows an increase for  $\theta \leq -1$  when T2 is long. This is because as the compromise rate drops, the effect of compromised responses on the

score inflation decreases, but the  $\theta$  posterior distribution becomes more flat due to the presence of more reversed Guttman patterns in low-proficiency examinees' response patterns. As for the comparison among three statistics, when T2 is short, such as when T2 only consists of 5 items, using the EAP leads to larger power than using the summed score, especially when the compromise rate is only 0.6. As T2 becomes longer, the power difference between the EAP and the summed score is very small, only up to the second decimal place in most conditions. The posterior variance demonstrates lower power than the other two statistics, partly because a two-sided test was conducted for it while one-sided tests were conducted for the other two statistics. Lastly, the difference among the three methods is at the second or third decimal places in many conditions. In conditions where the difference is as large as 0.1 or 0.2- such as when T2 contains 20 items and when T1 only consists of 5 items- the two Bayesian approaches have more similar power, while they exhibit slightly larger power than the fiducial approach.

Table 4.3: Power Rate Using Fiducial and Jeffreys Prior

| Stat | Rate <sup>1</sup> | $\theta$ | T1=5, T2=5 |       | T1=5, T2=20 |       | T1=20, T2=5 |       | T1=20, T2=20 |       |
|------|-------------------|----------|------------|-------|-------------|-------|-------------|-------|--------------|-------|
|      |                   |          | FID        | JEF   | FID         | JEF   | FID         | JEF   | FID          | JEF   |
| Sum  | 1                 | -2       | 0.398      | 0.426 | 0.737       | 0.775 | 0.702       | 0.701 | 0.993        | 0.994 |
|      |                   | -1       | 0.368      | 0.371 | 0.619       | 0.676 | 0.616       | 0.615 | 0.978        | 0.978 |
|      |                   | 0        | 0.262      | 0.262 | 0.600       | 0.653 | 0.374       | 0.386 | 0.929        | 0.934 |
|      |                   | 1        | 0.051      | 0.053 | 0.408       | 0.590 | 0.016       | 0.017 | 0.737        | 0.778 |
|      |                   | 2        | 0.001      | 0.002 | 0.144       | 0.377 | 0.000       | 0.000 | 0.291        | 0.354 |
|      | 0.6               | -2       | 0.140      | 0.177 | 0.308       | 0.457 | 0.373       | 0.379 | 0.833        | 0.842 |
|      |                   | -1       | 0.140      | 0.148 | 0.296       | 0.339 | 0.243       | 0.244 | 0.705        | 0.708 |
|      |                   | 0        | 0.099      | 0.099 | 0.216       | 0.254 | 0.083       | 0.083 | 0.464        | 0.478 |
|      |                   | 1        | 0.018      | 0.019 | 0.126       | 0.166 | 0.006       | 0.005 | 0.212        | 0.229 |
|      |                   | 2        | 0.001      | 0.002 | 0.057       | 0.104 | 0.000       | 0.000 | 0.075        | 0.089 |
| EAP  | 1                 | -2       | 0.430      | 0.516 | 0.746       | 0.788 | 0.786       | 0.786 | 0.996        | 0.996 |
|      |                   | -1       | 0.382      | 0.399 | 0.617       | 0.684 | 0.671       | 0.665 | 0.976        | 0.979 |
|      |                   | 0        | 0.262      | 0.262 | 0.609       | 0.668 | 0.387       | 0.395 | 0.933        | 0.940 |
|      |                   | 1        | 0.051      | 0.053 | 0.408       | 0.609 | 0.016       | 0.017 | 0.764        | 0.801 |
|      |                   | 2        | 0.001      | 0.002 | 0.144       | 0.377 | 0.000       | 0.000 | 0.291        | 0.354 |
|      | 0.6               | -2       | 0.268      | 0.340 | 0.351       | 0.469 | 0.587       | 0.598 | 0.876        | 0.881 |
|      |                   | -1       | 0.220      | 0.252 | 0.298       | 0.345 | 0.412       | 0.416 | 0.711        | 0.710 |
|      |                   | 0        | 0.127      | 0.134 | 0.217       | 0.258 | 0.122       | 0.125 | 0.479        | 0.489 |
|      |                   | 1        | 0.026      | 0.028 | 0.133       | 0.174 | 0.007       | 0.006 | 0.229        | 0.251 |
|      |                   | 2        | 0.001      | 0.002 | 0.058       | 0.107 | 0.000       | 0.000 | 0.094        | 0.108 |
| VAR  | 1                 | -2       | 0.134      | 0.171 | 0.231       | 0.365 | 0.274       | 0.265 | 0.705        | 0.720 |
|      |                   | -1       | 0.119      | 0.144 | 0.018       | 0.044 | 0.243       | 0.242 | 0.064        | 0.065 |
|      |                   | 0        | 0.113      | 0.168 | 0.058       | 0.124 | 0.225       | 0.226 | 0.765        | 0.782 |
|      |                   | 1        | 0.013      | 0.029 | 0.056       | 0.175 | 0.005       | 0.003 | 0.465        | 0.503 |
|      |                   | 2        | 0.001      | 0.001 | 0.001       | 0.031 | 0.000       | 0.000 | 0.034        | 0.041 |
|      | 0.6               | -2       | 0.197      | 0.223 | 0.242       | 0.396 | 0.321       | 0.316 | 0.763        | 0.770 |
|      |                   | -1       | 0.117      | 0.125 | 0.178       | 0.222 | 0.096       | 0.102 | 0.302        | 0.298 |
|      |                   | 0        | 0.021      | 0.037 | 0.006       | 0.009 | 0.043       | 0.045 | 0.064        | 0.069 |
|      |                   | 1        | 0.005      | 0.008 | 0.009       | 0.024 | 0.002       | 0.002 | 0.137        | 0.170 |
|      |                   | 2        | 0.001      | 0.001 | 0.009       | 0.035 | 0.000       | 0.002 | 0.036        | 0.040 |

Note. <sup>1</sup>Rate=compromise rate in T2

#### 4.4.4 Type-I Error at the item level

Item-level empirical type-I error rates were examined for the two items that were used in the evaluation of power and false positive rate- one item with  $a=1.0$  and  $b=1.0$ , and the other with  $a=1.0$  and  $b=0.0$ . Table 4.4 summarizes item-level type-I error rates

for several length combinations of T1 and T2. Results for the remaining conditions are summarized in Table A.7 in Appendix A. Results show that the item-level type-I error is likely to be inflated as T2 contains more items, and this inflation occurs regardless of the estimation method being used.

Table 4.4: Empirical Type-I Error at Item Level

| Estimation           | $\theta$ | T1=5, T2=5 |       | T1=5, T2=20 |        | T1=20, T2=5 |       | T1=20, T2=20 |        |
|----------------------|----------|------------|-------|-------------|--------|-------------|-------|--------------|--------|
|                      |          | b=1        | b=0   | b=1         | b=0    | b=1         | b=0   | b=1          | b=0    |
| FID                  | -2       | 0.031      | 0.002 | 0.030       | 0.041  | 0.037       | 0.015 | 0.054        | 0.057  |
|                      | -1       | 0.031      | 0.028 | 0.063       | 0.054  | 0.030       | 0.023 | 0.065*       | 0.051  |
|                      | 0        | 0.040      | 0.047 | 0.044       | 0.050  | 0.031       | 0.030 | 0.049        | 0.064* |
|                      | 1        | 0.020      | 0.024 | 0.040       | 0.038  | 0.026       | 0.022 | 0.053        | 0.042  |
|                      | 2        | 0.008      | 0.006 | 0.064*      | 0.062  | 0.022       | 0.011 | 0.046        | 0.058  |
| N(0,2 <sup>2</sup> ) | -2       | 0.028      | 0.004 | 0.036       | 0.052  | 0.030       | 0.019 | 0.046        | 0.050  |
|                      | -1       | 0.048      | 0.021 | 0.052       | 0.066* | 0.032       | 0.018 | 0.055        | 0.063  |
|                      | 0        | 0.048      | 0.037 | 0.062       | 0.056  | 0.039       | 0.026 | 0.052        | 0.041  |
|                      | 1        | 0.018      | 0.028 | 0.044       | 0.048  | 0.023       | 0.025 | 0.065*       | 0.069* |
|                      | 2        | 0.003      | 0.007 | 0.061       | 0.050  | 0.019       | 0.008 | 0.054        | 0.057  |
| JEF                  | -2       | 0.030      | 0.015 | 0.041       | 0.054  | 0.031       | 0.021 | 0.049        | 0.054  |
|                      | -1       | 0.059      | 0.030 | 0.062       | 0.072* | 0.035       | 0.018 | 0.053        | 0.062  |
|                      | 0        | 0.048      | 0.040 | 0.066*      | 0.056  | 0.040       | 0.023 | 0.051        | 0.038  |
|                      | 1        | 0.017      | 0.036 | 0.048       | 0.056  | 0.021       | 0.026 | 0.063        | 0.074* |
|                      | 2        | 0.004      | 0.010 | 0.060       | 0.053  | 0.020       | 0.014 | 0.058        | 0.067* |

#### 4.4.5 Power and False Positive Rate at the item level

Table 4.5 summarizes the power and false positive rate when the Jeffreys prior and the fiducial distribution are used. Results for the remaining conditions are summarized in Table A.8 to A.9 in Appendix A. It can be seen that there is not much power at the item level: the empirical power rate is between 0.2 and 0.3 at the lowest  $\theta$  level, and lower than 0.2 for the rest of the  $\theta$  levels. A higher compromise rate leads to a lower power, which is consistent with the hypothesis that the presence of compromised items after deleting one item will add noise to the detection. The impact of lengths of T1

and T2 on item-level power shows different patterns from those observed in the item-set level detection. The item-level power does not show a uniformly increasing pattern in different conditions as T1 becomes longer. For instance, when the compromise rate is 20%, increasing the length of T1 results in a smaller power at the lowest three  $\theta$  levels, but when the compromise rate is 40%, increasing the length of T1 leads to a larger power when T2 contains 10 or 20 items. Although the effect of T1 length is not consistent in different conditions, further analyses show that the power results do converge to those obtained from the true predictive distribution (i.e., evaluated at the true  $\theta$ ) of the test statistics as T1 becomes longer. The fact that a shorter length of T1 results in a larger power in certain conditions is probably due to the uncertainty in the predictive distribution constructed through  $p(\theta|y_1)$ . As for the effect of T2 length, when T1 is long, which approximates the situation where the predictive distribution is constructed using the true  $\theta$ , the power tends to increase first as the length of T2 increases from 5 to 10 and then the power decreases as the length of T2 increases from 10 to 20. The phenomenon is possibly due to the discrete nature of the test statistic's distribution, which may cause a non-monotonic change of the empirical rejection rate as the length of T2 changes. However, it is necessary to consider more test length conditions in order to verify this conjecture. Researchers such as Brown, Cai, and DasGupta (2001) also found this oscillation phenomenon when studying the coverage probability of the confidence interval for the binomial proportion. Regarding the comparison among three estimation methods, using the Jeffreys prior or  $N(0,2^2)$  results in slightly larger power than using the fiducial distribution.

The false positive rate is high at the lowest two  $\theta$  levels, and it increases as the compromise rate increases. Increase in the length of T2 tends to result in a larger false positive rate, especially when T1 is long, and the Jeffreys prior can result in a larger false positive rate than the other two procedures when the compromise rate is 40%.

Table 4.5: Item-Level Power and False Positive Rate

|                      | T2 | $\theta$ | Power    |       |          |       | False Positive |       |          |       |
|----------------------|----|----------|----------|-------|----------|-------|----------------|-------|----------|-------|
|                      |    |          | Rate=0.2 |       | Rate=0.4 |       | Rate=0.2       |       | Rate=0.4 |       |
|                      |    |          | T1=5     | T1=20 | T1=5     | T1=20 | T1=5           | T1=20 | T1=5     | T1=20 |
| JEF                  | 5  | -2       | 0.314    | 0.214 | 0.155    | 0.176 | 0.042          | 0.086 | 0.263    | 0.365 |
|                      |    | -1       | 0.172    | 0.068 | 0.087    | 0.055 | 0.067          | 0.065 | 0.224    | 0.251 |
|                      |    | 0        | 0.091    | 0.084 | 0.048    | 0.033 | 0.050          | 0.036 | 0.147    | 0.135 |
|                      |    | 1        | 0.026    | 0.038 | 0.006    | 0.007 | 0.032          | 0.012 | 0.061    | 0.027 |
|                      |    | 2        | 0.008    | 0.025 | 0.001    | 0.002 | 0.009          | 0.007 | 0.008    | 0.004 |
|                      | 20 | -2       | 0.334    | 0.270 | 0.213    | 0.305 | 0.056          | 0.197 | 0.246    | 0.511 |
|                      |    | -1       | 0.136    | 0.065 | 0.075    | 0.112 | 0.090          | 0.112 | 0.160    | 0.295 |
|                      |    | 0        | 0.071    | 0.060 | 0.047    | 0.056 | 0.076          | 0.083 | 0.131    | 0.168 |
|                      |    | 1        | 0.014    | 0.018 | 0.002    | 0.004 | 0.080          | 0.099 | 0.133    | 0.110 |
|                      |    | 2        | 0.001    | 0.013 | 0.000    | 0.005 | 0.074          | 0.075 | 0.090    | 0.085 |
| N(0,2 <sup>2</sup> ) | 5  | -2       | 0.266    | 0.203 | 0.152    | 0.166 | 0.020          | 0.074 | 0.105    | 0.208 |
|                      |    | -1       | 0.156    | 0.068 | 0.101    | 0.064 | 0.053          | 0.063 | 0.150    | 0.180 |
|                      |    | 0        | 0.097    | 0.088 | 0.057    | 0.046 | 0.044          | 0.033 | 0.126    | 0.122 |
|                      |    | 1        | 0.027    | 0.039 | 0.012    | 0.010 | 0.028          | 0.011 | 0.063    | 0.021 |
|                      |    | 2        | 0.007    | 0.024 | 0.001    | 0.007 | 0.007          | 0.003 | 0.007    | 0.001 |
|                      | 20 | -2       | 0.281    | 0.220 | 0.117    | 0.266 | 0.035          | 0.176 | 0.185    | 0.511 |
|                      |    | -1       | 0.129    | 0.076 | 0.076    | 0.114 | 0.070          | 0.115 | 0.132    | 0.302 |
|                      |    | 0        | 0.062    | 0.039 | 0.044    | 0.046 | 0.086          | 0.094 | 0.130    | 0.182 |
|                      |    | 1        | 0.014    | 0.035 | 0.003    | 0.007 | 0.080          | 0.064 | 0.126    | 0.096 |
|                      |    | 2        | 0.000    | 0.011 | 0.000    | 0.005 | 0.081          | 0.067 | 0.094    | 0.083 |
| FID                  | 5  | -2       | 0.283    | 0.229 | 0.152    | 0.197 | 0.021          | 0.071 | 0.099    | 0.210 |
|                      |    | -1       | 0.153    | 0.076 | 0.108    | 0.084 | 0.057          | 0.067 | 0.156    | 0.186 |
|                      |    | 0        | 0.081    | 0.052 | 0.054    | 0.037 | 0.060          | 0.045 | 0.150    | 0.153 |
|                      |    | 1        | 0.026    | 0.046 | 0.006    | 0.012 | 0.034          | 0.015 | 0.070    | 0.023 |
|                      |    | 2        | 0.010    | 0.028 | 0.001    | 0.003 | 0.005          | 0.004 | 0.007    | 0.002 |
|                      | 20 | -2       | 0.271    | 0.241 | 0.082    | 0.300 | 0.017          | 0.185 | 0.107    | 0.513 |
|                      |    | -1       | 0.116    | 0.070 | 0.052    | 0.119 | 0.060          | 0.124 | 0.131    | 0.306 |
|                      |    | 0        | 0.043    | 0.034 | 0.026    | 0.050 | 0.075          | 0.092 | 0.121    | 0.187 |
|                      |    | 1        | 0.003    | 0.019 | 0.000    | 0.003 | 0.082          | 0.065 | 0.132    | 0.103 |
|                      |    | 2        | 0.000    | 0.014 | 0.000    | 0.003 | 0.084          | 0.073 | 0.104    | 0.089 |

## 4.5 Discussion

Study 1 evaluated the empirical type-I error and power of the predictive checking method under different lengths of T1 and T2 as well as different test compromise rates. Considering the posterior distribution of  $\theta$  estimated from a short secure section might be largely affected by the use of an inappropriately specified prior distribution, study 1 investigated the performance of two less-informative Bayesian priors and the fiducial distribution which does not need to specify a prior distribution. The  $\theta$  recovery results showed that using the fiducial distribution led to smaller MSE for the point estimate of  $\theta$ , but the detection effectiveness among the three methods was quite similar, especially as the secure section consists of 10 or more items.

Regarding the detection effectiveness under different factors, results suggested that the length of both sections played an important role in the detection. In one extreme condition where the secure section only consisted of five items, the detection power was low except when a large set of items were compromised. This suggests using just five or ten secure items has the potential to detect severe test compromise situations. In the other extreme condition where the possibly compromised section only consisted of five items, high detection power could be achieved if all items in this section were compromised and if a long secure section was available, and using the EAP as the test statistic could lead to higher power than using the summed score or the posterior variance. In addition, if a large possibly compromised section only contains a small proportion of truly compromised items, the detection power of the summed score or the EAP could reduce to a large extent. For instance, compared to the condition where the possibly compromised set only consisted of five items but all of them were compromised, the detection power of

the summed score or the EAP was slightly lower when there were six truly compromised items in a 10-item possibly compromised set. Therefore, in order to maintain the detection power, one can apply predictive checking to items that are most likely to be compromised. Alternatively, one can only include relatively difficult items in the possibly compromised subset, as it is hard to detect preknowledge on easy items by any means.

Item-level detection does not turn out to be effective in this study. Low power was observed even at the lowest  $\theta$  level, and high false positive rate occurred as the compromise rate increased. This result is not surprising given that each item only has two response categories. Item-level detection may have larger power on a polytomous item with a larger number of response categories. However, given the “leave-one-out” nature of the item-level detection, it can be expected that the high false positive rate will remain a problem.

## CHAPTER 5

### STUDY 2: COMPARISON OF METHODS

#### 5.1 Background

Following study 1, which mainly aimed at understanding the statistical properties of the predictive checking method in different conditions, another simulation study (labeled as *Study 2* hereinafter) was conducted to evaluate this method in two simulated test compromise situations that are likely to happen in practice, and compare this method with other approaches, so as to add more practical implications to this project. Study 2 is specific to the research plan initiated by ETS under the Harold Gulliksen Psychometric Research Fellowship Program. The author would like to thank researchers at ETS for their contribution to this study.

Study 2 consisted of two smaller simulation designs, each mimicking a practical test compromise situation. The first situation has been discussed mostly up to now, which is preknowledge of items due to the use of a limited item pool in an on-demand testing program (hereinafter called as “*shallow pool situation*”), and the second situation is a more serious security breach problem in which the keys to an entire operational test section are exposed (hereinafter called as “*key exposure situation*”). This could occur on tests administered internationally, where the time-zone difference provides an opportunity for examinees taking the test later to get access to the keys to an entire operational test section. In both situations, the pretest section could serve as a baseline to infer whether an examinee has preknowledge on the operational section: in the first situation, the pretest section has very low exposure rate, so examinees rarely have chance

to see them beforehand; in the second situation, pretest sections are usually randomly assigned and thus different people usually get different pretest sections, so examinees may not have the keys or may use the wrong keys for the pretest section.

