ABSTRACT

Title of dissertation: METHODS OF INTEGRATING MULTI-MODAL DATA FOR DETECTING ABERRANT TEST-TAKING BEHAVIORS IN LARGE-SCALE ASSESSMENTS

Kaiwen Man
Doctor of Philosophy, 2020

Dissertation directed by: Professor Jeffrey R. Harring
Department of Human Development and Quantitative Methodology

Many schools, states, and countries use scores from large-scale assessments in making important high-stakes decisions in such areas as college admissions, academic performance evaluations, and job promotions among others. These decisions rely on accurate, reliable scores from which valid inferences about examinees can be assessed. However, aberrant test-taking behaviors, including copying from other test-takers and practicing with real items ahead of time, undermine the effectiveness of such assessments in yielding accurate, precise information on an examinee's performance. Also, with the wide adoption of technology-enhanced online learning and testing system, especially as I am writing my thesis while the outbreak of COVID-19 virus, it is critical to address an example question like "how to make the online-delivered tests more secure?" As a result, investigating ways to identify potential cheaters after these assessments or batteries have been taken and data collected is an important endeavor for the numerous administrators of

such assessments. The purpose of this line of research is to create, develop, investigate, and test new approaches that will incorporate bio-information technology, such as eye-tracking, into current machine-learning methods in the detection of cheating and other aberrant testing behaviors in computer-based testing scenarios. In other words, cheating detection for innovative large-scale assessments with big data techniques augmented by bio-information technologies will be explored. The eye-tracking systems, in particular, have the potential to capture cheating and other aberrant test-taking behaviors with visual information gathered through the analysis of eye movement patterns (saccades, fixations, pupil size). This type of data can be subtly gathered in real-time on test-takers as they attempt to answer each assessment item. To assess the visual attention nuances across test-takers, three negative binomial distribution-based visual fixation counts models will be presented. Moreover, a joint-modeling approach of integrating product data (e.g., item responses), process data (e.g., response times), and biometric information (visual fixation counts) will be demonstrated. By joint modeling the three types of information, we can assess test-takers' performance in a comprehensive way. Finally, selected supervised and unsupervised statistical learning methods will be explored for detecting different types of responding behaviors.

METHODS OF INTEGRATING MULTI-MODAL DATA FOR DETECTING ABERRANT TEST-TAKING BEHAVIORS IN LARGE-SCALE ASSESSMENTS

by

Kaiwen Man

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Jeffrey R. Harring, Chair/Advisor
Professor Hong Jiao
Professor Yang Liu
Professor Donald J. Bolger
Professor Colleen O'Neal

Dedication

I dedicate this dissertation to God Almighty my creator, my solid pillar, my source of wisdom, inspiration, knowledge, love and understanding.

"The fear of the Lord is the beginning of wisdom." – Proverbs 1:7

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible, and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I would like to thank my academic advisor, Professor Jeffrey Harring, for carrying me along this journey with patience and trust, who has always been supportive in my good and bad times. Since I joined the EDMS family, he has provided me significant opportunities to develop my professional career. My current academic accomplishments, if any, are mainly due to Dr. Harring's tremendous endorsement and supervision. He is not only my mentor but also my role model.

I would also like to thank our program director, Dr. Gregory Hancock. Because of his kindness, I had this valuable opportunity to chase my dream of being a professor in America. Because of his unshaken faith in me, I would be able to continue my studies after a bumpy ride during my first year in the program. Thanks for being such a wonderful person in my life! Moreover, I would also like to thank Dr. Huahua Chang, Dr. Hong Jiao, Dr. Kadriye Ercikan, Dr. Sandip Sinharay, Dr. Kun Yuan, Dr. Donald Bolger, Dr. Colleen O'Neal, and Dr. Aimo Hinkkanen for their support and guidance.

I would also like to acknowledge help and support from some of my friends. I would like to thank Dongbo Guo for his encouragement over the years, especially during the difficult times. Thanks, Sarah Thomas, for linking me with other professionals who are working in the test security field, which helped my research tremendously. Thanks, Peida Zhan, for his insights and support on many research projects. Thanks to my church friends Xiaofang Wang and Dan for their kind invitations for Thanksgiving and Christmas

parties over the years, which gave me a strong sense of belonging during the holidays. Thanks to my church elder Der-Chen Chang for his teachings about God, which brings me the true happiness and peace. Thanks, Joseph Feser, for his valuable insights and help on many things such as job interview preparation, thesis proofreading, and career decision-making. Also, I would like to thank my other friends: Yewon Lee, Yi Wei, Feng Yi, Monica Morell, Daniel Lee, and Tessa Johnson for their help along the journey.

Additionally, I owe my deepest thanks to my family - my mother and father who have always stood by me, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them.

Furthermore, I would like to acknowledge financial support from the Educational Testing Service by awarding me with the Harold Gulliksen Psychometric Research Fellowship to conduct my dissertation research.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank God for His unconditional love, blessings, and infinite mercy upon me.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables viii

List of Figures ix

1 Introduction 1
- 1.1 Limitations of Previous Works . . . 3
- 1.2 Conceptual Framework for Current Study . . . 5
- 1.3 Research Significance . . . 7

2 Literature Review 9
- 2.1 Test Security . . . 9
- 2.2 Statistical Methods for Identifying Aberrant Testing Behavior . . . 12
  - 2.2.1 Similarity Analysis (Collusion, Answer-Copy) . . . 12
    - 2.2.1.1 Identical Errors Analysis . . . 13
    - 2.2.1.2 Error Similarity Analysis . . . 14
    - 2.2.1.3 Generalized Binomial Similarity . . . 16
  - 2.2.2 Gain Score Analysis . . . 17
  - 2.2.3 Erasure (Answer Changing) Analysis . . . 18
  - 2.2.4 Person Fit Analysis . . . 20
    - 2.2.4.1 Representative Parametric Indices . . . 21
    - 2.2.4.2 Representative Non-Parametric Indices . . . 25
    - 2.2.4.3 Representative Response Time based Index . . . 28
  - 2.2.5 Use of Data Mining Methods to Detect Test Fraud . . . 30
    - 2.2.5.1 Unsupervised Machine Learning Methods . . . 31
    - 2.2.5.2 Supervised Machine Learning Methods . . . 36
- 2.3 Incorporating Biometrics to Detect Aberrant Testing Behaviors . . . 45
  - 2.3.1 Insights into Problem-Solving Using Eye Tracking . . . 45
    - 2.3.1.1 Fixation . . . 46

2.3.1.2 Pupil Diameter . . . 47
2.3.1.3 Blinking . . . 48
2.3.1.4 Saccades . . . 49
2.3.1.5 Regression . . . 49
2.3.2 Representative Ways to Integrate Process Data into Psychometric Methods to Identify Aberrant Test Takers . . . 50
2.3.2.1 Incorporating RT as a Variable Into the Item Response Model . . . 51
2.3.2.2 Joint Modeling of Item Responses and Response Times 53
2.4 Future Directions and Challenges . . . 59
2.5 Conclusion . . . 61

3 Methodology 62
3.1 Experimental Design . . . 62
3.1.1 Data Collection . . . 62
3.1.2 Experimental Conditions . . . 63
3.1.3 Data Recording . . . 64
3.2 New Test Engagement Model Based on Visual Fixation Counts . . . 64
3.2.1 The Negative Binomial Fixation Model . . . 65
3.2.2 Negative Binomial Fixation Model with Linear Trend . . . 68
3.2.3 Negative Binomial Fixation Model with Quadratic Trend . . . 71
3.3 A Three-way Joint Modeling Approach of Item Response, Response Time and Fixation Counts . . . 72
3.3.1 Measurement Models at Level 1 . . . 72
3.3.2 Modeling Item Domain and Person Domain models at Level 2 . . 73
3.3.2.1 Modeling Person Domain Parameters . . . 73
3.3.2.2 Modeling Item Domain Parameters . . . 74
3.3.3 Model Parameter Estimation . . . 76
3.3.4 Evaluating Model-data Fit: Posterior Predictive Model Checking . 78
3.4 Integration of Bio- and Psychometrical Infomation into Machine Learning Methods for Detecting Aberrant Behaviors . . . 83
3.4.1 Data Normalization . . . 84
3.4.2 Feature Selection . . . 85
3.4.3 Outcome Measures and Expected Results . . . 86
3.5 Research Significance . . . 87