## 5.2 Shallow Pool Simulation

The shallow pool situation was simulated using an MST design. The MST design was adopted because it is becoming increasingly popular and it is currently used by several operational testing programs (e.g., GRE). Since current testing programs that use MST designs are on-demand, they are likely to have to deal with the item exposure problem. A 1-3 MST design was employed to mimic the design used by the Revised GRE. The 1-3 MST design consisted of two operational stages. Based on a person's responses to the routing module in the first stage, subject to the routing rule, one of the three modules in the second stage was administered. In this study, the routing was based on a person's proficiency estimate (i.e.,  $\hat{\theta}$ ) in Stage 1. If  $\hat{\theta} < -0.43$ , the easy module was administered, and if  $\hat{\theta} > 0.43$ , the hard module was administered. Otherwise, the module with moderate difficulty was administered. The thresholds of  $\pm 0.43$  are the 33rd and 66th percentile in the standard normal distribution, and they were chosen as the routing thresholds such that each module would have roughly the same exposure rate among the examinee population with  $\theta \sim N(0,1)$ . The two stages served as the possibly compromised section (i.e. T2) in this study. Each stage consisted of 20 items, so T2 consisted of 40 items in total. In addition, a pretest section, served as the uncompromised section (i.e., T1), consisting of 20 items was administered. It was assumed that the administration of

the pretest section was not subject to any routing rule, and the items in the pretest section were chosen to have a broad range of item difficulty.

The 3PL IRT model was used to generate the responses and the true item parameters for each section are summarized in Table 5.1 below. Instead of using item parameter estimates from a real dataset or sampling item parameters from a distribution, each parameter was set to several fixed values, and each value for a given parameter was crossed with all possible values in the other parameters to create the data-generating parameter set for a given MST module. Specifically, two values were chosen for the  $a$ -parameter: 1.0 and 1.4 to represent moderate and high discriminating items on the logistic scale. Two values were chosen for the  $c$ -parameter: 0.0 and 0.15 to represent the  $c$ -parameter value for short constructed response items and multiple choice items. Five values were chosen for the  $b$ -parameter: for pretest section and stage 1-routing module, five values between -1.5 and 1.5 with an increment of 0.75 were chosen; for stage-2 easy module, five values between -1.5 and 0.5 with an increment of 0.5 were chosen; for stage-2 middle difficulty module, five values between -1 and 1 with an increment of 0.5 were chosen, and for stage-2 hard module, five values between -0.5 and 1.5 with an increment of 0.5 were chosen. These values represent the typical parameter values in a high-stakes large scale assessment, and this way of choosing item parameters aimed at representing a more general case, so as to increase the generalizability of the results.

Table 5.1: True Item Parameters in Shallow Pool Situation in Study 2

|   | Pretest section                       | Stage 1-<br>Routing                   | Stage 2-<br>Easy                   | Stage 2-<br>Middle                  | Stage 2-<br>Hard                 |
|---|---------------------------------------|---------------------------------------|------------------------------------|-------------------------------------|----------------------------------|
| a | 1, 1.4                                | 1, 1.4                                | 1, 1.4                             | 1, 1.4                              | 1, 1.4                           |
| b | -1.5,<br>-0.75,<br>0,<br>0.75,<br>1.5 | -1.5,<br>-0.75,<br>0,<br>0.75,<br>1.5 | -1.5,<br>-1,<br>-0.5,<br>0,<br>0.5 | -1.0,<br>-0.5,<br>0,<br>0.5,<br>1.0 | -0.5,<br>0,<br>0.5,<br>1,<br>1.5 |
| c | 0, 0.15                               | 0, 0.15                               | 0, 0.15                            | 0, 0.15                             | 0, 0.15                          |

*Note.* Each value for a given parameter is crossed with all possible values in the other parameters to create the data-generating parameter set for a given section. Item discrimination parameter is on logistic scale.

To simulate compromised responses in this situation, preknowledge was introduced to the hardest items in a given section, and a correct response was assigned to the compromised item. Three factors were manipulated to simulate different compromise conditions. The first factor was the proportion of compromised items in T2: 0%, 10%, 20% and 40%. The level of 10% was used since it represented a realistic level in practice, as not many items would be compromised in practice due to careful test designs and exposure rate control procedures. The level of 20% and 40% were used to represent worse scenarios and to investigate whether the detection methods had sufficient power to detect more serious test compromise situations. The average theta estimation error ( $\hat{\theta} - \theta$ ) across replications was calculated to evaluate the consequences of item preknowledge on score inflation, where  $\hat{\theta}$  was estimated using MLE based on responses to the entire T2 (i.e. two stages). In the null condition, the bias, variance, and root mean square error for  $\hat{\theta}$  were calculated as a baseline reference to compare to the score inflation in each compromise condition.

The second factor was the compromised MST stage: item preknowledge could occur at stage 1 or 2 or both. It was expected that preknowledge would be harder to detect

if it occurred in a stage that contained easier items (relative to a person’s ability) than that contained harder items. The third factor was the presence of item parameter errors in person-level analysis. Person-level analysis was conducted first using true item parameters, and then using item parameter estimates, so that the effect of error in item parameter estimation on detection was investigated. To obtain item parameter estimates, item calibrations were conducted using correctly or incorrectly specified models respectively. To be specific, response data were first generated for 5000 examinees with  $\theta \sim N(0,1)$ , and then item calibration was conducted by fitting 3PLM and 2PLM, respectively. This set-up was used to mimic the practical situation where item parameter estimates often contain error due to either sampling variability or model misfit. Table 5.2 below summarizes the conditions under investigation.

Table 5.2: Conditions in Shallow Pool Situation

| Factor                      | Condition  |
|-----------------------------|--|
| Compromised stage           | Stage 1  |
|                             | Stage 2  |
|                             | Stage 1 & 2  |
| % of compromised item in T2 | 0%   |
|                             | 10%  |
|                             | 20%  |
|                             | 40%  |
| Item parameter error        | No error: true item parameter used   |
|                             | Estimation error: item parameter estimates from the correctly specified model                  |
|                             | Model misfit & estimation error: item parameter estimates from the incorrectly specified model |

In the shallow pool situation, only the likelihood ratio test and the adapted KL divergence were chosen to compare to the predictive checking method. The regression method was not used mainly because it needs to be implemented based on responses by a

group of examinees. With an MST design, examinees in a group typically do not take the same test items, and thus the simple linear regression model built on their summed scores on each section may be affected to some extent by the fact that the summed scores are not calculated based on the same sets of items for everyone in the group<sup>3</sup>.

### **5.3 Key Exposure Simulation**

Different from the shallow pool situation, key exposure situation represents a much more serious security breach problem where a large number of keys are exposed. It is more likely to happen on a fixed-form test, as there is less overlap among the items administered to different examinees in an adaptive test. Therefore, a fixed-form test with 60 items was simulated. Out of the 60 items, the pretest section consisted of 20 items, and the operational section consisted of 40 items. It was assumed that there were multiple pretest sections and they were randomly administered to examinees in different test administrations, while the same operational section was repeatedly used across test administrations.

In key exposure situation, two scenarios were considered: in the first scenario, an examinee answered the pretest items based on his/her real proficiency, instead of relying on the keys. This could happen when examinees have some information about the items when they get the keys, so they can judge which keys are for which items and when they realize that the pretest items they are administered do not match those provided by the

---

<sup>3</sup> This problem could be overcome by building the simple linear regression based on examinees'  $\hat{\theta}$  instead of summed scores. However, in practice, the regression is typically run on simple summed score to avoid fitting an IRT model. Therefore, the regression method is not considered in the shallow pool situation.

source of the keys (due to the random assignment of the pretest sections), they respond to the pretest section based on their real proficiency. In the second scenario, an examinee's responses on the pretest section were based on the wrong key or based on random responses. This could happen when examinees do not have information about the items but only have the keys, so they will apply the keys directly to all items they are administered. It could also be the case where examinees recognize the keys provided by the source do not match the pretest section they are administered, and thus they realize it is the pretest section that does not count towards their scores, so they simply randomly respond to those items. In this scenario, we assume an examinee's responses to the pretest section are correct by chance and their responses to the operational section are based on the keys provided by a high-proficiency source examinee.

Within each of the two scenarios above, two conditions were considered: in one condition, an examinee memorized the keys provided by the source perfectly (i.e. 0% incorrect memorization), and in another condition, an examinee memorized 20% keys on the operational section incorrectly (i.e. 20% incorrect memorization), as memorizing keys on all items could be hard in a moderate to long test. The incorrect memorization was simulated to 20% items in the operational section with the largest item difficulty. The hardest items were chosen for three reasons: (1) an examinee is more likely to recognize a key is wrong on easier items than on a harder item; (2) an examinee is more likely to make gridding errors later in a section than earlier in the section, and the items at the end of a section tend to be harder items; (3) an examinee is more likely to forget keys at the end of a list.

The two factors above- responses to the pretest section and the operation section- are crossed, resulting in four simulation conditions. In addition to the four conditions, the null condition in which responses to both subsets are based on one's real proficiency was also simulated to investigate the type-I error rate; and the condition in which responses to the pretest section are random or based on wrong answers, but responses to the operational section are based on one's real proficiency was also simulated to evaluate the false positive detection when responses to T1 do not reflect one's proficiency. Table 5.3 below summarizes the six simulation conditions.

Table 5.3: Simulation Conditions in Key Exposure Situation

| Responses to the pretest section    | Responses to the operational section    | Condition Label |
|-------------------------------------|---|-----------------|
| Based on real proficiency           | Based on real proficiency               | T1N_T2N         |
|                                     | Identical with the keys provided        | T1N_T2C         |
|                                     | With 20% memorization error on the keys | T1N_T2C2        |
| Random responses/based on wrong key | Based on real proficiency               | T1R_T2N         |
|                                     | Identical with the keys provided        | T1R_T2C         |
|                                     | With 20% memorization error on the keys | T1R_T2C2        |

As for data simulation, to simulate responses on the pretest section, in the first scenario where the pretest section was not compromised, the responses were simulated using the 3PLM, whereas in the second scenario where the pretest section was randomly responded to, responses on the pretest section were simulated to be correct with probability of 0.25. To simulate the compromised responses on the operational section,

instead of assuming an examinee had perfect keys, keys were simulated by using a response pattern from a source examinee of high-proficiency. Specifically,  $\theta=2$  was used as the source ability, and if  $P(Y_i=1|\theta=2)$  is greater than 0.7, a correct response was assigned, otherwise, an incorrect response was assigned. In the condition with 0% incorrect memorization, an examinee's responses were identical with the keys, while in condition with 20% incorrect memorization, an incorrect response was assigned to the incorrectly memorized items. The true item parameters in each section are summarized in Table 5.4.

Table 5.4: Summary of True Item Parameters in Key Exposure Generation

|   | Pretest section   | Operational section  |
|---|---|--|
| a | 1, 1.4  | 1, 1.4   |
| b | b/w -1.5 and 1.5, with<br>mean=0<br>(-1.5, -0.75, 0, 0.75, 1.5) | b/w -1.5 and 1.5, with<br>mean=0<br>(-1.5, -1.17, -0.83, -0.5, -<br>0.17, 0.17, 0.5, 0.83, 1.17,<br>1.5) |
| c | 0, 0.15   | 0, 0.15  |

In the key exposure scenario, the likelihood ratio test, the adapted KL divergence, and the regression-based method were all used to compare to the predictive checking method. The regression-based method is different from the other three methods in that it needs to be conducted based on a group of examinees, while the analysis in the other methods is simply based on the response vector of one examinee. When evaluating the regression-based method, responses in the null condition were generated for 5000 examinees with  $\theta \sim N(0,1)$ . In the cheating condition, compromised responses were simulated on 5% examinees. A relatively small number of cheaters were chosen here because cheating is still a low-probability event in a high-stakes operational testing

program. Furthermore, considering the estimation of the regression line would be affected to a different extent if cheaters follow different proficiency distributions, two distributions were assumed for the proficiency of cheaters:  $\theta \sim N(0,1)$  and  $N(-1,1)$ <sup>4</sup>. These two distributions were chosen based on the assumption that examinees with low to moderate proficiency levels have a higher motivation to cheat. Other assumptions could be adapted but they were not in this study.

#### 5.4 Evaluation Criteria

For each method proposed here, detection rate at both person level and group level are evaluated. Person-level detection is evaluated at nine  $\theta$  levels from -2 to 2 with equal increment. For each given theta, response generation is replicated for 1000 times, and the detection rate under no preknowledge condition is calculated as the false positive rate and that under preknowledge condition is the hit rate. For the regression-based approach, considering the sampled examinee group does not contain theta values that equal exactly to the eleven theta levels under investigation, the theta value closest to each theta level under investigation is replaced by it, and data generation is replicated for 1000 times for the entire group.

---

<sup>4</sup> The cheating group with  $\theta \sim N(0,1)$  was chosen by randomly sampling 250 examinees from the 5000 examinees with  $\theta \sim N(0,1)$ . For the cheating group with  $\theta \sim N(-1,1)$ , a group of 250 was simulated from  $\theta \sim N(-1,1)$  first, and then another group of 4750 was simulated from  $\theta \sim N(250/4750,1)$ , and these two groups were combined to mimic a group of 5000 examinees with  $\theta \sim N(0,1)$ . The mixture distribution does not follow exactly a standard normal distribution. The mean of the mixture distribution is 0, and the standard deviation is slightly larger than 1, but the difference is small enough to have any practical consequences for the purpose of this study.

For the group-level detection rate in all methods except the regression analysis, the detection rates at 13  $\theta$  levels from -3 to 3 were first obtained, and then the person-level detection rate was multiplied by the population weight at each  $\theta$  in a given distribution. The weighted sum was the group-level detection rate. Two distributions were assumed for the ability of examinees with preknowledge:  $\theta \sim N(0,1)$  and  $N(-1,1)$ . The group-level detection rate in the regression method was simply calculated by the average detection rate in each cheating group across replications.

## **5.5 Results in Shallow Pool Simulation**

### **5.5.1 Theta estimation error**

Table 5.5 summarizes the  $\theta$  recovery results - the bias, variance (Var) and root mean squared error (RMSE) for  $\hat{\theta}$  in the null condition. Table 5.6 presents the  $\hat{\theta}$  inflation in different test compromise conditions. Results on  $\theta$  recovery in the null condition suggest that using item parameter estimates from either the 3PLM or the 2PLM does not lead to a large difference from using the true item parameters at most  $\theta$  levels. Moderate differences occur at the lower or higher end of the  $\theta$  continuum. Specifically, using item parameter estimates from the 3PLM leads to larger negative bias for  $\theta$  at the lower end, while using item parameters from the 2PLM leads to slightly larger positive bias for  $\theta$  at the higher end. Results on the variance of  $\hat{\theta}$  and the RMSE also suggest that using the 3PLM estimates leads to less efficient estimates for lower  $\theta$  levels, and using estimates from the 2PLM leads to less efficient estimates for higher  $\theta$  levels.

Table 5.5: BIAS and RMSE in Null Condition

| $\theta$ | BIAS  |       |       | VAR  |      |      | RMSE |      |      |
|----------|-------|-------|-------|------|------|------|------|------|------|
|          | True* | 3PLM* | 2PLM* | True | 3PLM | 2PLM | True | 3PLM | 2PLM |
| -2       | -0.15 | -0.54 | 0.03  | 0.35 | 0.67 | 0.16 | 0.61 | 0.98 | 0.40 |
| -1.5     | -0.04 | -0.19 | 0.01  | 0.16 | 0.29 | 0.11 | 0.40 | 0.57 | 0.33 |
| -1       | -0.02 | -0.03 | -0.03 | 0.14 | 0.21 | 0.11 | 0.38 | 0.46 | 0.34 |
| -0.5     | 0.01  | 0.05  | -0.02 | 0.11 | 0.12 | 0.10 | 0.33 | 0.35 | 0.32 |
| 0        | 0.00  | 0.04  | -0.03 | 0.11 | 0.10 | 0.11 | 0.33 | 0.32 | 0.33 |
| 0.5      | 0.01  | 0.04  | 0.00  | 0.10 | 0.09 | 0.12 | 0.32 | 0.31 | 0.34 |
| 1        | -0.01 | -0.01 | 0.02  | 0.12 | 0.10 | 0.14 | 0.34 | 0.31 | 0.37 |
| 1.5      | 0.05  | -0.01 | 0.14  | 0.14 | 0.12 | 0.19 | 0.38 | 0.34 | 0.46 |
| 2        | 0.08  | -0.02 | 0.28  | 0.19 | 0.15 | 0.29 | 0.44 | 0.39 | 0.61 |

\*Represents the type of item parameter used. True=true item parameters, 3PLM=parameter estimates from 3PLM, and 2PLM=parameter estimates from 2PLM.

Table 5.6: Average  $\hat{\theta}$  Inflation ( $\hat{\theta} - \theta$ ) under Test Compromise Conditions

| $\theta$ | %   | True Item Parameters |      |      | 3PLM Estimates |      |      | 2PLM Estimates |      |      |
|----------|-----|----------------------|------|------|----------------|------|------|----------------|------|------|
|          |     | S1*                  | S2*  | S12* | S1             | S2   | S12  | S1             | S2   | S12  |
| -2       | 0.1 | 0.57                 | 0.52 | 0.50 | 0.52           | 0.45 | 0.43 | 0.48           | 0.52 | 0.38 |
| -1.5     |     | 0.50                 | 0.44 | 0.44 | 0.54           | 0.46 | 0.47 | 0.40           | 0.41 | 0.30 |
| -1       |     | 0.44                 | 0.38 | 0.38 | 0.50           | 0.44 | 0.44 | 0.33           | 0.32 | 0.23 |
| -0.5     |     | 0.40                 | 0.36 | 0.36 | 0.46           | 0.42 | 0.42 | 0.29           | 0.31 | 0.20 |
| 0        |     | 0.36                 | 0.32 | 0.32 | 0.42           | 0.37 | 0.38 | 0.27           | 0.27 | 0.17 |
| 0.5      |     | 0.35                 | 0.33 | 0.32 | 0.37           | 0.35 | 0.34 | 0.31           | 0.31 | 0.21 |
| 1        |     | 0.31                 | 0.31 | 0.29 | 0.29           | 0.28 | 0.27 | 0.33           | 0.33 | 0.22 |
| 1.5      |     | 0.36                 | 0.36 | 0.33 | 0.29           | 0.28 | 0.26 | 0.47           | 0.48 | 0.35 |
| 2        |     | 0.39                 | 0.39 | 0.35 | 0.26           | 0.26 | 0.23 | 0.61           | 0.63 | 0.49 |
| -2       | 0.2 | 1.19                 | 1.00 | 1.06 | 1.27           | 1.06 | 1.14 | 0.97           | 0.94 | 0.90 |
| -1.5     |     | 1.02                 | 0.85 | 0.93 | 1.11           | 0.93 | 1.03 | 0.83           | 0.78 | 0.77 |
| -1       |     | 0.89                 | 0.71 | 0.84 | 0.98           | 0.80 | 0.93 | 0.72           | 0.65 | 0.68 |
| -0.5     |     | 0.77                 | 0.67 | 0.77 | 0.83           | 0.72 | 0.85 | 0.64           | 0.60 | 0.63 |
| 0        |     | 0.67                 | 0.60 | 0.73 | 0.71           | 0.63 | 0.78 | 0.58           | 0.55 | 0.61 |
| 0.5      |     | 0.63                 | 0.61 | 0.72 | 0.63           | 0.59 | 0.73 | 0.60           | 0.59 | 0.67 |
| 1        |     | 0.55                 | 0.57 | 0.68 | 0.51           | 0.51 | 0.64 | 0.59           | 0.62 | 0.70 |
| 1.5      |     | 0.56                 | 0.62 | 0.76 | 0.47           | 0.50 | 0.66 | 0.70           | 0.78 | 0.90 |
| 2        |     | 0.58                 | 0.64 | 0.78 | 0.44           | 0.47 | 0.63 | 0.85           | 0.94 | 1.06 |
| -2       | 0.4 | 2.14                 | 1.73 | 2.04 | 2.24           | 1.81 | 2.15 | 2.00           | 1.65 | 1.73 |
| -1.5     |     | 1.77                 | 1.45 | 1.84 | 1.86           | 1.53 | 1.93 | 1.66           | 1.40 | 1.57 |
| -1       |     | 1.50                 | 1.23 | 1.70 | 1.57           | 1.30 | 1.76 | 1.41           | 1.19 | 1.46 |
| -0.5     |     | 1.24                 | 1.12 | 1.56 | 1.29           | 1.15 | 1.57 | 1.19           | 1.09 | 1.39 |
| 0        |     | 1.02                 | 1.02 | 1.46 | 1.04           | 1.02 | 1.43 | 1.00           | 1.02 | 1.36 |
| 0.5      |     | 0.89                 | 1.00 | 1.44 | 0.87           | 0.95 | 1.37 | 0.92           | 1.06 | 1.44 |
| 1        |     | 0.76                 | 0.93 | 1.39 | 0.71           | 0.83 | 1.27 | 0.87           | 1.07 | 1.50 |
| 1.5      |     | 0.73                 | 0.94 | 1.48 | 0.62           | 0.78 | 1.29 | 0.93           | 1.20 | 1.74 |
| 2        |     | 0.72                 | 0.93 | 1.40 | 0.56           | 0.72 | 1.18 | 1.05           | 1.35 | 1.81 |

\*Represents the test compromise stage. S1=stage 1 only, S2=stage 2 only, S12=stage 1 and 2

Results on the  $\hat{\theta}$  inflation show that even with 10% compromised responses,  $\hat{\theta}$  could be inflated for about one standard error, and 20% or 40% compromise could lead to inflation up to two to four standard errors at all  $\theta$  levels. Using item parameter estimates from the 3PLM tends to result in slightly larger inflation for lower  $\theta$  levels, and smaller inflation for higher  $\theta$  levels compared to using true item parameters, while using item parameter estimates from the 2PLM demonstrates a reversed pattern. Using item

parameter estimates does not lead to a large difference in  $\hat{\theta}$  inflation for medium  $\theta$  levels from using true item parameters, but it could lead to a difference as large as 0.2 to 0.4 for  $\theta$  at the higher or lower end.