4 Results 89
4.1 Summary Statistics of the Collected Data across conditions . . . 89
4.2 Data Visualization Across Different Experimental Conditions . . . 90
4.3 Negative Binomial Visual Fixation Models . . . 94
4.3.1 Item Parameter Estimates . . . 94
4.3.2 Person Parameter Estimates . . . 95
4.4 Three-way Factor Model Parameter Estimates . . . 98
4.4.1 Item Parameter Estimates . . . 99
4.4.2 Variance-Covariance Estimates . . . 100

4.4.3 Item-Side Variance-Covariance Structure 100
4.4.4 Person-Side Variance-Covariance Structure 102
4.4.5 Accessing the Item-Wise Data Model Fit 102
4.5 Assessing Test-Taking Behaviors Across Different Experimental Conditions 104
4.5.1 Impact of Having Pre-knowledge of Test Items on Item Characteristics 105
4.5.2 Impact of Having Pre-knowledge of Test Items on Test-Takers Behavior 108
4.6 Use of Person-fit Statistics to Classify Different Responding Behaviors . . 111
4.7 Use of Data Mining Methods to Classify Different Responding Behaviors 113
4.7.1 Representative Unsupervised Learning Methods 116
4.7.2 Representative Supervised Learning Methods 119

5 Discussion 127
5.0.1 Limitations for the Current Work 130
5.0.2 Recommendations for Future Directions 131

A List of Variable Names 134

B Summary Statistics of All the Variables 136

List of Tables

1.1 General guidelines of data forensics Analysis 2

2.1 Representative person fit indices 21
2.2 Guttman scale and index calculation 27
2.3 Advantages and disadvantages of various data mining methods for detecting aberrant test taking behaviors 44

3.1 Input psychological and biological variables for data mining methods . . 86

4.1 Number of subjects in each condition 90
4.2 Item parameter estimates 95
4.3 Variance covariance estimates 96
4.4 Item parameter estimates of three-way factor model 99
4.5 Variance-covariance estimates of three-way model 101
4.6 Item parameter estimates across different experimental conditions 107
4.7 Person-side correlation matrix estimates 109
4.8 Sensitivity and specificity for PFS IRT- and RT-based methods 112
4.9 Classification accuracy for K-Means methods with three groups 116
4.10 Sensitivity and specificity for K-means methods with two groups 117
4.11 Classification accuracy for KNN methods with three groups 121
4.12 Sensitivity and specificity for KNN Methods with two groups 121
4.13 Classification accuracy for RF Methods with three groups 124
4.14 Sensitivity and specificity for RF methods with two groups 125

A.1 Variable names 135

B.1 Summary Statistics of All the Variables 137

List of Figures

2.1 Relationship between several representative indicators 50
2.2 Conditional independence of item responses and response times given latent ability and speediness 56
2.3 A mixture modeling approach to investigate the intraindividual variation in responses and response times 57
2.4 A conditional joint modeling approach for locally dependent item responses and response times 58

3.1 A graphical representation of the negative binomial fixation model 68
3.2 Two items with fitted fixation counts 69
3.3 A graphical representation of the negative binomial fixation model 70
3.4 Trivariate joint model approach of item response, response time, and visual fixation counts 76
3.5 Graphical demonstration of posterior predictive model checking (PPMC) Method 81