Regarding the comparisons of  $\hat{\theta}$  inflation across different compromise stages, when the compromise rate is 10%, the comparison does not show a consistent pattern across different  $\theta$  levels or across the use of different item parameters. As the compromise rate gets to 20% or higher, more consistent patterns are observed. Specifically, when the compromise rate is 20%, the preknowledge on both stages results in the largest score inflation for  $\theta \geq -0.5$ , and the preknowledge in stage 1 only results in the largest score inflation for  $\theta < -0.5$ . When the compromise rate is 40%, the preknowledge at both stages results in the largest score inflation for most  $\theta$  levels. The different patterns at different compromise rates are related to the routing error. For instance, with the compromise rate being 0.2, when test compromise occurs at both stages, for low  $\theta$  levels, the  $\hat{\theta}$  inflation due to the compromised responses in stage 1 may not be large enough to change the routing. Therefore, low-ability examinees will still be routed to the easy module in stage 2, and the compromised items for them will be easier than those if the preknowledge occurs at stage 1 only. Aberrantly correct response on easier items are likely to result in smaller inflation in  $\hat{\theta}$ . In comparison, for high  $\theta$  levels, especially for those near the routing threshold, when the preknowledge occurs at both stages, they are very likely to be routed to a harder module. Therefore, they will have correct responses on harder items than if the preknowledge only occurs in stage 1, which is likely to result in larger inflation in  $\hat{\theta}$ . Similarly, when the compromise rate increases to 40%, when the preknowledge occurs at both stages, the inflation in  $\hat{\theta}$  with stage-1

compromise responses may be large enough for most  $\theta$  levels to be routed to a harder module, so it will lead to aberrantly correct responses on harder items than if the preknowledge only occurs in stage 1, and thus result in larger inflation in  $\hat{\theta}$ .

When the preknowledge occurs at one stage only, with the compromise rate being 20% or higher, the preknowledge on stage 1 results in larger  $\hat{\theta}$  inflation at low or medium  $\theta$  levels ( $\theta \leq 0.5$ ) and the preknowledge at stage 2 results in larger inflation for higher  $\theta$  levels ( $\theta > 0.5$ ). This is because with the item parameters used in this study and with the items to which preknowledge is simulated, for low or medium  $\theta$  levels, stage-1 compromised items are harder than stage-2 compromised items, while for high  $\theta$  levels, stage-2 compromised items are harder than those at stage-1.

### **5.5.2 Detection rate at the person-level**

The empirical type-I error rate at individual  $\theta$  levels are presented in Figure 5.1. The nominal level for the type-I error rate is 0.05, and a 95% normal-approximation confidence interval for the empirical type-I error rate out of 500 replications is (0.031, 0.069). Based on this criterion, the empirical type-I error rates for the predictive checking approach and the KL divergence are conservative at the lower end of the  $\theta$  continuum when item parameters estimates from the 3PLM are used, but when true item parameters or item parameter estimates from the 2PLM are used, the empirical type-I error fall in the 95% confidence interval at most  $\theta$  levels. In comparison, the likelihood ratio test has higher type-I error rates. The type-I error of the likelihood ratio test exceeds the upper bound of the nominal level at  $\theta=2$  when true item parameters are used, but it falls in the 95% confidence interval at all  $\theta$  levels when item parameter estimates are used.

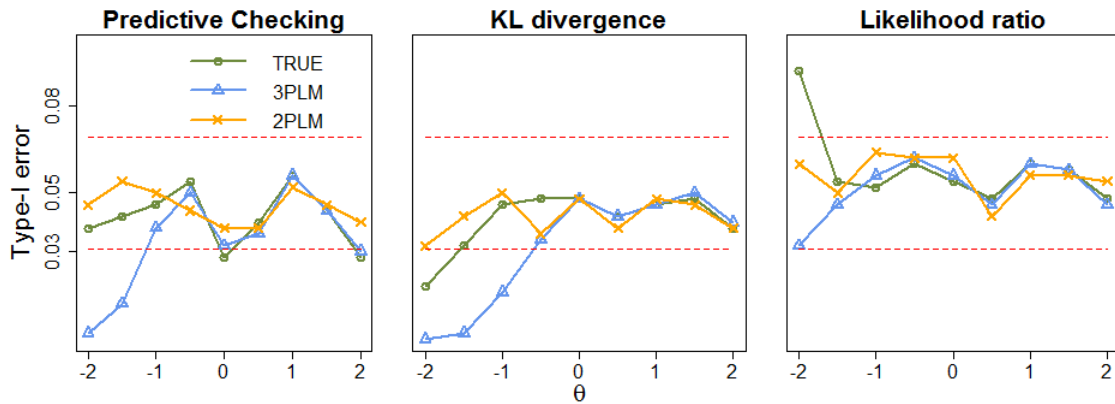


Figure 5.1: Type-I error rate at person level. The two dotted lines represent the upper and lower bound of the 95% normal-approximation confidence interval for the empirical type-I error rate at nominal level of 0.05. “True” represents using true item parameters, 3PLM and 2PLM represent using item parameter estimates from 3PLM and 2PLM respectively.

Table 5.7 to Table 5.9 summarize the person-level power with true item parameters, and item parameter estimates from different models.

Table 5.7: Power with True Item Parameters

|      |     | Predictive Checking |       |       | KL    |       |       | Likelihood Ratio |       |       |
|------|-----|---------------------|-------|-------|-------|-------|-------|------------------|-------|-------|
|      | %   | S1                  | S2    | S12   | S1    | S2    | S12   | S1               | S2    | S12   |
| -2   | 0.1 | 0.162               | 0.140 | 0.126 | 0.112 | 0.120 | 0.026 | 0.220            | 0.220 | 0.126 |
| -1.5 |     | 0.182               | 0.148 | 0.160 | 0.094 | 0.116 | 0.054 | 0.122            | 0.140 | 0.094 |
| -1   |     | 0.164               | 0.152 | 0.140 | 0.132 | 0.114 | 0.060 | 0.158            | 0.130 | 0.098 |
| -0.5 |     | 0.162               | 0.132 | 0.138 | 0.138 | 0.116 | 0.084 | 0.166            | 0.142 | 0.118 |
| 0    |     | 0.126               | 0.112 | 0.124 | 0.114 | 0.114 | 0.072 | 0.144            | 0.140 | 0.104 |
| 0.5  |     | 0.132               | 0.116 | 0.116 | 0.108 | 0.088 | 0.072 | 0.132            | 0.102 | 0.094 |
| 1    |     | 0.114               | 0.112 | 0.106 | 0.130 | 0.112 | 0.104 | 0.148            | 0.128 | 0.122 |
| 1.5  |     | 0.082               | 0.096 | 0.084 | 0.096 | 0.100 | 0.082 | 0.118            | 0.128 | 0.102 |
| 2    |     | 0.070               | 0.070 | 0.054 | 0.078 | 0.082 | 0.062 | 0.110            | 0.108 | 0.096 |
| -2   | 0.2 | 0.484               | 0.398 | 0.440 | 0.360 | 0.348 | 0.324 | 0.438            | 0.416 | 0.398 |
| -1.5 |     | 0.472               | 0.336 | 0.384 | 0.320 | 0.298 | 0.296 | 0.394            | 0.356 | 0.366 |
| -1   |     | 0.426               | 0.298 | 0.386 | 0.332 | 0.258 | 0.272 | 0.384            | 0.306 | 0.324 |
| -0.5 |     | 0.330               | 0.298 | 0.328 | 0.276 | 0.246 | 0.266 | 0.330            | 0.294 | 0.310 |
| 0    |     | 0.254               | 0.230 | 0.306 | 0.248 | 0.238 | 0.260 | 0.292            | 0.280 | 0.296 |
| 0.5  |     | 0.250               | 0.224 | 0.300 | 0.202 | 0.186 | 0.250 | 0.230            | 0.222 | 0.292 |
| 1    |     | 0.188               | 0.196 | 0.230 | 0.186 | 0.178 | 0.224 | 0.206            | 0.204 | 0.268 |
| 1.5  |     | 0.144               | 0.180 | 0.220 | 0.160 | 0.168 | 0.202 | 0.198            | 0.204 | 0.244 |
| 2    |     | 0.092               | 0.118 | 0.146 | 0.106 | 0.124 | 0.164 | 0.160            | 0.172 | 0.222 |
| -2   | 0.4 | 0.960               | 0.840 | 0.940 | 0.936 | 0.824 | 0.860 | 0.958            | 0.864 | 0.890 |
| -1.5 |     | 0.878               | 0.718 | 0.906 | 0.860 | 0.722 | 0.812 | 0.896            | 0.770 | 0.844 |
| -1   |     | 0.828               | 0.638 | 0.888 | 0.750 | 0.610 | 0.776 | 0.786            | 0.668 | 0.824 |
| -0.5 |     | 0.664               | 0.590 | 0.860 | 0.650 | 0.522 | 0.778 | 0.704            | 0.602 | 0.826 |
| 0    |     | 0.498               | 0.494 | 0.754 | 0.510 | 0.472 | 0.720 | 0.546            | 0.502 | 0.776 |
| 0.5  |     | 0.378               | 0.446 | 0.704 | 0.356 | 0.424 | 0.636 | 0.384            | 0.480 | 0.700 |
| 1    |     | 0.264               | 0.352 | 0.558 | 0.266 | 0.344 | 0.562 | 0.314            | 0.392 | 0.632 |
| 1.5  |     | 0.188               | 0.286 | 0.488 | 0.204 | 0.254 | 0.444 | 0.234            | 0.304 | 0.552 |
| 2    |     | 0.118               | 0.198 | 0.354 | 0.144 | 0.190 | 0.288 | 0.206            | 0.258 | 0.440 |

Table 5.8: Power with Item Parameter Estimates from 3PLM

|      | %   | Predictive Checking |              |              | KL           |              |              | Likelihood Ratio |              |              |
|------|-----|---------------------|--------------|--------------|--------------|--------------|--------------|------------------|--------------|--------------|
|      |     | S1                  | S2           | S12          | S1           | S2           | S12          | S1               | S2           | S12          |
| -2   | 0.1 | 0.070               | <b>0.060</b> | 0.062        | <b>0.006</b> | <b>0.000</b> | 0.000        | 0.076            | 0.104        | 0.052        |
| -1.5 |     | 0.138               | 0.112        | 0.116        | 0.016        | 0.024        | 0.004        | 0.100            | 0.132        | 0.072        |
| -1   |     | 0.144               | 0.144        | 0.144        | 0.062        | 0.072        | 0.026        | 0.146            | 0.134        | 0.096        |
| -0.5 |     | 0.168               | 0.130        | 0.136        | 0.108        | 0.102        | 0.072        | 0.156            | 0.148        | 0.114        |
| 0    |     | 0.126               | 0.110        | 0.114        | 0.104        | 0.114        | 0.076        | 0.136            | 0.138        | 0.108        |
| 0.5  |     | 0.144               | 0.114        | 0.112        | 0.104        | 0.094        | 0.078        | 0.130            | 0.102        | 0.096        |
| 1    |     | 0.106               | 0.114        | 0.108        | 0.126        | 0.114        | 0.108        | 0.152            | 0.130        | 0.124        |
| 1.5  |     | 0.080               | 0.096        | 0.090        | 0.096        | 0.102        | 0.080        | 0.122            | 0.124        | 0.106        |
| 2    |     | 0.070               | 0.074        | 0.062        | 0.086        | 0.082        | 0.068        | 0.110            | 0.110        | 0.092        |
| -2   | 0.2 | 0.408               | <b>0.280</b> | <b>0.328</b> | <b>0.050</b> | <b>0.040</b> | <b>0.024</b> | <b>0.280</b>     | <b>0.314</b> | <b>0.224</b> |
| -1.5 |     | 0.428               | 0.314        | 0.402        | <b>0.108</b> | <b>0.140</b> | <b>0.080</b> | 0.312            | 0.330        | <b>0.254</b> |
| -1   |     | 0.424               | 0.300        | 0.404        | <b>0.198</b> | 0.200        | 0.190        | 0.336            | 0.302        | 0.288        |
| -0.5 |     | 0.332               | 0.306        | 0.342        | 0.244        | 0.240        | 0.232        | 0.316            | 0.306        | 0.312        |
| 0    |     | 0.246               | 0.226        | 0.294        | 0.236        | 0.236        | 0.244        | 0.288            | 0.276        | 0.302        |
| 0.5  |     | 0.242               | 0.226        | 0.298        | 0.204        | 0.184        | 0.244        | 0.228            | 0.220        | 0.288        |
| 1    |     | 0.186               | 0.200        | 0.226        | 0.186        | 0.184        | 0.234        | 0.212            | 0.204        | 0.266        |
| 1.5  |     | 0.138               | 0.170        | 0.220        | 0.162        | 0.166        | 0.204        | 0.204            | 0.198        | 0.250        |
| 2    |     | 0.094               | 0.118        | 0.146        | 0.108        | 0.124        | 0.172        | 0.152            | 0.174        | 0.220        |
| -2   | 0.4 | 0.956               | 0.838        | 0.938        | 0.856        | <b>0.644</b> | <b>0.504</b> | 0.950            | 0.848        | 0.818        |
| -1.5 |     | 0.882               | 0.732        | 0.910        | 0.806        | 0.648        | <b>0.668</b> | 0.888            | 0.768        | 0.814        |
| -1   |     | 0.828               | 0.668        | 0.894        | 0.726        | 0.602        | 0.710        | 0.786            | 0.696        | 0.794        |
| -0.5 |     | 0.666               | 0.598        | 0.842        | 0.646        | 0.534        | 0.762        | 0.700            | 0.624        | 0.828        |
| 0    |     | 0.494               | 0.504        | 0.744        | 0.516        | 0.486        | 0.710        | 0.552            | 0.520        | 0.774        |
| 0.5  |     | 0.364               | 0.442        | 0.690        | 0.362        | 0.432        | 0.646        | 0.390            | 0.478        | 0.688        |
| 1    |     | 0.266               | 0.344        | 0.550        | 0.276        | 0.330        | 0.566        | 0.324            | 0.378        | 0.636        |
| 1.5  |     | 0.192               | 0.288        | 0.480        | 0.210        | 0.246        | 0.456        | 0.238            | 0.298        | 0.552        |
| 2    |     | 0.114               | 0.194        | 0.338        | 0.152        | 0.180        | 0.296        | 0.204            | 0.252        | 0.438        |

*Note.* The italic bold numbers represent the power rates that are different from the power with true item parameters by 0.1.

Table 5.9: Power with Item Parameter Estimates from 2PLM

|      | %   | Predictive Checking |       |       | KL    |       |       | Likelihood Ratio |       |       |
|------|-----|---------------------|-------|-------|-------|-------|-------|------------------|-------|-------|
|      |     | S1                  | S2    | S12   | S1    | S2    | S12   | S1               | S2    | S12   |
| -2   | 0.1 | 0.178               | 0.142 | 0.148 | 0.150 | 0.174 | 0.110 | 0.184            | 0.208 | 0.144 |
| -1.5 |     | 0.190               | 0.166 | 0.156 | 0.106 | 0.134 | 0.090 | 0.138            | 0.154 | 0.104 |
| -1   |     | 0.174               | 0.132 | 0.148 | 0.124 | 0.120 | 0.104 | 0.158            | 0.144 | 0.120 |
| -0.5 |     | 0.160               | 0.130 | 0.136 | 0.130 | 0.106 | 0.100 | 0.156            | 0.138 | 0.122 |
| 0    |     | 0.126               | 0.108 | 0.108 | 0.098 | 0.096 | 0.072 | 0.132            | 0.126 | 0.096 |
| 0.5  |     | 0.130               | 0.118 | 0.110 | 0.094 | 0.084 | 0.066 | 0.118            | 0.096 | 0.082 |
| 1    |     | 0.106               | 0.116 | 0.112 | 0.106 | 0.094 | 0.078 | 0.134            | 0.134 | 0.104 |
| 1.5  |     | 0.090               | 0.098 | 0.100 | 0.084 | 0.092 | 0.068 | 0.116            | 0.124 | 0.084 |
| 2    |     | 0.082               | 0.088 | 0.064 | 0.068 | 0.068 | 0.056 | 0.116            | 0.110 | 0.094 |
| -2   | 0.2 | 0.490               | 0.390 | 0.418 | 0.392 | 0.392 | 0.360 | 0.438            | 0.430 | 0.404 |
| -1.5 |     | 0.442               | 0.346 | 0.410 | 0.332 | 0.334 | 0.334 | 0.396            | 0.396 | 0.392 |
| -1   |     | 0.408               | 0.302 | 0.404 | 0.324 | 0.258 | 0.276 | 0.372            | 0.318 | 0.330 |
| -0.5 |     | 0.318               | 0.290 | 0.322 | 0.274 | 0.242 | 0.244 | 0.312            | 0.302 | 0.304 |
| 0    |     | 0.246               | 0.218 | 0.286 | 0.226 | 0.192 | 0.230 | 0.268            | 0.260 | 0.274 |
| 0.5  |     | 0.240               | 0.228 | 0.288 | 0.168 | 0.162 | 0.196 | 0.216            | 0.208 | 0.262 |
| 1    |     | 0.196               | 0.208 | 0.232 | 0.168 | 0.166 | 0.196 | 0.206            | 0.192 | 0.240 |
| 1.5  |     | 0.154               | 0.190 | 0.238 | 0.142 | 0.160 | 0.174 | 0.184            | 0.192 | 0.234 |
| 2    |     | 0.120               | 0.128 | 0.176 | 0.092 | 0.100 | 0.134 | 0.164            | 0.174 | 0.222 |
| -2   | 0.4 | 0.958               | 0.838 | 0.904 | 0.934 | 0.812 | 0.838 | 0.952            | 0.844 | 0.884 |
| -1.5 |     | 0.872               | 0.734 | 0.892 | 0.858 | 0.724 | 0.770 | 0.886            | 0.760 | 0.828 |
| -1   |     | 0.824               | 0.650 | 0.882 | 0.730 | 0.594 | 0.736 | 0.772            | 0.658 | 0.776 |
| -0.5 |     | 0.664               | 0.586 | 0.854 | 0.632 | 0.504 | 0.746 | 0.680            | 0.588 | 0.774 |
| 0    |     | 0.494               | 0.488 | 0.744 | 0.496 | 0.428 | 0.662 | 0.536            | 0.502 | 0.706 |
| 0.5  |     | 0.374               | 0.448 | 0.704 | 0.320 | 0.378 | 0.552 | 0.386            | 0.460 | 0.626 |
| 1    |     | 0.288               | 0.364 | 0.578 | 0.252 | 0.318 | 0.498 | 0.306            | 0.374 | 0.572 |
| 1.5  |     | 0.214               | 0.322 | 0.512 | 0.186 | 0.218 | 0.400 | 0.236            | 0.308 | 0.518 |
| 2    |     | 0.150               | 0.222 | 0.394 | 0.134 | 0.170 | 0.238 | 0.212            | 0.282 | 0.450 |

The power for all three methods is small when the compromise rate is only 10% - the power among the lowest  $\theta$  levels is less than 0.2 for the predictive checking approach, and less than 0.1 for the other two approaches. The power increases to a large extent as the compromise rate increases to 40% - moderate to high power rates are observed among medium to low  $\theta$  levels. The three methods demonstrate similar power when true item parameters and item parameter estimates from 2PLM are used: the power difference between the predictive checking and the likelihood ratio test is smaller than 0.1 and both

of them tend to have slightly higher power than the KL divergence at lower  $\theta$  levels. When item parameter estimates from 3PLM are used, using the predictive checking could have slightly larger power at lower  $\theta$  levels than the likelihood ratio test, and both of them have much higher power at lower  $\theta$  levels than the KL divergence. As for the power comparisons across different compromised stages for each method, when the compromise rate is 20% or higher, the power comparisons demonstrate very similar patterns as the score inflation across different compromised stages for the predictive checking approach. This is consistent with the expectation that a larger score inflation is easier to be detected. As for the KL divergence and the likelihood ratio test, the power comparison patterns become similar to the pattern of score inflation when the compromise rate is 40%. The lack of similarity between the power pattern and the score inflation pattern when the compromise rate is small could be attributed to the fact that both methods measure more than the shift in the point estimate of  $\theta$ : the KL divergence measures the discrepancy between the  $\theta$  posterior distributions on the two types of items, and the likelihood ratio test measures the difference in likelihoods under null and alternative hypotheses. When the compromise rate is small, the value of the KL divergence statistic or the likelihood ratio statistic does not simply reflect the shift of  $\hat{\theta}$ , but when the compromise rate gets larger, the large value of either statistic is dominated by the shift of  $\hat{\theta}$ .

In terms of the effect of using item parameter estimates on the detection power, as can be seen from Table 5.8 and 5.9, the power differences between using the true item parameters and the 2PLM item parameter estimates are very small at most  $\theta$  levels: all differences are up to the second decimal place. Using 3PLM item parameter estimates could result in power reduction for greater than 0.1 at the lowest one or two  $\theta$  levels ( $\theta=-$

2 or -1.5) for all three methods, especially for the KL divergence: that the power reduction for the KL divergence could be as large as 0.2 to 0.5 when the 3PLM item parameter estimates are used. This effect is likely to be caused by the fact that the presence of the  $c$ -parameter reduces the Fisher information, and thus the likelihood function becomes flatter in the 3PLM. Particularly, for the KL divergence method, as illustrated from Figure 5.2, the presence of the  $c$ -parameter makes the  $\theta$  posterior distributions flatter than those estimated from the true item parameters or 2PLM item parameter estimates, and the log posterior ratios between the two posterior distributions ( $\ln \frac{P(\theta|y_1)}{P(\theta|y_2)}$ ), which are part of the integrand in the KL divergence calculation, also become much smaller. These two outcomes together result in smaller values for the KL divergence, and thus the power is reduced with the 3PLM item parameter estimates.

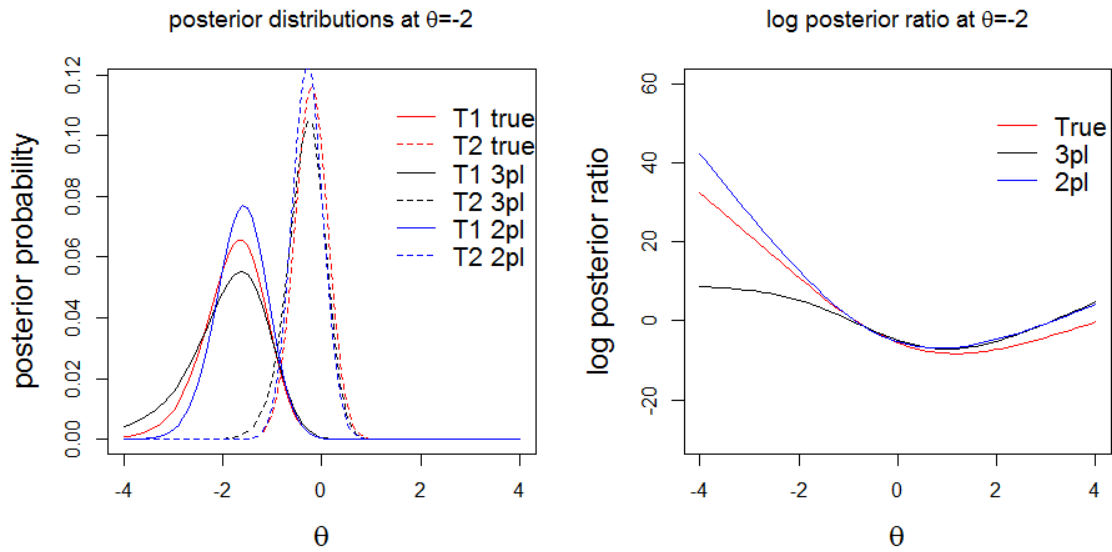


Figure 5.2: The left panel shows the  $\theta$  posterior distributions on T1 and T2 in the condition with 40% compromised responses in both stages. The red, black and blue lines represent the use of true item parameters, 3PLM item parameter estimates and 2PLM item parameter estimates. The right panel shows the log posterior ratio between T1 and T2 when the three types of item parameters are used.

### 5.5.3 Detection rate at the group level

Figure 5.3 displays the type-I error at group level, and Figures 5.4 and 5.5 display the detection power among examinees with ability distributions of  $N(0,1)$  and  $N(-1,1)$ , respectively. Figure 5.3 shows that the empirical type-I error for the KL divergence is slightly conservative among  $N(-1,1)$ , while the empirical type-I error rates for the other two methods both fall into the 95% normal approximation confidence interval. Same as the pattern at the person-level, the KL divergence has the lowest type-I error rate, due to the use of the most conservative cutoff value, and the likelihood ratio test has the highest type-I error rate. Figures 5.4 and 5.5 show that the detection power among both ability groups is low when there is a small compromise rate – the detection power is below 0.4 for both groups when the compromise rate is 10% or 20%. With 40% compromised

items, moderate to high detection power is observed. Consistent with the detection rate at individual  $\theta$  levels, the predictive checking and the likelihood ratio test have very similar power, and the KL divergence has the lowest power. When the compromise rate is less than 40%, the difference in the detection power among different MST stages is small for both ability groups. With 40% compromise rate, the detection power is slightly higher when the preknowledge occurs in both stages, due to the fact that preknowledge at both stages results in larger  $\hat{\theta}$  inflation among most  $\theta$  levels.

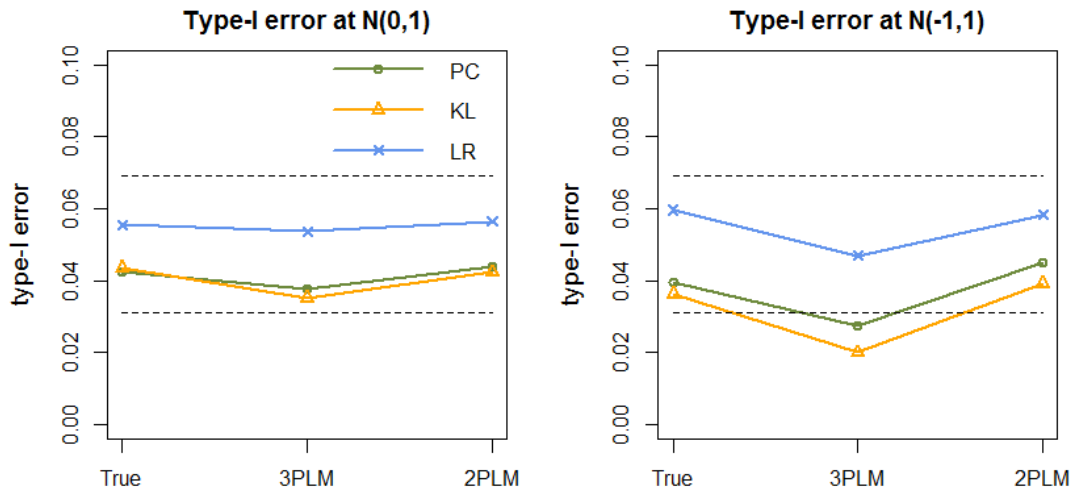


Figure 5.3: Type-I error rate of the three methods (PC=Predictive checking, KL=KL divergence, and LR=likelihood ratio) among different examinee ability groups. The x-axis represents the type of item parameter used, and the two dotted lines represent the upper and lower bound of the 95% normal-approximation confidence interval for the empirical type-I error rate at nominal level of 0.05.

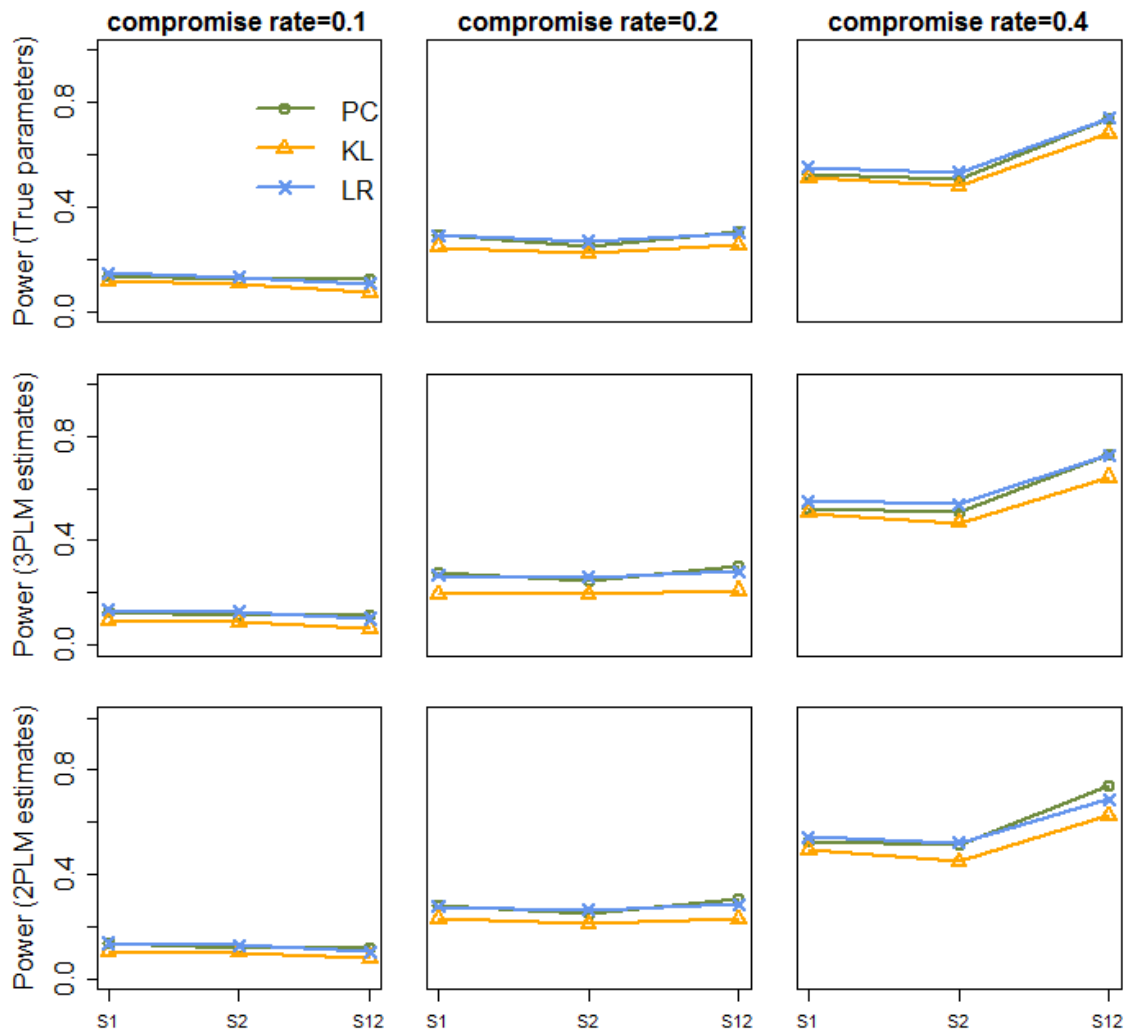


Figure 5.4: Detection power among examinees with ability distribution of  $N(0,1)$ . Each plot shows the detection rates of the three methods at different compromised stages (S1=stage 1 only, S2=stage 2 only, S12=Stage 1 and 2). The first row shows the detection power with true item parameters, the second row shows the power with 3PLM item parameter estimates, and the third row shows the power with 2PLM item parameter estimates.

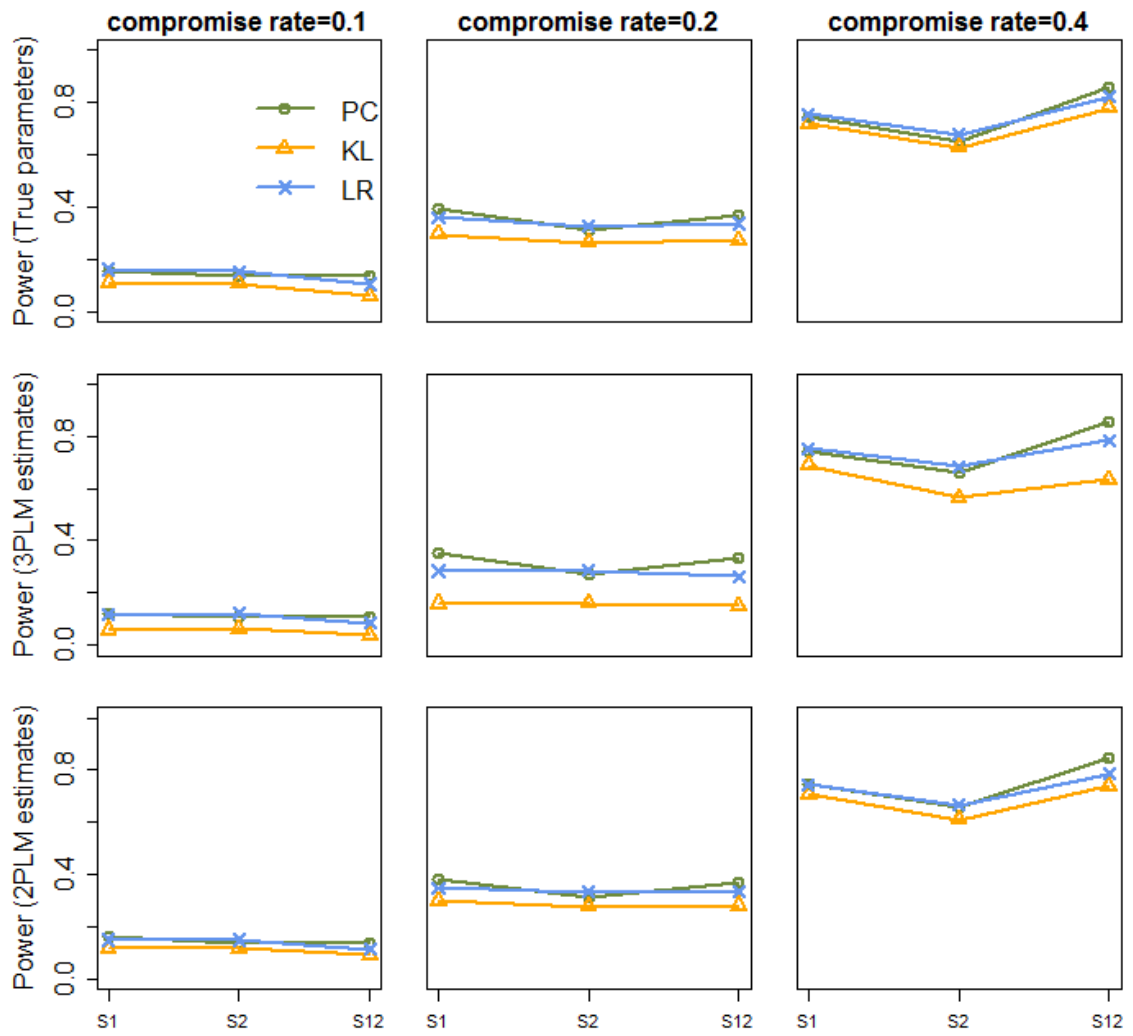


Figure 5.5: power among examinees with ability distribution of  $N(-1,1)$ .

## 5.6 Results in Key Exposure Simulation

### 5.6.1 Person-level Detection Result

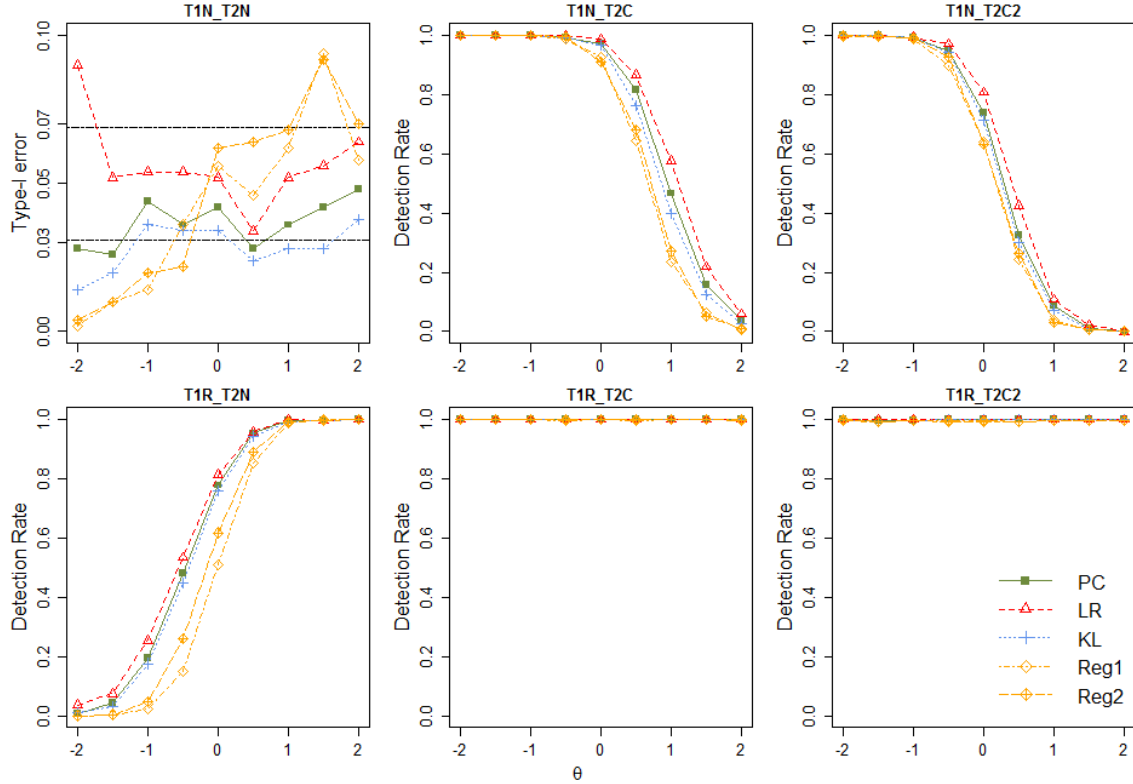


Figure 5.6: Person-level detection rate across different conditions. Reg1 represents the regression method when the cheating group follows  $\theta \sim N(0,1)$ , and Reg2 represents the regression method when the cheating group follows  $\theta \sim N(-1,1)$ .

The detection rates of different methods in the six conditions are summarized in Figure 5.6. Regarding the type-I error of different methods, the type-I error of the predictive checking method is slightly conservative at lower  $\theta$  levels, and the KL divergence has conservative type-I error at many  $\theta$  levels, due to the use of the most conservative cut-off value. Similar to the findings in the shallow pool situation, the likelihood ratio test has larger type-I error than the predictive checking and the KL divergence, and it has inflated type-I error at  $\theta = -2$ . The type-I error of the simple linear regression shows an increasing pattern as  $\theta$  increases. This is due to the fact that the

residuals tend to be smaller at lower  $\theta$  levels and larger at higher  $\theta$  levels, as demonstrated in Figure 5.7.