4.1 Scatterplots of essential variables under condition 1. 91
4.2 Scatterplots of essential variables under condition 2. 92
4.3 Scatterplots of essential variables under condition 3. 93
4.4 Individual test engagement estimates based on the NBFM. 97
4.5 Individual test engagement estimates based on the NBFM-LT. 97
4.6 Individual test engagement estimates based on the NBFM-QT. 98
4.7 Scatter-plots for item parameter estimates 101
4.8 Scatter-plots for person parameter estimates 102
4.9 Posterior predictive p-values for 1-PL IRT model, log-normal response time model, and negative binomial visual fixation counts model. 104
4.10 Item parameter estimates across distinct experimental conditions, and negative binomial visual fixation counts model. 108
4.11 Scatterplots for person-side parameter estimates. A loess non-parametric smoothed curve is plotted for each scatterplot 110
4.12 PFS's performance of classifying different type of responding behaviors . 112
4.13 Pair-wise correlations between features 114
4.14 Feature importance 115

4.15 Number of optimal clusters based on K-Means method 117
4.16 Segregations among the three groups based on K-means method 118
4.17 The optimal number of neighborhood based on the KNN 120
4.18 Segregations among the three groups based on KNN method 123
4.19 Number of trees in random forest 124
4.20 Classification Tree as a demonstration of classifying different types of responding behaviors 125

Chapter 1: Introduction

Over the last decade, the number of cheating-related test security events has grown (Wollack & Fremer, 2013), especially on tests that aim to assess student achievement. These incidents have become more discernible with the use of computer-based testing (CBT), probably due to the high number of tests given, which could lead to higher rates of item-exposure. Cheating behavior on educational and psychological tests has been known to compromise the accuracy of results on assessments of student achievement (Cizek & Wollack, 2017; Meijer & Sijtsma, 1995; W. J. van der Linden & Guo, 2008; van Krimpen-Stoop & Meijer, 2001), and thus influence the inferences drawn from these scores. These undesirable outcomes are exacerbated in high-stakes, competitive assessment scenarios in which fraudulent test-taking behavior not only influences the scoring of the deviant test-taker, but causes harm to other test-takers as these questionable scores impact others scores with whom they are directly compared (Sinharay, 2017). Meijer (1997) suggested that the existence of misfitting responses in test data could negatively impact the reliability of test scores and validity of the inferences drawn from the scores on these tests. Hendrawan, Glas, and Meijer (2005) also indicated that model data misfit due to aberrant response patterns negatively impact item parameter estimation, which could result in inaccurate latent ability estimates of examinees.

In order to secure our exams from unethical test-taking behaviors like cheating, in recent years, a number of statistical methods have been proposed for detecting different types of test-taking behaviors. The Council of Chief State School Officers (CCSSO) and Association of Test Publishers (ATP) have suggested the following guidelines for searching for misconduct (see Table 1.1).

Table 1.1: General guidelines of data forensics Analysis

Data Forensics Analysis	**Types of Testing Irregularities**
Unusual score gains or losses (test-retake)	Coaching on actual test content, "helping" during an examination
Eraser (answer changing) analysis	Changing answers by educators, Inappropriate assistance during testing
Similarity analysis (collusion)	Sharing answers during testing, teachers helping before or during testing, illicit use of stolen test questions
Person fit analysis (aberrant wrong and right answer patterns)	Inconsistent response patterns such as answering difficult questions correctly while missing easy questions

(Olson & Fremer, 2013)

Based on these guidelines, many statistical indices were created for detecting aberrant test-taking behaviors. Response similarity indices are used to evaluate the agreement of two response vectors from two test takers, mainly focusing on flagging answer copy cheating or collusion among individual test-takers. The representative indices are the K index (Holland, 1996a), K_1, K_2, S_1 indices (Sijtsma & Meijer, 1992; Sotaridona, van der Linden, & Meijer, 2006a), Generalized Binomial index (W. J. van der Linden & Sotaridona, 2006) and the ω index (Wollack, 1997). Person-fit indices are computed to assess

different response patterns of test-takers, which could also be used for detecting copy cheating and other types of behaviors such as pre-knowledge cheating and item stealing (e.g., Belov & Armstrong, 2010; Fox & Marianti, 2016; Sinharay, 2017). Eraser detection indices (EDIs) are utilized to detect suspicious answer changing behaviors (e.g., Sinharay & Johnson, 2017; Wollack & Cizek, 2016). Gain scores analysis (GSA) are more focused on flagging teacher cheating. In other words, GSA is mainly used to catch the unexpected test score fluctuations at the group level. Representative GSA methods include those proposed by Bishop and Egan (2016) and Skorupski and Egan (2011).