As for the detection power when T1 responses are based on one's real proficiency but the entire T2 is compromised, all methods have large power to detect security breach on T2 for  $\theta \leq 0$ . When responses to T2 are identical with the keys (i.e. no memorization error), the three IRT-based methods have larger power than the regression method: at relatively high  $\theta$  levels ( $\theta = 0.5$  or  $1$ ), the power of the predictive checking and the likelihood ratio test is about 0.2 higher than the regression method, and the power of the KL divergence is about 0.1 higher than the regression method. The comparison among the three IRT-based methods shows that the likelihood ratio test has the highest power, and the KL divergence has the lowest power. When responses to T2 contain 20% memorization error, all methods only have moderate or low power to detect preknowledge among medium to high  $\theta$  levels ( $\theta \geq 0.5$ ). Under this condition, the power comparison among all four methods shows the same pattern as in the condition where responses to T2 contain no memorization errors, but the difference between methods is much smaller.

When responses to T1 are random or based on the wrong key, all methods have high power to detect score difference between T1 and T2. On one hand, when T2 is uncompromised, all methods have a very high chance of falsely identifying a high-ability person as a cheater. On the other hand, when the entire T2 is compromised, all methods will correctly identify a cheater with the probability of 1 for all  $\theta$  levels.

It is also seen that although the regression is a group-based method and the regression line could be affected to different extent when the cheating group follows

different ability distributions, in this study, the person-level detection rate in the regression method is very similar when the cheating group follows different ability distributions. This is mainly due to the fact that there is not a large proportion of cheating examinees in this study, so the impact of outliers on the regression line is very small.

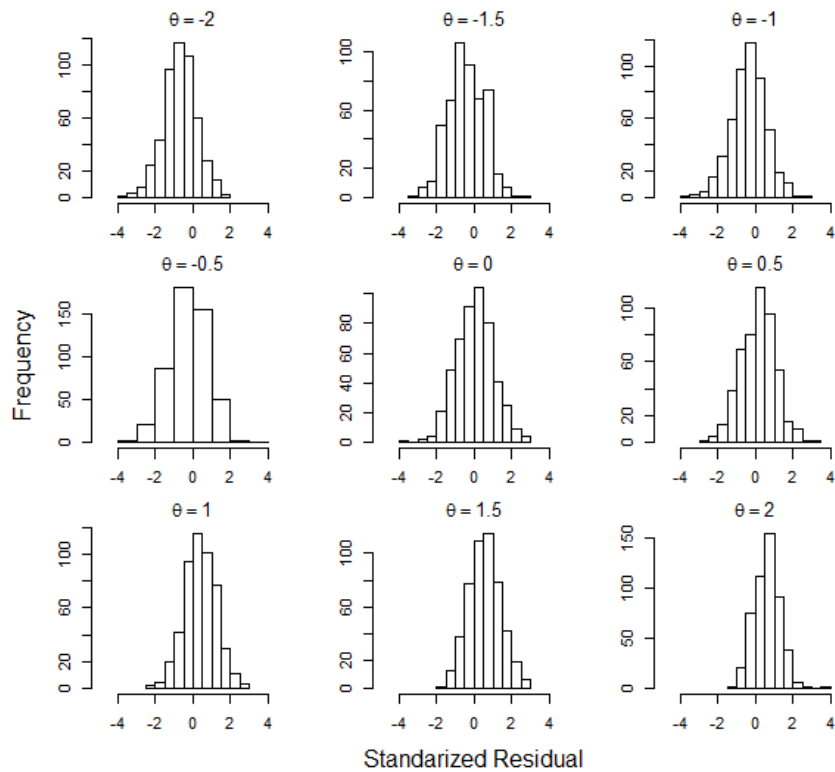


Figure 5.7: Distribution of standardized residuals at different  $\theta$  levels.

## 5.6.2 Group-level Detection Results

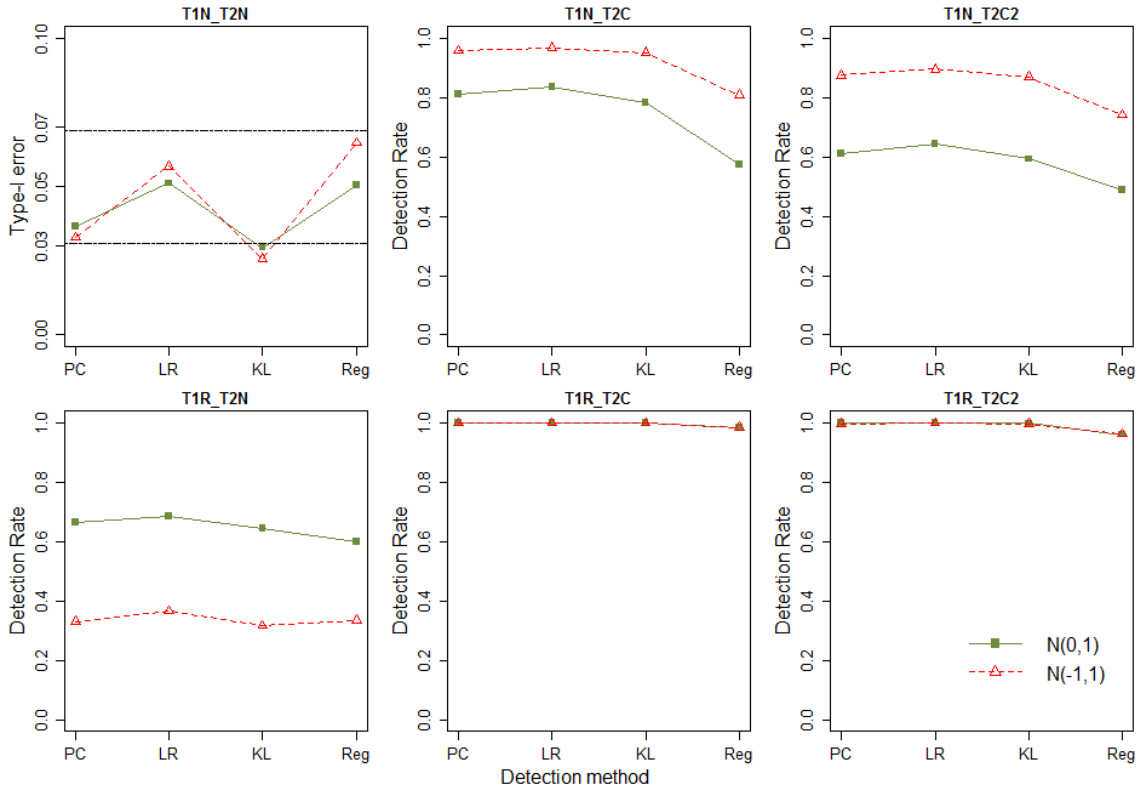


Figure 5.8: Group-level detection rate across different conditions.

Regarding the detection rate at the group level, the type-I error rate for all methods fall in the 95% normal approximation confidence interval for both ability groups. The likelihood ratio test and the regression method have larger type-I error rates than the other two methods. Same as person-level detection results, when T1 responses are based on one's real proficiency, all methods have high power to detect security breach: the power is above 0.6 for the group with  $\theta \sim N(0,1)$ , and above 0.8 for the group with  $\theta \sim N(-1,1)$ . When responses to T1 are random or based on the wrong key, if security breach happens, all methods will correctly detect it with a probability of almost 1, but they will also falsely detect it with a high probability if the security breach does not happen.

## 5.7. Discussion

Study 2 mimicked two test compromise situations that are likely to happen in practice, and compared the predictive checking method with several other approaches. In the shallow pool situation with small to moderate test compromise rates, the likelihood ratio test and the adapted KL divergence were compared to the predictive checking in an MST design. The results suggest that the predictive checking and the likelihood ratio test have very similar power in all conditions, and both of them slightly outperformed the adapted KL divergence. The slightly lower power of the adapted KL divergence could be partly attributed to the fact that the most conservative sampling distribution is used to ensure the type-I error rate does not exceed the nominal level in the null space of  $\theta$ . Both the predictive checking method and the adapted KL divergence method use simulation to approximate the exact distribution of the test statistic in the null condition, while the likelihood ratio test relies on an asymptotic distribution. The results in this study suggest that using the asymptotic distribution does not lead to either too conservative or too liberal type-I error rate at the nominal level of 0.05 for most  $\theta$  levels, but the type-I error of the likelihood ratio test is slightly larger than the other two methods. Considering the likelihood ratio test is less computationally intensive than the other two methods, one can consider using the likelihood ratio test if both subtests (T1 and T2) contain more than 20 items in practice. However, it should be noted that the asymptotic distribution for the likelihood ratio test may not work well when either subtest is not long enough, and in that case, the predictive checking method is a useful alternative, as it does not need an asymptotic distribution.

In the shallow pool simulation, the detection effectiveness was also investigated under different test compromise rates, different compromised MST stages, and different types of item parameters. The findings suggest that none of the three methods has sufficient power to detect preknowledge on only 10% items, even though 10% test compromise rate could result in  $\hat{\theta}$  inflation of one standard error. The predictive checking method and the likelihood ratio test have moderate power to detect 20% test compromise among low-ability examinees, and both of them have moderate to large power to detect 40% test compromise among low to medium-ability examinees. Preknowledge at different stages will result in different amount of  $\hat{\theta}$  inflation, depending on the difficulty of the compromised items relative to a person's ability, and in general, the detection power increases with larger  $\hat{\theta}$  inflation. Lastly, regarding the use of different types of item parameters, the power of the KL divergence at the lower end of  $\theta$  is greatly reduced by the use of the 3PLM item parameter estimates, possibly due to the impact of  $c$ -parameter on the likelihood function and the  $\theta$  posterior distribution. The predictive checking method and the likelihood ratio test are only slightly affected by the use of 3PLM item parameter estimates, but not by the use of the 2PLM estimates, even though the 2PLM has slight misfit to the data.

In the security breach situation with extreme test compromise, the predictive checking was compared to the likelihood ratio test, the adapted KL divergence and the simple linear regression in a fixed-form test. Results suggest that all methods are effective in detecting the extreme security breach. The regression method and the likelihood ratio test have inflated type-I error rate at the higher or lower end of  $\theta$  continuum, while the type-I error of the predictive checking and the KL divergence are

below the upper bound of the nominal level at all  $\theta$  levels. Although previous research suggests that the regression method is ineffective to detect test compromise, the regression method showed sufficient power to detect score difference between T1 and T2 in this study and its power is not substantially lower than the other IRT-based methods. The difference between findings in this study and in the previous study could be attributed to the difference in the level of simulated security breach. In this study, we considered an extreme situation where the keys to all 40 items in T2 are exposed. If such an extreme situation occurs in reality, our findings suggest that the regression method can effectively identify those cheaters. The regression method has slightly lower power than the IRT-based methods partly because the assumptions of simple linear regression is slightly violated with this simulated dataset. Additional analysis (see Figure B.1 in Appendix B) suggest that the homoscedasticity assumption is slightly violated: compared to the error variance at the middle range of T1 scores, the error variance at the two ends of T1 scores is smaller. Furthermore, because the data were simulated with the IRT model, the comparison is slightly biased towards the IRT-based methods.

The results in the security breach simulation also suggest that if responses on T1 are aberrant, the detection methods are still effective in detecting the score difference between T1 and T2 in the extreme security breach condition. However, some examinees may be falsely identified as cheaters. On the one hand, as mentioned earlier, the statistical evidence should only be used as a screening procedure to flag out some problematic examinees. Further investigation needs to be conducted to gather more evidence to conclude an examinee conducts cheating or not. On the other hand, this finding suggests that it is important to check the person-fit on T1 responses when using all the methods

proposed here. If person-fit analysis suggests that T1 responses involve some aberrancy, it implies that we cannot use T1 responses as a valid baseline to estimate a person's real proficiency, and thus a higher score on T2 does not necessarily imply that the person conducts cheating on T2.

## CHAPTER 6

### REAL DATA APPLICATION

#### 6.1 Data Description

A real dataset was used to illustrate the practical use of the predictive checking method and the other detection methods. The dataset comes from a state assessment measuring students' math proficiency in grade 4. The original dataset consists of 23583 examinees' responses to 63 items. Since some items are randomly assigned to examinees, there are a lot of missing responses in the file. To remove the missingness, 21 items that are randomly assigned among examinees were deleted. The remaining 42 items consisted of 38 dichotomous items and 4 polytomous items, and the polytomous items were all scored from 0 to 3. In addition, as the analysis was conducted at the person level, instead of using the entire examinee population, a sample of 5000 randomly selected examinees was used to simplify the analysis.

Instead of using the original response dataset, responses to some items were modified to create an artificially compromised dataset. Specifically, 21 items (19 dichotomous items and 2 polytomous items) were randomly selected from the 42 items as unsecure items. Then 5% examinees were randomly selected from the examinee sample and their responses to 8 items (7 dichotomous items and 1 polytomous item) in the unsecure subset were modified to create the compromised responses. The 8 compromised items were randomly selected from the unsecure subset for each examinee, so different examinees had compromised responses on different items. An examinee's response on a compromised dichotomous item was modified to be correct and one's score on the

compromised polytomous item was increased by 2-if the score after manipulation exceeded the maximum possible score, it was set to the maximum.

## 6.2 Data Analysis

As the first step of the analysis, the person-fit of responses on the artificially secure section (denoted “T1”) was evaluated for each examinee. As the predictive checking method assumes that the responses on T1 fit the IRT model, and considering that all methods could have high false positive rates if the responses to T1 do not reflect one’s real proficiency (as shown in study 2), only response vectors not identified as having misfit problems were kept in further analysis. The person-fit of responses on “T1” was evaluated using the popular person-fit statistic  $l_z$  (Drasgow et al., 1985; Sinharay, 2015), which has been shown to perform at least as well as or better than many other person-fit statistics (e.g. Drasgow, Levine, & McLaughlin, 1987). Previous studies have shown that when  $\hat{\theta}$  is used in the  $l_z$  calculation, the empirical distribution of  $l_z(\hat{\theta})$  deviates from the asymptotic distribution derived for  $l_z(\theta)$ . Sinharay (2015) constructed the null distribution of  $l_z(\hat{\theta})$  using the Bayesian PPC approach in a mixed-format test, and found it led to more power than using the asymptotic distribution, and the PPC  $p$ -value did not have the problem of being conservative in the case with  $l_z(\hat{\theta})$ . Therefore, PPC was used to construct the null distribution of  $l_z$  in the present study. A nominal level of 0.05 was used to identify misfitting responses.

The four methods discussed in this chapter- the predictive checking method, likelihood ratio test, simple linear regression and modified KL divergence- were applied to the modified dataset. In the implementation of the predictive checking method, the

summed score and the posterior variance were used as test statistics. The EAP was not used as study 1 showed that the summed score and the EAP had very similar performance when both subtests consisted of 20 items or more, and the summed score was computationally much easier than the EAP.  $N(0,2^2)$  was used as the prior distribution as the simulation in study 1 showed that the difference between prior configurations was small when both T1 and T2 were long, and using  $N(0,2^2)$  led to an easier computation. Item parameter estimates from the examinee population were used in person-level analysis. The 3PL model and the graded response model (GRM; Samejima, 1997) were used as the scoring models for dichotomous and polytomous items, respectively.

To evaluate the detection effectiveness among different methods, first of all, the detection rates among the examinees whose responses were modified (i.e. modified examinees) and the remaining examinees (i.e. unmodified examinees) were calculated. Second, the classification consistency- the agreement of classifying an examinee as a cheater or non-cheater- between every two methods was computed. The classification consistency was calculated in two approaches. The first approach was to count the observed proportion of detection agreement between two methods directly, denoted  $P_O$ . The second approach used Cohen's Kappa (Cohen, 1960), which corrected for the decision consistency by chance ( $P_C$ ). Cohen's Kappa takes the form of

$$\kappa = \frac{P_O - P_C}{1 - P_C} \quad (37)$$

$P_C$  can be calculated with the formula  $P_C = P_{1.}P_{.1} + P_{0.}P_{.0}$ , where  $P_{1.}$  represents the detection power by the first method, and  $P_{.1}$  represents the detection power by the second method,  $P_{0.} = 1 - P_{1.}$  and  $P_{.0} = 1 - P_{.1}$ . Lastly, the common cases flagged by all

methods as well as the unique cases flagged by each method only were identified. The posterior distributions of  $\theta$  on T1 and T2 were plotted for each of the cases to explore the characteristics of cases detected by different methods. If there were more than one cases flagged by all methods, or by one method uniquely, the case with the smallest  $p$ -value was chosen to make the plot. In addition, the unique cases detected by different methods were also highlighted in the scatterplots of examinees' summed scores/ EAP scores on T1 and T2 to further explore the detection characteristics of different methods. For comparison purposes, only cases showing an increase in  $\hat{\theta}$  were included in the plots.

## 6.3 Results

### 6.3.1 Detection rate

Table 6.1: Detection Rate (EAP change) of Different Methods

|                      | PC(sum)          | PC(var)           | LRT              | KL               | Reg              |
|----------------------|------------------|-------------------|------------------|------------------|------------------|
| Modified examinees   | 0.183<br>(1.883) | 0.085<br>(0.049)  | 0.248<br>(1.766) | 0.171<br>(1.928) | 0.301<br>(1.602) |
| Unmodified examinees | 0.012<br>(1.719) | 0.098<br>(-0.517) | 0.036<br>(1.487) | 0.015<br>(1.733) | 0.032<br>(1.422) |

*Note.* PC(sum) = predictive checking using the summed score, PC(var) = predictive checking using the posterior variance, LRT=likelihood ratio test, KL=KL divergence, and Reg=regression. The numbers in the parenthesis represent the average EAP increase from T1 to T2 among the examinees detected by each method.

Table 6.1 shows the detection rates among modified and unmodified examinees by each method. Different from the simulation results in study 2, the regression method has a better detection rate than the IRT-based method. This is probably because in the simulation in study 2, response data were generated based an IRT model, and there were some violations to the assumptions of the simple linear regression. However, with real data, the IRT model inevitably tends to have some misfit to the dataset, while the

assumptions of simple linear regression are well satisfied, as shown in Figure B.2 in the Appendix B. The comparisons among the three IRT-based methods demonstrate the same pattern among modified examinees as in the simulation in study 2: the likelihood ratio test has the largest detection rate, followed by the predictive checking method with the summed score; the KL divergence has slightly lower power than the predictive checking with the summed score and the predictive checking with the posterior variance has the lowest power, partly due to the use of a two-sided test for the posterior variance. Among unmodified examinees, the predictive checking using the posterior variance has the highest detection rate. The likelihood ratio test and the regression method detect more examinees than the KL divergence and the predictive checking with the summed score. It is also observed that the detection rates of the predictive checking with the summed score and the KL divergence are both substantially below the nominal type-I error rate. This could happen for several reasons. One possible cause is that a right-tailed test was conducted for these two methods to detect score increase on T2, but the observed score on T2 could be lower than the expected score due to some aberrant responses patterns such as careless responding or lack of motivation. Therefore, it is possible that the empirical distribution of the test statistic shifts to the left of the null distribution instead of to the right, and thus the empirical detection rate is lower than the type-I error from a right-tailed test. Another possible cause is that the test statistic's null distribution in the two methods were simulated from an IRT model, but the IRT model has some misfit to the observed responses in reality, and thus there is some discrepancy between the null distribution used here and the truth.

Results on the EAP change from T1 to T2 suggest that the predictive checking with the posterior variance could detect examinees with only slight change or even negative change in  $\hat{\theta}$ . This is because this method is not aimed at detecting score change, but at evaluating the flatness of the likelihood function. This implies that in order to use this method to detect preknowledge in particular, one can first use this statistic to identify examinees with flat likelihood functions (as implemented in this study), and then based on the direction of  $\hat{\theta}$  change, only examinees with positive  $\hat{\theta}$  change are further flagged out as possible “cheaters”. Among the rest of the methods, on average, examinees detected by the regression method and the likelihood ratio test show slightly smaller EAP increase than those detected by the KL divergence and the predictive checking with the summed score. This suggests that the regression and the likelihood ratio test are better at detecting examinees with a smaller score increase than the other two methods.