1.1 Limitations of Previous Works

The use and effectiveness of many of those indices used for aberrant behavior detection have been compared and reported in many studies (e.g. Karabatsos, 2003; Reise, 1990; Sinharay, 2017). Many of these methods have been found to be sensitive in detecting different types of behaviors. However, despite nuanced successes of these methods to detect aberrant testing behavior, they have a number of limitations. The detection power of the previously introduced indexes and statistics are constrained by their corresponding inputs, such as item responses or response times, without further considering other information of the test taker. Also, many testing programs have transferred from conventional paper-pencil tests to computer-based tests and testing environments in these years. A wealth of test-taker related multimodal behavioral data (item responses, response times and process information) are collected in real time during the administration of tests. Yet, most of the previously mentioned indices have limited power to incorporate the high

dimensional behavioral information into their functional forms. Furthermore, current detection methods are isolated from each other on detecting different types of aberrant behaviors due to the nature of their designs. As a result, a unified platform for aggregating the detection power of the most effective methods is lacking.

Some recent methodological investigations have attempted to address the aforementioned limitations (e.g. Dai, 2013; Fox & Marianti, 2017a). Dai (2013) introduced a mixture Rasch model to explore underlying latent groups by incorporating covariates, collateral information, which could positively impact classification accuracy. Fox and Marianti (2017a) proposed a person-fit test, which accounts for relations between item and person characteristics by jointly modeling item responses and response times. Though these methods are good examples of integrating auxiliary information for more accurate classification of aberrant-behaved test takers from the normally-behaved group, they are still deficient in a number of ways. First, incorporating high-dimensional input variables with complex structures is challenging under current modeling frameworks. For example, it is increasingly challenging for models to converge if many covariates are added in the mixture model. Also, many of these variables could be nonlinearly-related to an outcome of interest. Thus, ignoring such nonlinearities existing among all the covariates may imply that the fitted linear model fails to accurately capture a systematic pattern between the outcome variable and a set of predictors (Huber, 1991). Secondly, the power to detect aberrant behaviors based on these methods for item responses or response times alone has been quite modest (e.g., Fox & Marianti, 2017a). Inclusion of essential behavioral indicators has the potential for improving the power to detect aberrant test-taking behaviors.

1.2 Conceptual Framework for Current Study

In order to address the previous challenges, this study aims to create, develop, investigate, and test new approaches that would incorporate bio-information technology, such as eye-tracking, into traditional psychometric models (e.g., IRT and RT models). Also, the bio-information measures and other psychometric measures can be used as inputs for various machine-learning methods in the detection of cheating and other aberrant testing behaviors in computer-based testing scenarios. In other words, cheating detection on innovative large-scale assessments with big data techniques augmented by bio-information technologies will be explored. Data mining algorithms, a class of methods for clustering cases, can be a convenient methodological platform for detecting fraudulent test taking behaviors that can overcome noted limitations of traditional methods. Sensitivity to detect aberrant behavior can be potentially increased by incorporating not only process and biometrical data as inputs into these algorithms, but also indices based on traditional approaches. Additionally, in contrast to applications involving traditional IRT based, and RT based methods, data mining algorithms have the facility to examine both linear as well as nonlinear relations among variables. Therefore, complex interactions between background, psychometric, and biometric variables would be better approximated than simply assuming linear relationships among them.

Eye-gaze pattern related variables recorded by an eye-tracking system, in particular, will be incorporated into the data-mining platform as essential biometric indicators. The eye-tracking system has the potential to capture cheating and other aberrant test-taking behaviors with bio-information gathered through the analysis of eye movement patterns

(e.g., saccades, fixations, pupil size). Such information in the context of large-scale assessment testing scenarios may help address and answer some interesting questions such as: (a) where does an examinee look and what does this information tell us about aberrant testing behavior? (b) What information can be ignored? (c) When does blinking occur, and what information does that convey about the examinees behavior? (d) How does the pupil react to different stimuli? This type of bio-information can be subtly gathered in real-time on test-takers as they answer assessment items in a CBT environment. Thus, many supervised and unsupervised statistical learning methods will be explored in this study by incorporating biometric information. Also, a new eye-gaze pattern based model will be proposed. Moreover, the new model would be further jointly modeled with item response model (IRT) and(or) response time models (RT).

Given the purpose of this research, several research questions to be addressed are:

1. Do differences exist in classification accuracy measured by sensitivity and specificity rates across the different data mining methods?

2. Which data source - item responses, response times, or eye-tracking data - is most predictive of aberrant test taking behavior detection?

3. Do differences exist in classification accuracy of aberrant testing behaviors among conventional standalone approaches and data mining methods? If differences exist, can they be partially explained by incorporating both biometric and psychological data?

1.3 Research Significance

This study explores the ways of incorporating biological information into traditional psychological methods by utilizing data-mining algorithms to better understand test takers' behaviors. This methodological work has the potential to aid administrators of large scale assessments to ferret out aberrant behaving examinees, but can lead to future research in the area of test security. Second, this study has the potential to create new eye-tracking measure-oriented models and develop methods that may flag aberrances with increased accuracy by incorporating biometric measures into current framework. An advantage of the newly proposed statistical methods is that they would not be solely based on one source of information, such as item responses, but rather on multiple sources of information about test-takers including data stemming from bio-information technologies and integrating log file information. Ideally, all of this information could be the inputs to be aggregated using highly efficient computational methods such as cloud computing in the data-mining framework. Third, through the methodological investigations and analyses using empirical data, the signal-to-noise ratio (SNR) could be increased, which means greater and more accurate classification with high sensitivity to different aberrant testing behaviors. Fourth, the performances of the different methods (e.g., item response based person-fit analysis, response time based fraud detection methods, K-means clustering, supported vector machine, random forest, finite mixture modeling, neural networks) in detecting different types of aberrant behaviors such as pre-knowledge cheating and copy cheating in terms of classification accuracy would be further manifested by this study. Also, this study could edify the research community by comparing cheating behavior de-

tection accuracy rate by incorporating both biometric and psychological information with results from the traditional aberrant test-taking detection methods. Moreover, new models, such as gaze fixation counts based models or mixed effects models for jointly modeling the eye-tracking data with traditional psychometric item response models would be proposed. Additionally, some machine learning methods yield results that can be interpreted in a straightforward manner, which could then be communicated and perhaps for easily understood by stakeholders. With high quality measures such as eye-tracking data and indexes computed from the psychometric model, the rates of false negatives could be better controlled.

Given these potential impacts on the enhancement of security of high-stake assessment, this dissertation begins with an overview of the various methods on aberrant behavior detection, including traditional item response and response time based methods as well as a set of data mining methods, followed by the introduction of eye-tracking technologies. Chapter 3 describes the detailed research design and the proposed methods. Results from the current study are presented in Chapter 4. Implications, limitations, and future directions of the present study are discussed in Chapter 5.

Chapter 2: Literature Review

Chapter 2 begins with a brief overview of recent developments in test security. Many item response and response time based detection methods along with their unique features will be discussed. Next, this chapter introduces various data mining methods, which could be utilized for test security purposes. In addition, different types of eye-tracking measures will be introduced. Ultimately, jointly modeling distinct latent constructs and their potential applications for incorporating biometric information will be reviewed.