### **6.3.2 Classification Consistency**

Table 6.2 below shows the classification consistency between every two methods. The observed classification consistency is high between every two methods, but after correcting for the chance agreement, it is clear that the predictive checking with the posterior variance detected quite difference cases than the rest of the methods. Both the predictive checking with summed score and the likelihood ratio test have the highest classification consistency with the KL divergence, and the regression method has the highest classification consistency with the likelihood ratio test.

Table 6.2: Classification consistency among different methods

|         | PC(Sum) | PC(Var) | LRT    | KL     | Reg    |
|---------|---------|---------|--------|--------|--------|
| PC(Sum) | 1       | -0.024  | 0.561  | 0.733  | 0.559  |
| PC(Var) | 0.884   | 1       | -0.025 | -0.011 | -0.032 |
| LRT     | 0.972   | 0.862   | 1      | 0.656  | 0.627  |
| KL      | 0.989   | 0.883   | 0.977  | 1      | 0.558  |
| Reg     | 0.972   | 0.862   | 0.968  | 0.971  | 1      |

*Note.* The upper triangle shows the kappa coefficient, and the lower triangle shows the observed classification consistency.

### 6.3.3 Detection Characteristics by Different Methods

Figures 6.1 and 6.2 below display the  $\theta$  posterior distributions of modified and unmodified examinees detected by all methods, and the examinees detected by one method only. It turns out the KL divergence does not have unique detection- all the cases detected by the KL divergence were detected by the other methods, and among unmodified examinees, the predictive checking with the summed score does not have unique detection. Both Figures 6.1 and 6.2 suggest that for the case detected by all methods, the  $\theta$  posterior distribution shows a large shift from T1 to T2. Both figures also indicate that the predictive checking with the posterior variance can detect only slight shift in  $\theta$  posterior distributions; the likelihood ratio test can better detect  $\theta$  posterior distribution shift for very low or very high  $\theta$  levels, and the regression can better detect  $\theta$  posterior distribution shift for medium  $\theta$  levels. This point is further supported by the pattern shown in Figure 6.3, which highlights the unique cases detected by different methods. It is clear from Figure 6.3 that the likelihood ratio test tends to have unique detection at the lower or higher end of the ability continuum, while the regression method tends to have unique detection in the middle of the ability continuum. The predictive

checking with posterior variance detects many cases close to the fitted simple linear regression line.

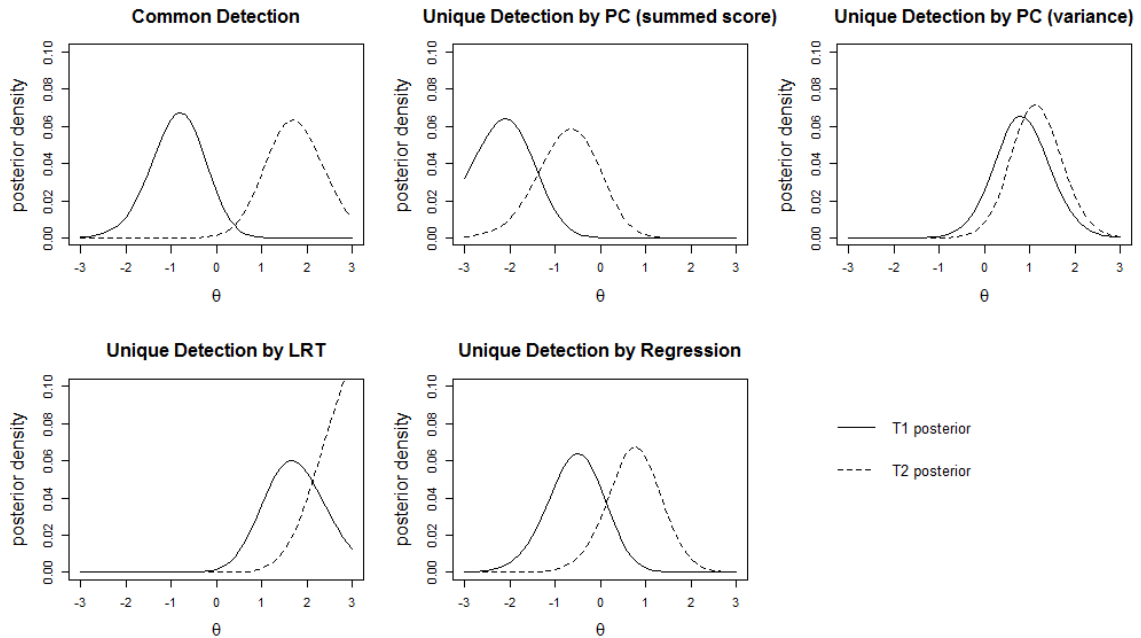


Figure 6.1: Posterior distribution of  $\theta$  for cases detected by different methods among modified examinees. The first plot shows the case detected by all methods, and the rest of the plots show cases detected by one method only but not by the other methods. PC (summed score) represents predictive checking using the summed score, PC (variance) represents predictive checking using the posterior variance, LRT represents likelihood ratio test.

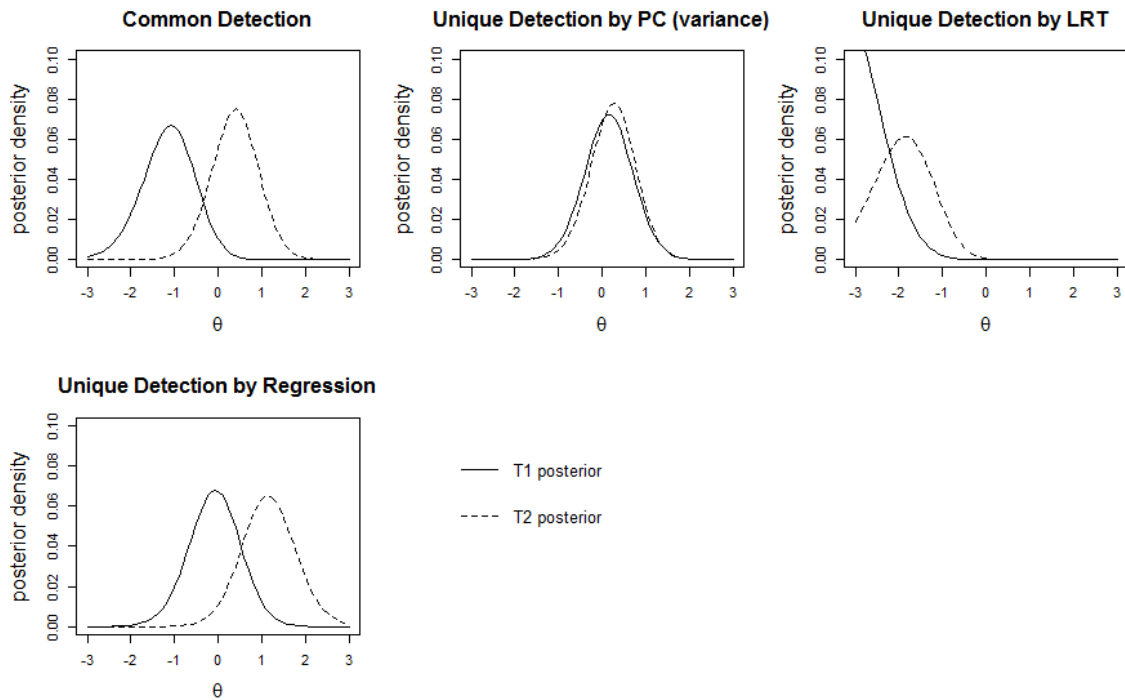


Figure 6.2: Posterior distribution of  $\theta$  for cases detected by different methods among unmodified examinees.

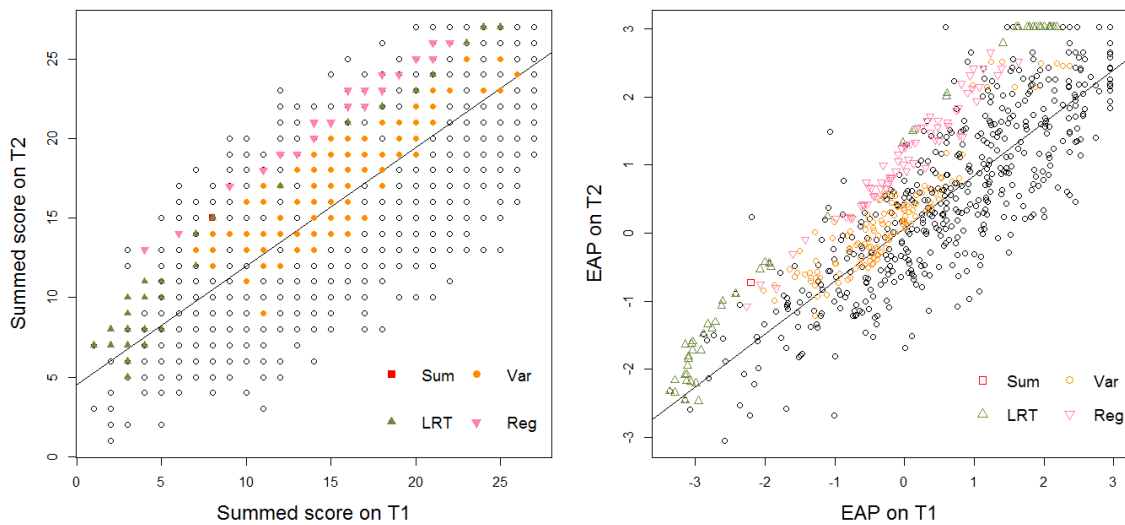


Figure 6.3: Scatterplot of summed score (left) and EAP (right) on T1 and T2. The solid line on each plot represents the fitted simple linear regression line based on the summed score or EAP.

## CHAPTER 7

### DISCUSSION AND CONCLUSIONS

This study focuses on a test security problem where examinees might have preknowledge on repeatedly-used items in a continuous testing program. A predictive checking method was proposed to detect whether an examinee uses preknowledge on exposed items by using information from secure items. To investigate the effectiveness of this method, three studies were conducted in sequence. The first study aims at understanding the statistical properties of this method, and a simulation study was designed to evaluate the empirical type-I error and power of this method under different lengths of secure and insecure sections, by considering different prior configurations and different test statistics. The second study aims at comparing this method to three other methods in two simulated test compromise situations that are likely to happen in practice. The third study aims at demonstrating the practical use of different methods, and investigating their detection consistency in a real dataset. The following three paragraphs summarize some general findings from the three studies.

First of all, regarding the detection effectiveness of the predictive checking method at the item-set level, both study 1 and 2 have suggested that the type-I error of this method gets close to the nominal level of 0.05 when both subsets are long (e.g. containing 20 items or more). With a secure subset containing 20 or more items, this method has moderate to large power to detect preknowledge on 20% or more items in the insecure subset among low to medium-ability individuals by using the summed score or the EAP as test statistics. In the extreme situation where a large number of insecure items are compromised, using only five or ten secure items can lead to sufficient detection

power. Using the variance of the  $\theta$  posterior distribution does not exhibit large power by itself, but both the simulation and the real data analysis suggest using it along with the summed score or the EAP can help detect examinees with a relatively small score increase. However, caution needs to be taken if two statistics are used to evaluate the same response vector at the same time, as multiple hypothesis tests are conducted simultaneously and the family-wise error rate needs to be controlled. In this study, the detection difference between prior configurations is small, and this suggests that one can simply choose a less-informative prior that leads to easier computation in practice. The item-level detection does not turn out to be successful in this study, as the power is too low and the false positive rate is too high.

Second, regarding the comparison between the predictive checking and three other methods (i.e. the likelihood ratio test, adapted KL divergence, and the simple linear regression), the simulation results suggest that the predictive checking and the likelihood ratio test have larger power than the other methods. The predictive checking has very similar power to the likelihood ratio test, but the likelihood ratio test has a larger type-I error rate. However, as mentioned in the discussion section in Chapter 5, the likelihood ratio test relies on an asymptotic sampling distribution, which may not work well when either subtest is short. The predictive checking, in comparison, can be used with short subtests at a cost of being more computationally intensive. Study 2 also suggests that the power of the adapted KL divergence may be affected to a large extent when item parameter estimates from the 3PLM are used, which could limit its effectiveness in detecting preknowledge among low-ability examinees. The simple linear regression also demonstrates large power in detecting the extreme security breach. Its performance in the

simulation study is slightly worse than the other methods in that it has inflated type-I error among high  $\theta$  levels, and it has slightly lower power than the other methods, but this result could be related to the data generation model adopted in study 2.

Third, the real data application gives some different results regarding the comparison among different methods. The simple linear regression has the highest detection rates than the other methods among artificially cheating examinees. The comparison among the three IRT-based methods in the real dataset demonstrates a similar pattern as in the simulation study: the likelihood ratio test flags more suspicious examinees than the predictive checking, and the KL divergence flags slightly fewer examinees among artificial cheaters than the predictive checking. The real data analysis also suggest that although the four methods (i.e. predictive checking with the summed score, adapted KL divergence, likelihood ratio test, and simple linear regression) have high detection consistency, one method may perform better in detecting a particular type of score increase than the others. For instance, the likelihood ratio is more likely to detect score increase among high- or low-ability individuals, while the simple linear regression is more likely to detect score increase in the middle of the ability continuum.

The findings in this study have several practical implications. The results in the key exposure situation suggest that if one's primary goal is to detect severe security breach due to item preknowledge, the simple linear regression can be considered as the first choice, since it is the easiest to implement among all the methods under investigation in this study, and it does not need to be based on the IRT model. To detect small to moderate test compromise, the likelihood ratio test can be used when both subsets have more than 20 items, while the predictive checking method can be used when either subset

is short. To increase the detection power under small to moderate test compromise situations, one can narrow the possibly compromised subset down to items that are most likely to be compromised or items with moderate to high difficulty level. Items that are more likely to be compromised can be determined through some external information regarding item leakage. Alternatively, it can be determined through some systematic investigation, such as monitoring the statistical behavior of each item over time and identifying items that become aberrantly easier at a certain time point. Those items can be flagged as potentially compromised items, and can be included in the possibly compromised section at the person-level investigation.

This study has several limitations. First of all, except for one condition in the security breach simulation in study 2, the simulations in this project assume the responses to the secure items fit the IRT model well. This is unlikely to happen in practice. For instance, those responses are likely to be affected by some aberrant response behaviors other than preknowledge, such as careless responding, test speededness, etc. The only condition in study 2 that incorporates aberrant responses in secure items considered an extreme case where the responses to all items in the 20-item secure section only have a 0.25 chance of being correct. This makes the secure section useless in inferring a person's proficiency. Future study can evaluate the performance of the predictive checking and other methods when there is some small to moderate level of misfit in responses to the secure items, such that the responses to the secure items still provide some valid information to infer a person's ability, but contain some noises.

The second limitation is that all studies in this project only evaluated the type-I error and power of the detection methods at the nominal level of 0.05. This particular

nominal level was chosen due to its common use in hypothesis testing, but it only provides partial information for the performance of a method in the null and alternative condition. To better justify the use of a statistical test, one can evaluate the  $p$ -value distribution of the test statistic in the null condition and see whether it asymptotically approximates a uniform distribution<sup>5</sup>. In addition, to more comprehensively evaluate the power of a statistical test, one can look at the ROC curve which plots the power against different type-I error rates and report the area-under-curve (AUC) indices for each method.

The third limitation relates to the use of the data-generating model in the simulation studies. As implied by the findings from the real-data analysis, where the comparison between the regression method and the other methods showed a different pattern from that in the simulation study, the comparison among methods may depend on the model we used to generate the data in the null and alternative condition. In all simulation studies, the IRT model was used to generate response data in the null condition, which gave a bias to all IRT-based approaches. Future studies could consider using a non-parametric model to generate response data. In addition, the data-generating models in the preknowledge condition are relatively simple in study 1 and in the shallow pool simulation in study 2: the compromised responses were generated by adding a constant to the probability in the null condition or changing the responses to correct. Different data-generating models are likely to have an impact on the performance of each method and the comparison between methods. Future studies can consider several

---

<sup>5</sup> The author did evaluate the  $p$ -value distribution of the predictive checking in the null condition, but did not evaluate this for the other methods in the null condition.

different data-generating models in the preknowledge condition by using some existing models in the literature, such as the model proposed by McLeod et al. (2003) and the one by Shu et al. (2013), and then evaluate the impact of different data-generating models on the detection effectiveness of different methods.

The fourth limitation is related to the use of different item parameters in the person-level analysis. In the shallow pool simulation in study 2, it was found that the use of item parameters from different models did not have a large impact on the detection effectiveness for the predictive checking method and the likelihood ratio test. This was partly because a large calibration sample (i.e. with 5000 sample size) was used, and thus the estimation error in item parameters might be negligible compared to the variance of the  $\theta$  posterior distribution. Future studies can further evaluate this factor by considering a smaller calibration sample, such as a sample size of 2000 or fewer.

As for future research directions, in addition to overcoming the four limitations above, future studies can consider three more directions. The first is to incorporate response time information into the predictive checking. By adopting a response time model and by considering the joint distribution of the latent person-level variables in the IRT model and the response time model, one can both predict a person's responses and response times on the possibly compromised items. This will provide more information for a person's response behavior than using the responses alone. In addition, as response time is a continuous quantity, the item-level detection using response times is expected to be more promising as we can have a continuous predictive distribution of response time on a single item, instead of working with dichotomous response variables.

Second, in this study, the item-level detection does not show any success, mainly because each response only has two categories and the item-level detection is based on the idea of “leaving-one-out”. As mentioned above, item-level detection is expected to have some promise if working with some continuous variables (such as response time) or variables with multiple categories (such as polytomous item with more than five categories). In that case, it is not necessary to use the “leave-one-out” method to conduct item-level detection. One can just obtain the predictive distribution of the relevant variable on a single item, and compare the observed value to the predictive distribution. This can overcome the problem of high false positive rates.

Third, the item-level detection in this project is at the person-level. In other words, it is aimed at detecting whether a person uses preknowledge on a single item. As shown in study 1 that addressed dichotomous responses, this type of item-level detection does not work well. Future study can extend the idea of item-level detection to the group level. In other words, instead of focusing on detecting suspicious examinees, the interest can be shifted to monitoring a single item’s performance. This is another type of quality control procedure, aiming at identifying problematic items over test administrations. Predictive checking may be applied to this context as well by first predicting a person’s responses to an item and then aggregating the prediction to a group level.

Lastly, there is a final cautionary point that should be made when using statistical methods to detect examinee preknowledge. All the methods considered in this project belong to the frequentist hypothesis testing procedures. In general, in hypothesis testing, rejecting the null hypothesis does not necessarily mean the alternative hypothesis is true, and similarly, not rejecting the null hypothesis does not necessarily mean the null

hypothesis is true. This implies that all the methods can only be used as a screening procedure to identify suspicious examinees who might have used preknowledge. If an examinee is flagged by a certain method, further investigation needs to be conducted to gather additional evidence to prove that an examinee does use preknowledge to gain unfair score increase or not. It is not our intention to recommend using statistical methods alone to make a judgement about an examinee's testing taking behavior. Instead, we recommend using these methods as a quick screening process and using them to infer the extent of test compromise in a certain examinee group. By identifying potential test security problems in an examinee group, one can take actions to improve the security for their testing programs as early as possible, so as to create ensure the validity and fairness of testing.

## APPENDIX A

### TABLES

Table A.1: True Item Parameters in Study 1

| T1=5  |       | T2=5  |        |
|-------|-------|-------|--------|
| a     | b     | a     | b      |
| 0.960 | 0.960 | 1.656 | -0.311 |
| 1.199 | 1.199 | 1.782 | -1.517 |
| 0.985 | 0.985 | 1.279 | 0.525  |
| 1.237 | 1.237 | 1.244 | 0.387  |
| 1.161 | 1.161 | 0.793 | -0.640 |
| T1=10 |       | T2=10 |        |
| a     | b     | a     | b      |
| 1.625 | 1.625 | 0.893 | 0.172  |
| 1.286 | 1.286 | 0.872 | -0.017 |
| 1.490 | 1.490 | 1.310 | 0.741  |
| 1.066 | 1.066 | 1.024 | 0.097  |
| 1.214 | 1.214 | 1.419 | -1.244 |
| 1.680 | 1.680 | 1.118 | -0.191 |
| 1.336 | 1.336 | 1.347 | 1.692  |
| 1.008 | 1.008 | 1.961 | -0.418 |
| 1.875 | 1.875 | 1.021 | 1.147  |
| 1.547 | 1.547 | 1.430 | -0.198 |
| T1=20 |       | T2=20 |        |
| a     | b     | a     | b      |
| 0.872 | 0.872 | 1.752 | -0.743 |
| 1.166 | 1.166 | 0.897 | -1.385 |
| 1.131 | 1.131 | 1.269 | 0.052  |
| 1.192 | 1.192 | 0.837 | 0.009  |
| 1.164 | 1.164 | 0.936 | -1.090 |
| 1.640 | 1.640 | 1.025 | -1.248 |
| 1.976 | 1.976 | 0.759 | -0.160 |
| 1.755 | 1.755 | 1.022 | -0.817 |
| 1.031 | 1.031 | 1.592 | -1.622 |
| 1.679 | 1.679 | 0.990 | -0.461 |
| 0.880 | 0.880 | 1.104 | -0.115 |
| 1.606 | 1.606 | 1.213 | -0.536 |
| 1.298 | 1.298 | 1.114 | 0.779  |
| 0.785 | 0.785 | 0.879 | 0.878  |
| 1.161 | 1.161 | 1.511 | 1.106  |

---

|       |       |       |        |
|-------|-------|-------|--------|
| 1.831 | 1.831 | 1.288 | -1.049 |
| 1.043 | 1.043 | 1.456 | -1.019 |
| 1.618 | 1.618 | 0.825 | -0.539 |
| 1.443 | 1.443 | 1.355 | -0.357 |
| 0.780 | 0.780 | 1.045 | -0.406 |

---

Table A.2: Empirical Type-I Error at Item-set Level Using Fiducial and Jeffreys Prior

|              | $\theta$ | Sum   |       | EAP   |       | Var    |        |
|--------------|----------|-------|-------|-------|-------|--------|--------|
|              |          | JEF   | FID   | JEF   | FID   | JEF    | JEF    |
| T1=5, T2=10  | -2       | 0.002 | 0.000 | 0.015 | 0.000 | 0.013  | 0.008  |
|              | -1       | 0.024 | 0.009 | 0.043 | 0.021 | 0.034  | 0.019  |
|              | 0        | 0.030 | 0.025 | 0.048 | 0.034 | 0.067* | 0.061  |
|              | 1        | 0.029 | 0.021 | 0.045 | 0.033 | 0.044  | 0.035  |
|              | 2        | 0.018 | 0.013 | 0.034 | 0.025 | 0.001  | 0.001  |
| T1=20, T2=10 | -2       | 0.023 | 0.018 | 0.050 | 0.047 | 0.020  | 0.018  |
|              | -1       | 0.021 | 0.020 | 0.044 | 0.044 | 0.035  | 0.034  |
|              | 0        | 0.022 | 0.018 | 0.041 | 0.034 | 0.040  | 0.039  |
|              | 1        | 0.026 | 0.021 | 0.045 | 0.037 | 0.055  | 0.051  |
|              | 2        | 0.013 | 0.009 | 0.017 | 0.015 | 0.017  | 0.011  |
| T1=10, T2=5  | -2       | 0.003 | 0.008 | 0.015 | 0.021 | 0.006  | 0.007  |
|              | -1       | 0.024 | 0.026 | 0.046 | 0.045 | 0.029  | 0.025  |
|              | 0        | 0.029 | 0.027 | 0.038 | 0.035 | 0.029  | 0.028  |
|              | 1        | 0.011 | 0.010 | 0.012 | 0.011 | 0.027  | 0.030  |
|              | 2        | 0.000 | 0.000 | 0.000 | 0.000 | 0.008  | 0.010  |
| T1=10, T2=10 | -2       | 0.004 | 0.001 | 0.018 | 0.003 | 0.007  | 0.001  |
|              | -1       | 0.026 | 0.020 | 0.050 | 0.031 | 0.041  | 0.025  |
|              | 0        | 0.026 | 0.027 | 0.056 | 0.053 | 0.046  | 0.045  |
|              | 1        | 0.030 | 0.027 | 0.042 | 0.041 | 0.041  | 0.040  |
|              | 2        | 0.009 | 0.007 | 0.013 | 0.011 | 0.017  | 0.009  |
| T1=10, T2=20 | -2       | 0.005 | 0.000 | 0.009 | 0.002 | 0.021  | 0.016  |
|              | -1       | 0.039 | 0.019 | 0.051 | 0.028 | 0.045  | 0.028  |
|              | 0        | 0.046 | 0.039 | 0.057 | 0.055 | 0.062  | 0.064* |
|              | 1        | 0.045 | 0.041 | 0.062 | 0.056 | 0.042  | 0.032  |
|              | 2        | 0.020 | 0.014 | 0.042 | 0.032 | 0.028  | 0.015  |

Table A.3: Empirical Type-I Error at Item-set Level Using Normal Prior

| Stat | $\theta$ | T2=5  |       |       | T2=10 |       |       | T2=20  |        |       |
|------|----------|-------|-------|-------|-------|-------|-------|--------|--------|-------|
|      |          | T1=5  | T1=10 | T1=20 | T1=5  | T1=10 | T1=20 | T1=5   | T1=10  | T1=20 |
| Sum  | -2       | 0.002 | 0.002 | 0.020 | 0.002 | 0.004 | 0.017 | 0.003  | 0.004  | 0.038 |
|      | -1       | 0.014 | 0.014 | 0.019 | 0.022 | 0.026 | 0.020 | 0.020  | 0.039  | 0.031 |
|      | 0        | 0.033 | 0.027 | 0.021 | 0.028 | 0.029 | 0.021 | 0.034  | 0.047  | 0.024 |
|      | 1        | 0.012 | 0.009 | 0.001 | 0.028 | 0.032 | 0.022 | 0.033  | 0.047  | 0.033 |
|      | 2        | 0.000 | 0.000 | 0.000 | 0.018 | 0.010 | 0.010 | 0.039  | 0.020  | 0.037 |
| EAP  | -2       | 0.004 | 0.006 | 0.039 | 0.005 | 0.016 | 0.044 | 0.004  | 0.005  | 0.053 |
|      | -1       | 0.023 | 0.033 | 0.040 | 0.033 | 0.046 | 0.040 | 0.028  | 0.049  | 0.044 |
|      | 0        | 0.039 | 0.038 | 0.027 | 0.040 | 0.059 | 0.038 | 0.043  | 0.069* | 0.037 |
|      | 1        | 0.015 | 0.010 | 0.001 | 0.043 | 0.045 | 0.039 | 0.048  | 0.064* | 0.043 |
|      | 2        | 0.000 | 0.000 | 0.000 | 0.033 | 0.014 | 0.014 | 0.069* | 0.042  | 0.051 |
| Var  | -2       | 0.003 | 0.005 | 0.015 | 0.013 | 0.005 | 0.019 | 0.021  | 0.015  | 0.041 |
|      | -1       | 0.040 | 0.024 | 0.024 | 0.025 | 0.039 | 0.035 | 0.024  | 0.042  | 0.046 |
|      | 0        | 0.036 | 0.033 | 0.024 | 0.059 | 0.048 | 0.039 | 0.095* | 0.063* | 0.046 |
|      | 1        | 0.021 | 0.025 | 0.033 | 0.038 | 0.043 | 0.053 | 0.014  | 0.039  | 0.058 |
|      | 2        | 0.001 | 0.003 | 0.004 | 0.001 | 0.014 | 0.011 | 0.016  | 0.024  | 0.034 |

Table A.4: Power at Item-set Level Using Fiducial and Jeffreys Prior When Compromise Rate is 100%

|              | $\theta$ | Sum   |       | EAP   |       | Var   |       |
|--------------|----------|-------|-------|-------|-------|-------|-------|
|              |          | JEF   | FID   | JEF   | FID   | JEF   | FID   |
| T1=5, T2=10  | -2       | 0.691 | 0.627 | 0.760 | 0.692 | 0.375 | 0.274 |
|              | -1       | 0.583 | 0.544 | 0.636 | 0.593 | 0.079 | 0.055 |
|              | 0        | 0.466 | 0.424 | 0.502 | 0.446 | 0.010 | 0.000 |
|              | 1        | 0.258 | 0.253 | 0.268 | 0.255 | 0.014 | 0.000 |
|              | 2        | 0.066 | 0.066 | 0.066 | 0.066 | 0.003 | 0.000 |
| T1=20, T2=10 | -2       | 0.938 | 0.935 | 0.961 | 0.960 | 0.647 | 0.640 |
|              | -1       | 0.886 | 0.879 | 0.930 | 0.929 | 0.109 | 0.099 |
|              | 0        | 0.764 | 0.755 | 0.792 | 0.790 | 0.356 | 0.339 |
|              | 1        | 0.431 | 0.409 | 0.432 | 0.412 | 0.272 | 0.240 |
|              | 2        | 0.062 | 0.054 | 0.062 | 0.054 | 0.023 | 0.021 |
| T1=10, T2=5  | -2       | 0.533 | 0.475 | 0.658 | 0.583 | 0.191 | 0.162 |
|              | -1       | 0.552 | 0.520 | 0.590 | 0.571 | 0.222 | 0.224 |
|              | 0        | 0.363 | 0.356 | 0.370 | 0.364 | 0.192 | 0.191 |
|              | 1        | 0.028 | 0.029 | 0.028 | 0.029 | 0.009 | 0.008 |
|              | 2        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| T1=10, T2=10 | -2       | 0.838 | 0.804 | 0.878 | 0.856 | 0.444 | 0.356 |
|              | -1       | 0.804 | 0.787 | 0.852 | 0.847 | 0.095 | 0.075 |
|              | 0        | 0.700 | 0.694 | 0.735 | 0.721 | 0.199 | 0.182 |
|              | 1        | 0.498 | 0.478 | 0.503 | 0.482 | 0.291 | 0.257 |
|              | 2        | 0.050 | 0.039 | 0.050 | 0.039 | 0.013 | 0.012 |
| T1=10, T2=20 | -2       | 0.941 | 0.930 | 0.947 | 0.939 | 0.463 | 0.354 |
|              | -1       | 0.912 | 0.902 | 0.911 | 0.904 | 0.057 | 0.032 |
|              | 0        | 0.920 | 0.900 | 0.922 | 0.915 | 0.576 | 0.535 |
|              | 1        | 0.789 | 0.782 | 0.826 | 0.810 | 0.561 | 0.510 |
|              | 2        | 0.305 | 0.270 | 0.305 | 0.270 | 0.036 | 0.025 |

Table A.5: Power at Item-set Level Using Fiducial and Jeffreys Prior When Compromise Rate is 60%

|              | $\theta$ | Sum   |       | EAP   |       | Var   |       |
|--------------|----------|-------|-------|-------|-------|-------|-------|
|              |          | JEF   | FID   | JEF   | FID   | JEF   | FID   |
| T1=5, T2=10  | -2       | 0.326 | 0.197 | 0.340 | 0.202 | 0.264 | 0.170 |
|              | -1       | 0.255 | 0.215 | 0.254 | 0.199 | 0.266 | 0.244 |
|              | 0        | 0.192 | 0.167 | 0.178 | 0.149 | 0.127 | 0.124 |
|              | 1        | 0.092 | 0.079 | 0.096 | 0.071 | 0.010 | 0.008 |
|              | 2        | 0.037 | 0.032 | 0.037 | 0.033 | 0.001 | 0.000 |
| T1=20, T2=10 | -2       | 0.650 | 0.641 | 0.662 | 0.648 | 0.528 | 0.518 |
|              | -1       | 0.485 | 0.487 | 0.452 | 0.448 | 0.291 | 0.296 |
|              | 0        | 0.289 | 0.287 | 0.248 | 0.241 | 0.109 | 0.099 |
|              | 1        | 0.116 | 0.108 | 0.117 | 0.106 | 0.059 | 0.048 |
|              | 2        | 0.022 | 0.018 | 0.022 | 0.018 | 0.008 | 0.005 |
| T1=10, T2=5  | -2       | 0.220 | 0.167 | 0.420 | 0.350 | 0.233 | 0.206 |
|              | -1       | 0.209 | 0.186 | 0.345 | 0.325 | 0.107 | 0.106 |
|              | 0        | 0.092 | 0.092 | 0.137 | 0.133 | 0.034 | 0.035 |
|              | 1        | 0.010 | 0.012 | 0.011 | 0.012 | 0.005 | 0.003 |
|              | 2        | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| T1=10, T2=10 | -2       | 0.431 | 0.352 | 0.455 | 0.351 | 0.333 | 0.207 |
|              | -1       | 0.391 | 0.364 | 0.361 | 0.334 | 0.281 | 0.274 |
|              | 0        | 0.262 | 0.260 | 0.224 | 0.214 | 0.105 | 0.110 |
|              | 1        | 0.132 | 0.119 | 0.131 | 0.118 | 0.045 | 0.039 |
|              | 2        | 0.027 | 0.021 | 0.027 | 0.021 | 0.010 | 0.009 |
| T1=10, T2=20 | -2       | 0.617 | 0.503 | 0.649 | 0.542 | 0.494 | 0.351 |
|              | -1       | 0.509 | 0.488 | 0.536 | 0.516 | 0.239 | 0.214 |
|              | 0        | 0.418 | 0.404 | 0.415 | 0.395 | 0.037 | 0.027 |
|              | 1        | 0.236 | 0.219 | 0.268 | 0.247 | 0.160 | 0.135 |
|              | 2        | 0.077 | 0.058 | 0.102 | 0.080 | 0.040 | 0.034 |

Table A.6: Power at Item-set Level Using Normal Prior

| Stat | Rate | $\theta$ | T2=5  |       |       | T2=10 |       |       | T2=20 |       |       |
|------|------|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|      |      |          | T1:5  | T1:10 | T1:20 | T1:5  | T1:10 | T1:20 | T1:5  | T1:10 | T:20  |
| Sum  | 1    | -2       | 0.400 | 0.517 | 0.699 | 0.678 | 0.841 | 0.936 | 0.751 | 0.946 | 0.992 |
|      |      | -1       | 0.368 | 0.558 | 0.606 | 0.576 | 0.812 | 0.881 | 0.663 | 0.915 | 0.977 |
|      |      | 0        | 0.262 | 0.370 | 0.374 | 0.461 | 0.708 | 0.749 | 0.654 | 0.924 | 0.930 |
|      |      | 1        | 0.053 | 0.029 | 0.016 | 0.259 | 0.499 | 0.419 | 0.599 | 0.800 | 0.784 |
|      |      | 2        | 0.002 | 0.000 | 0.000 | 0.066 | 0.053 | 0.056 | 0.384 | 0.318 | 0.400 |
|      | 0.6  | -2       | 0.140 | 0.191 | 0.371 | 0.308 | 0.435 | 0.639 | 0.388 | 0.627 | 0.836 |
|      |      | -1       | 0.141 | 0.215 | 0.235 | 0.238 | 0.405 | 0.473 | 0.307 | 0.525 | 0.698 |
|      |      | 0        | 0.099 | 0.096 | 0.080 | 0.182 | 0.274 | 0.266 | 0.240 | 0.432 | 0.457 |
|      |      | 1        | 0.019 | 0.010 | 0.005 | 0.092 | 0.134 | 0.111 | 0.165 | 0.244 | 0.213 |
|      |      | 2        | 0.002 | 0.000 | 0.000 | 0.037 | 0.028 | 0.018 | 0.106 | 0.083 | 0.093 |
| EAP  | 1    | -2       | 0.467 | 0.650 | 0.783 | 0.719 | 0.882 | 0.960 | 0.763 | 0.950 | 0.996 |
|      |      | -1       | 0.384 | 0.593 | 0.657 | 0.617 | 0.859 | 0.925 | 0.663 | 0.915 | 0.977 |
|      |      | 0        | 0.262 | 0.379 | 0.380 | 0.498 | 0.740 | 0.778 | 0.672 | 0.926 | 0.936 |
|      |      | 1        | 0.053 | 0.029 | 0.016 | 0.268 | 0.506 | 0.425 | 0.636 | 0.835 | 0.810 |
|      |      | 2        | 0.002 | 0.000 | 0.000 | 0.066 | 0.053 | 0.056 | 0.384 | 0.318 | 0.400 |
|      | 0.6  | -2       | 0.314 | 0.411 | 0.585 | 0.306 | 0.452 | 0.653 | 0.419 | 0.657 | 0.879 |
|      |      | -1       | 0.230 | 0.353 | 0.409 | 0.214 | 0.380 | 0.439 | 0.311 | 0.552 | 0.704 |
|      |      | 0        | 0.132 | 0.142 | 0.114 | 0.163 | 0.242 | 0.236 | 0.242 | 0.428 | 0.469 |
|      |      | 1        | 0.028 | 0.011 | 0.006 | 0.090 | 0.135 | 0.112 | 0.174 | 0.273 | 0.230 |
|      |      | 2        | 0.002 | 0.000 | 0.000 | 0.037 | 0.028 | 0.018 | 0.123 | 0.108 | 0.113 |
| Var  | 1    | -2       | 0.182 | 0.238 | 0.324 | 0.320 | 0.420 | 0.632 | 0.308 | 0.474 | 0.716 |
|      |      | -1       | 0.153 | 0.246 | 0.251 | 0.066 | 0.088 | 0.085 | 0.035 | 0.062 | 0.066 |
|      |      | 0        | 0.140 | 0.201 | 0.208 | 0.018 | 0.178 | 0.289 | 0.162 | 0.570 | 0.774 |
|      |      | 1        | 0.025 | 0.009 | 0.002 | 0.032 | 0.295 | 0.256 | 0.392 | 0.760 | 0.688 |
|      |      | 2        | 0.001 | 0.000 | 0.000 | 0.007 | 0.013 | 0.021 | 0.177 | 0.214 | 0.202 |
|      | 0.6  | -2       | 0.198 | 0.239 | 0.326 | 0.206 | 0.320 | 0.513 | 0.331 | 0.492 | 0.765 |
|      |      | -1       | 0.118 | 0.121 | 0.100 | 0.254 | 0.290 | 0.283 | 0.196 | 0.247 | 0.291 |
|      |      | 0        | 0.031 | 0.036 | 0.041 | 0.158 | 0.113 | 0.100 | 0.010 | 0.036 | 0.065 |
|      |      | 1        | 0.008 | 0.004 | 0.000 | 0.012 | 0.043 | 0.053 | 0.024 | 0.155 | 0.156 |
|      |      | 2        | 0.001 | 0.000 | 0.000 | 0.004 | 0.009 | 0.008 | 0.052 | 0.056 | 0.055 |