2.1 Test Security

Maintaining the security and confidentiality of student tests in any assessment program is critical for ensuring valid test scores and providing standard and equal testing opportunities for all students. Cizek (1999) made 18 recommendations for ensuring test security on large-scale educational achievement testing programs. His recommendations centered on how to establish rigorous standards for test safety and administration procedures (e.g., seal each test booklet, preventing test administrators from accessing the test booklets prior to the test). These recommendations provide a concrete foundation for further discussions on the problem of test security.

According to Wollack and Fremer (2013), test security generally requires policies, procedures, and guidelines involving the following characteristics:

1. Test safety should be guaranteed. Steps to ensure this might include the sealing of individual test booklets prior to delivery and pre-inspecting examination locations.

2. Test administrations must be strict, meaning examination instructions and policies must be made clear to all test takers.

3. Potential cheating behavior, before, during, and after test administrations should be carefully monitored and prevented.

4. Test administrators should be provided training to appropriately proctor and report potential cheating behaviors.

The ultimate purpose of establishing test security is to ensure test results that are valid and accurate for assessing the performance of examinees.

Test security is vital to maintaining the fairness of the test for all test takers. Without test security, the validity and reliability of the test scores would be questioned for all test takers, regardless of whether they engaged in cheating behaviors (e.g., Hendrawan et al., 2005; Nering & Meijer, 1998). Insecure test environments also involve problems of ethical issues. Test takers are more likely to cheat, especially in high-stake tests, when the test is not secured (Wollack & Fremer, 2013). The insecure test can also negatively impact the measurement classification validity (Hendrawan et al., 2005). Hendrawan et al. (2005) indicated that some misfitting of response patterns caused by cheating behaviors, such as pre-knowledge cheating and cheating by copying answers could result in bi-

ased estimates of item parameters. Sotaridona, van der Linden, and Meijer (2006b), who conducted a similar study focusing on this issue, found that item difficulty and discrimination parameters were consistently larger, and that standard errors of estimates were larger when the cheating behavior occurred. Hendrawan et al. (2005) also showed that the misfitting response patterns caused by cheating behavior lead to inaccurate mastery classification decisions. These studies have shown that the validity and reliability of the test results are sabotaged if test security is not maintained. Test security is not only about pre-administration prevention and the detection of aberrant testing behaviors that occur during the test itself, but this term also encompasses training professional and ethical administrators and exam supervisors to implement the test. According to Caveons webinar (Schoenig, Geraets, & Mulkey, 2016), test security has a conceptual framework. Before test administration, exam booklets should be acquired and distributed in a safe and confidential way. During the test administration, supervisors should proctor examinees closely and report aberrant behaviors honestly and promptly. At the completion of the assessment, test materials should be managed appropriately and undisclosed for preventing changing item responses. Wollack and Fremer (2013) described unethical behaviors by overseers of test administration (teachers or exam supervisors) such as failing to report violations and giving extra help during test administrations. Such actions could undermine ethical standards, which could have a corrupting influence on professional conduct and damage institutional confidence in using test results to make high-stake decisions. Therefore, maintaining test fairness is absolutely necessary for correctly interpreting test scores and for preserving the integrity of the testing regimen that organizations rely upon to make critical decisions.

2.2 Statistical Methods for Identifying Aberrant Testing Behavior

In recent years, a number of statistical methods have been developed for detecting different types of aberrant test-taking behaviors. The Council of Chief State School Officers (CCSSO) and Association of Test Publishers (ATP) have suggested the following guidelines for searching for misconduct (see Table 1.1). Based on this guidance, specific indices are suggested for detecting aberrant testing behavior that aligns to the different testing environments.

2.2.1 Similarity Analysis (Collusion, Answer-Copy)

Detecting collusion and answer copying has become an essential issue in high-stakes testing. In order to maintain the fairness of the test, many scholars are developing a variety of methods to detect and prevent answer copying and collusion behaviors among test takers. Wollack (2011) indicated the following types of test collusion: (1) illegal coaching by a teacher or test-prep school, (2) examinees accessing stolen test content posted on a study forum, (3) examinees copying answers from each other during an exam, (4) examinees harvesting and sharing exam content using e-mail or internet, and (5) teachers or administrator changing answers after the test has been administered. Based on these behavioral features, many methods have been proposed to identify examinees who engaged in colluded behavior (Bay, 1995; Belleza & Belleza, 1989; Cody, 1985; Frary, Tideman, & Watts, 1977; Hanson, Harris, & Brennan, 1987; Holland, 1996b; Sotaridona & Meijer, 2002a, 2003; Wollack, 1997).