Table A.7: Empirical Type-I Error at Item Level

|              | $\theta$ | JEF   |        | N(0,2 <sup>2</sup> ) |        | FID   |        |
|--------------|----------|-------|--------|----------------------|--------|-------|--------|
|              |          | $b=1$ | $b=0$  | $b=1$                | $b=0$  | $b=1$ | $b=0$  |
| T1=5, T2=10  | -2       | 0.043 | 0.030  | 0.039                | 0.024  | 0.035 | 0.018  |
|              | -1       | 0.053 | 0.047  | 0.047                | 0.044  | 0.041 | 0.040  |
|              | 0        | 0.052 | 0.072* | 0.049                | 0.071* | 0.046 | 0.059  |
|              | 1        | 0.063 | 0.054  | 0.059                | 0.051  | 0.043 | 0.033  |
|              | 2        | 0.020 | 0.040  | 0.018                | 0.041  | 0.013 | 0.047  |
| T1=20, T2=10 | -2       | 0.054 | 0.031  | 0.053                | 0.027  | 0.057 | 0.032  |
|              | -1       | 0.043 | 0.032  | 0.043                | 0.030  | 0.042 | 0.029  |
|              | 0        | 0.040 | 0.069* | 0.042                | 0.070* | 0.040 | 0.068* |
|              | 1        | 0.058 | 0.053  | 0.056                | 0.049  | 0.057 | 0.052  |
|              | 2        | 0.026 | 0.059  | 0.019                | 0.049  | 0.018 | 0.046  |
| T1=10, T2=5  | -2       | 0.025 | 0.015  | 0.020                | 0.013  | 0.026 | 0.002  |
|              | -1       | 0.046 | 0.031  | 0.043                | 0.029  | 0.036 | 0.018  |
|              | 0        | 0.042 | 0.029  | 0.041                | 0.029  | 0.036 | 0.036  |
|              | 1        | 0.026 | 0.034  | 0.026                | 0.032  | 0.024 | 0.028  |
|              | 2        | 0.018 | 0.010  | 0.016                | 0.009  | 0.016 | 0.010  |
| T1=10, T2=10 | -2       | 0.044 | 0.026  | 0.041                | 0.022  | 0.038 | 0.018  |
|              | -1       | 0.050 | 0.052  | 0.049                | 0.054  | 0.038 | 0.041  |
|              | 0        | 0.057 | 0.066* | 0.057                | 0.067* | 0.054 | 0.062  |
|              | 1        | 0.059 | 0.064* | 0.060                | 0.063  | 0.057 | 0.051  |
|              | 2        | 0.024 | 0.046  | 0.023                | 0.040  | 0.017 | 0.037  |
| T1=10, T2=20 | -2       | 0.036 | 0.042  | 0.033                | 0.040  | 0.026 | 0.040  |
|              | -1       | 0.050 | 0.056  | 0.050                | 0.054  | 0.057 | 0.043  |
|              | 0        | 0.058 | 0.064* | 0.058                | 0.068* | 0.063 | 0.063  |
|              | 1        | 0.061 | 0.061  | 0.061                | 0.060  | 0.051 | 0.056  |
|              | 2        | 0.063 | 0.045  | 0.061                | 0.041  | 0.053 | 0.040  |

Table A.8: Empirical Power at Item Level

|              | $\theta$ | Rate=0.2 |                      |       | Rate=0.4 |                      |       |
|--------------|----------|----------|----------------------|-------|----------|----------------------|-------|
|              |          | JEF      | N(0,2 <sup>2</sup> ) | FID   | JEF      | N(0,2 <sup>2</sup> ) | FID   |
| T1=5, T2=10  | -2       | 0.363    | 0.319                | 0.304 | 0.251    | 0.174                | 0.320 |
|              | -1       | 0.164    | 0.148                | 0.140 | 0.113    | 0.097                | 0.128 |
|              | 0        | 0.071    | 0.070                | 0.061 | 0.051    | 0.046                | 0.052 |
|              | 1        | 0.029    | 0.025                | 0.005 | 0.003    | 0.002                | 0.010 |
|              | 2        | 0.003    | 0.002                | 0.001 | 0.001    | 0.000                | 0.006 |
| T1=20, T2=10 | -2       | 0.302    | 0.285                | 0.313 | 0.313    | 0.297                | 0.029 |
|              | -1       | 0.100    | 0.099                | 0.105 | 0.126    | 0.125                | 0.068 |
|              | 0        | 0.046    | 0.048                | 0.047 | 0.047    | 0.045                | 0.065 |
|              | 1        | 0.038    | 0.039                | 0.037 | 0.013    | 0.013                | 0.057 |
|              | 2        | 0.024    | 0.015                | 0.011 | 0.011    | 0.008                | 0.059 |
| T1=10, T2=5  | -2       | 0.287    | 0.235                | 0.216 | 0.140    | 0.144                | 0.121 |
|              | -1       | 0.117    | 0.112                | 0.109 | 0.066    | 0.092                | 0.099 |
|              | 0        | 0.083    | 0.080                | 0.066 | 0.036    | 0.056                | 0.038 |
|              | 1        | 0.038    | 0.038                | 0.036 | 0.006    | 0.013                | 0.010 |
|              | 2        | 0.025    | 0.022                | 0.022 | 0.004    | 0.001                | 0.002 |
| T1=10, T2=10 | -2       | 0.332    | 0.306                | 0.282 | 0.258    | 0.231                | 0.159 |
|              | -1       | 0.127    | 0.117                | 0.107 | 0.115    | 0.116                | 0.103 |
|              | 0        | 0.058    | 0.057                | 0.055 | 0.045    | 0.048                | 0.040 |
|              | 1        | 0.038    | 0.036                | 0.030 | 0.009    | 0.009                | 0.004 |
|              | 2        | 0.016    | 0.016                | 0.005 | 0.014    | 0.013                | 0.004 |
| T1=10, T2=20 | -2       | 0.291    | 0.220                | 0.215 | 0.196    | 0.144                | 0.072 |
|              | -1       | 0.090    | 0.081                | 0.079 | 0.090    | 0.095                | 0.081 |
|              | 0        | 0.061    | 0.063                | 0.060 | 0.055    | 0.049                | 0.049 |
|              | 1        | 0.019    | 0.023                | 0.022 | 0.002    | 0.005                | 0.003 |
|              | 2        | 0.015    | 0.008                | 0.014 | 0.006    | 0.005                | 0.004 |

Table A.9: False Positive Rate at Item Level

|              | $\theta$ | Rate=0.2 |                      |       | Rate=0.4 |                      |       |
|--------------|----------|----------|----------------------|-------|----------|----------------------|-------|
|              |          | JEF      | N(0,2 <sup>2</sup> ) | FID   | JEF      | N(0,2 <sup>2</sup> ) | FID   |
| T1=5, T2=10  | -2       | 0.084    | 0.058                | 0.029 | 0.285    | 0.239                | 0.188 |
|              | -1       | 0.086    | 0.069                | 0.068 | 0.183    | 0.165                | 0.159 |
|              | 0        | 0.084    | 0.078                | 0.065 | 0.148    | 0.141                | 0.140 |
|              | 1        | 0.060    | 0.058                | 0.057 | 0.081    | 0.080                | 0.087 |
|              | 2        | 0.053    | 0.054                | 0.059 | 0.057    | 0.053                | 0.059 |
| T1=20, T2=10 | -2       | 0.159    | 0.148                | 0.150 | 0.468    | 0.460                | 0.467 |
|              | -1       | 0.100    | 0.097                | 0.094 | 0.292    | 0.291                | 0.300 |
|              | 0        | 0.078    | 0.077                | 0.075 | 0.176    | 0.168                | 0.169 |
|              | 1        | 0.062    | 0.055                | 0.058 | 0.074    | 0.070                | 0.072 |
|              | 2        | 0.060    | 0.052                | 0.048 | 0.042    | 0.044                | 0.042 |
| T1=10, T2=5  | -2       | 0.053    | 0.048                | 0.030 | 0.304    | 0.159                | 0.099 |
|              | -1       | 0.073    | 0.077                | 0.055 | 0.237    | 0.189                | 0.159 |
|              | 0        | 0.042    | 0.043                | 0.051 | 0.135    | 0.127                | 0.145 |
|              | 1        | 0.026    | 0.025                | 0.022 | 0.030    | 0.037                | 0.035 |
|              | 2        | 0.002    | 0.002                | 0.004 | 0.002    | 0.002                | 0.000 |
| T1=10, T2=10 | -2       | 0.090    | 0.082                | 0.050 | 0.343    | 0.333                | 0.266 |
|              | -1       | 0.108    | 0.109                | 0.087 | 0.271    | 0.273                | 0.247 |
|              | 0        | 0.077    | 0.080                | 0.072 | 0.156    | 0.162                | 0.151 |
|              | 1        | 0.057    | 0.057                | 0.049 | 0.076    | 0.075                | 0.077 |
|              | 2        | 0.045    | 0.041                | 0.042 | 0.042    | 0.038                | 0.040 |
| T1=10, T2=20 | -2       | 0.081    | 0.075                | 0.026 | 0.308    | 0.304                | 0.176 |
|              | -1       | 0.102    | 0.084                | 0.056 | 0.213    | 0.241                | 0.195 |
|              | 0        | 0.073    | 0.091                | 0.085 | 0.160    | 0.168                | 0.151 |
|              | 1        | 0.079    | 0.088                | 0.077 | 0.141    | 0.121                | 0.113 |
|              | 2        | 0.064    | 0.060                | 0.065 | 0.081    | 0.079                | 0.084 |

## APPENDIX B

### FIGURES

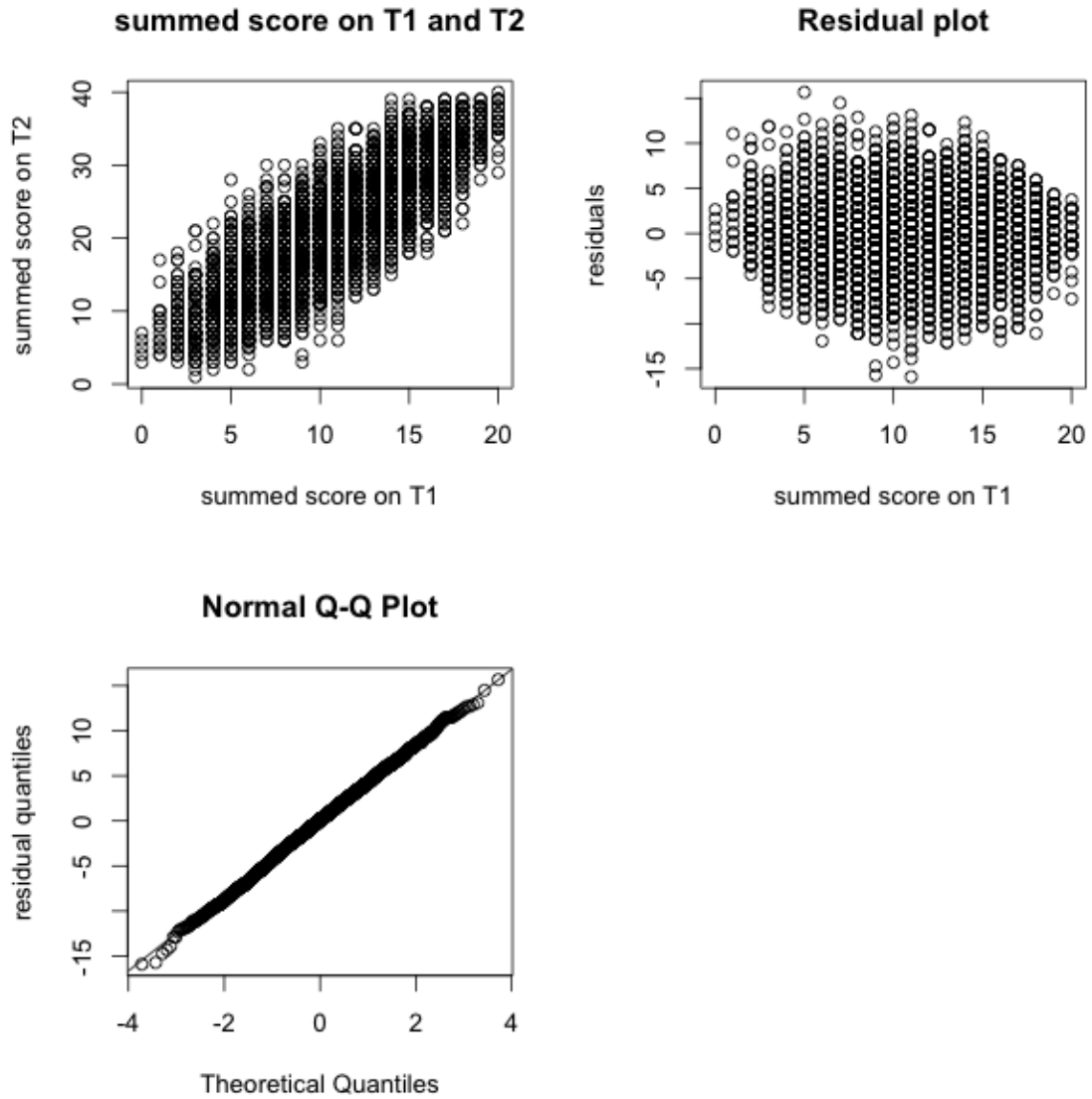


Figure B.1: Plots to check assumptions in simple linear regression in study 2.

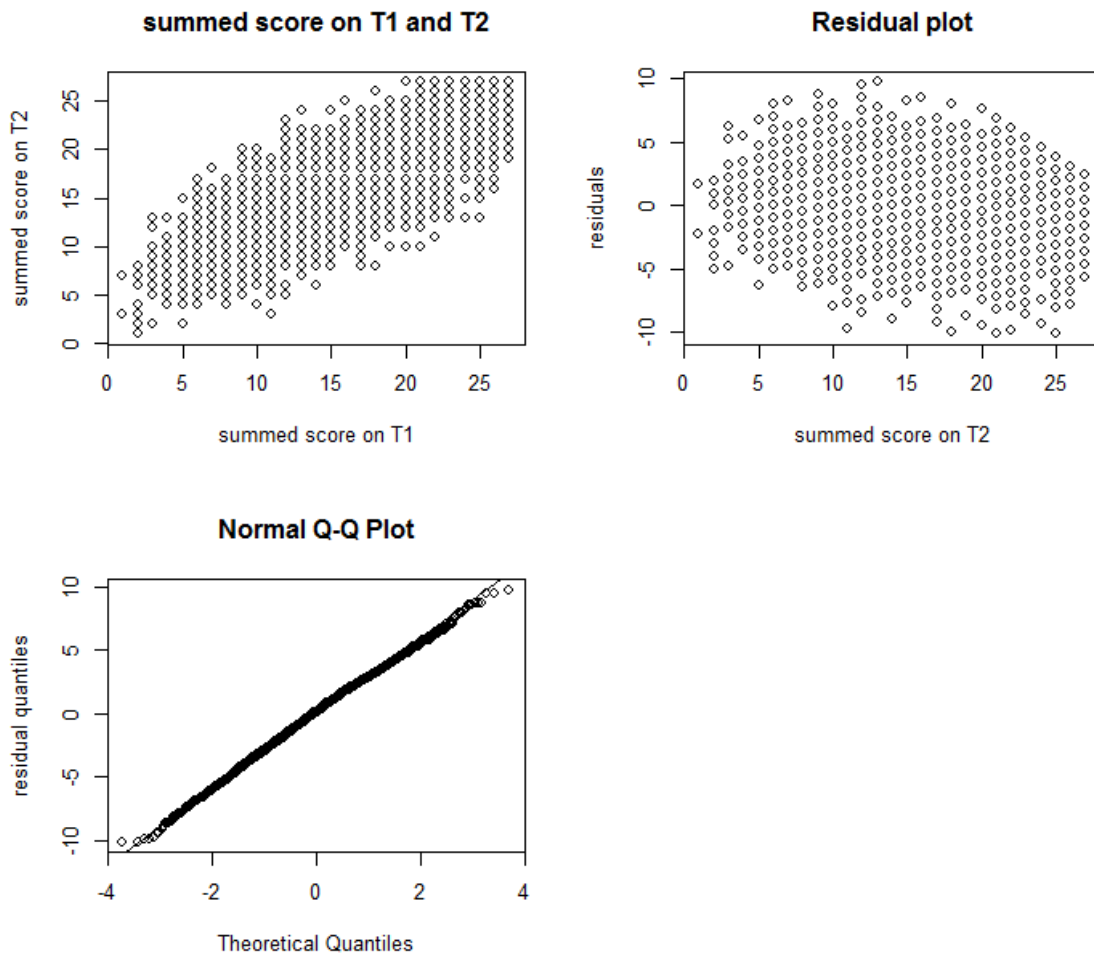


Figure B.2: Plots to check assumptions in simple linear regression in real data analysis.

## BIBLIOGRAPHY

- Bayarri, M. J., & Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 452, 1127-1142.
- Belov, D. I. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, 50, 141–163.
- Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2, 3, 37-58.
- Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7–14). Tokyo, Japan: Universal Academy Press.
- Billingsley, P. (1986). *Probability and measure*. New York: Wiley.
- Binet, A. & Simon, Th. A. (1905). M'ethodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *l'Anne'e Psychologie*, 11, 191–336.
- Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci*, 16, 2, 101-133.
- Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37–46
- Donlon, T.F., & Fischer, F.E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement*, 28, 105-113.
- Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11(1), 59-79.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* xxvi, 528-535.
- Geisser S. (1993). *Predictive Inference: An Introduction*. New York: Chapman and Hall.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D.B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. New York: Chapman & Hall.
- Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–807.
- Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8).
- Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, 27(3), 217–233.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.

- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.
- Haberman, S. (2008). *Outliers in assessments (RR-08-41)*. Princeton, NJ: Educational Testing Service.
- Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, *19*, 491-544.
- Hannig, J. (2013). Generalized Fiducial Inference via Discretization, *Statistica Sinica*, *23*, 489 – 514.
- Hannig, J., Iyer, H., Lai, R. C. S., & Lee, T. C. M. (2015). *Generalized fiducial inference: A review*. Manuscript under review.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA, US: Sage Publications, Inc.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*(3), 133-146.
- Hechinger, J. (2008). *Wall street journal*. Retrieved from <http://www.wsj.com/articles/SB122109733923122015>
- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*(1), 105-126.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.
- Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, *56*, 535–547.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*, 161–176.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290.
- Lewis, C., Lee, Y., & von Davier, A. A. (2012, May). *Test security for multistage tests: A quality control perspective*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Li, F., Gu, L., & Manna, V. (2004, April). *Methods to detect group-level aberrance in state standardized assessment*. Paper presented at the National Council on Measurement in Education Meeting, Philadelphia, PA.
- Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*, 215–231.
- Liu, Y. (2015). Generalized fiducial inference for graded response models (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses (UMI No.: 3703849).
- Liu, Y., & Hannig, J. (2016). Generalized fiducial inference for binary logistic item response models. *Psychometrika*, Advanced online publication. doi: 10.1007/s11336-015-9492-7.
- Luecht, R. M. (2003). *Exposure control using adaptive multistage item bundles*. Paper Presented at National Council on Measurement in Education, Chicago, IL.

- McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, 23, 147–160.
- McLeod, L. D., Lewis, C. & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27, 121–137.
- Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, 21(4), 321-336.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135.
- Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75–106.
- Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, 21(2), 115-127.
- Nering, M.L., Meijer, R.R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, 22(1): 53-69.
- Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement*, 16(4), 345-352.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19, 213–229.
- Robins, J. M., Vaart, A., & Ventura, V. (2000). Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association*, 95, 452, 1143-1156.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.
- Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, 55 (3), 3-38.
- Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tosho.
- Segall, D. (2002). An item response model for characterizing test comprise. *Journal of Educational and Behavioral Statistics*, 27(2), 163–179.
- Shu, Z. (2010). *Detecting test cheating using a Deterministic, gated item response theory model*. Unpublished Dissertation. University of North Carolina at Greensboro.
- Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, 78, 3, 481-97.
- Samejima, F. (1997). Graded response model. In van der Linden, W. J., & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7, 131-145.
- Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, 42, 4, 375-394.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30,4, 298-321.
- Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models*. Princeton, NJ, ETS. Retrieved September 10, 2015, from [http://www.ets.org/research/policy\\_research\\_reports/publications/report/2003/hsho](http://www.ets.org/research/policy_research_reports/publications/report/2003/hsho)
- Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, 40(4), 343-365.
- Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 1(23), 57-75.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, 7, 81-96.
- van der Flier, H. (1977). Environmental factors and deviant response patterns (pp. 30-35). In Y.P. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 30-35). Amsterdam: Swets and Zeitlinger.
- van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23(4), 327-345.
- Wang, C., Zheng, Y., & Chang, H.-H. (2014). Does standard deviation matter? Using “standard deviation” to quantify security of multistage testing. *Psychometrika*, 79(1), 154-174.
- Wang, X., Li, F., & Gu, L. (2015, April) Using Person-fit statistics to detect cheating due to item preknowledge. Paper presented at the 2015 American Educational Research Association meeting, Chicago, IL.
- Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security*. New York: Routledge.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago: Mesa.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. New York: CRC Press.