In general, response similarity analysis attempts to calculate the likelihood of agree-

ment between two response vectors (Zopluoglu, 2016). Some of these methods focus on matching incorrect responses between two response vectors; some use both matched incorrect and matched correct responses as evidence of collusion. The following section provides a historical and technical overview of the methods used for detecting unusual response similarities.

2.2.1.1 Identical Errors Analysis

Bird (1927) derived an empirical null distribution of the number of identical errors by randomly pairing test takers across different locations. The distribution of the number of identical errors for each pair was used as a norm for a specific test. To determine who the cheaters were, a cut-off value of the mean plus one standard deviation of the distribution was used. If the number of identical errors within any pair who were taking the specific exam was larger than the cut-off value, they are flagged as cheaters for having an unusual degree of agreement.

This method laid the foundation of many similarity analyses used in educational testing. It can be easily implemented on a variety of tests in different formats. However, its weakness is that it does not include a general index to flag cheaters, and the cut-off value varies across different tests.

Many other scholars have tried different methods to improve this work (Angoff, 1974; Crawford, 1930; Dickenson, 1945; Saupe, 1960). These methods were based on the idea of using empirical distributions and have been relatively under-researched. Zopluoglu (2016) provided three reasons for this: (1) lack of access by researchers to a

large-scale datasets, (2) limitations of computational power, and (3) null distributions were exam-specific, and thus could not be generalized to other tests in a straightforward manner.

2.2.1.2 Error Similarity Analysis

The error-similarity analysis index (Belleza & Belleza, 1989) based on tests using multiple choice items was proposed for detecting test collusion by analyzing the probability of choosing the same series of incorrect alternative choices for every possible pair of students. The index is defined as follows:

$$\frac{N!}{k!\,(N-K)!}P^k(1-P)^{N-k}, \tag{2.1}$$

where P stands for the probability of any two students choosing the same wrong distractor in a multiple choice question. This value was assumed to be 0.4 since it is not possible to expect that all incorrect alternatives will have the same likelihood of being selected. N signifies the total number of items in the test, while k is the number of items that received the same incorrect answers.

The probability of choosing the same incorrect distractor by chance can be calculated for each pair of students. If there are S students, then the number of comparisons is $S(S-1)/2$. The probability for each pair of students is used to determine the probability of collusion behavior occurring by chance. When the sample size is large, the probability of collusion could follow the normal distribution with mean of NP and standard deviation of $\sqrt{NP(1-P)}$. Test takers who have possible error-similarity scores above two standard deviations from the mean on the distribution of error-similarity scores would be consid-

ered people who had colluded. For example, the probability of two students choosing the same wrong answer is assumed to be 0.4. The probability of choosing the same five answers out of 10 wrong items is $10!/(5!5!)0.4^5(0.6)^5 = 0.20065$. By calculating all possible pairs, those who have high error-similarity scores that are two standard deviations above the mean would be flagged.

This line of research has several limitations. First, the value of P is assumed and fixed due to the reality that it is not possible to expect with equal likelihood that all incorrect alternatives will be selected. Second, this procedure requires the use of the binomial distribution in order to obtain valid results; the sample size, therefore, needs to be sufficiently large. Third, the error-similarity scores are test-length dependent, making it difficult to compare across tests with different number of items. All of these methods are specifically designed for paper-pencil tests due to the limited computational power at the time that the research was carried out. However, the advantage of this method is that it is easy to understand and calculate, especially when a computer is available for analyzing the data.

Based on this work and that of (Saupe, 1960), a series of indices were derived: K_1, K_2, S (Holland, 1996b; Sotaridona & Meijer, 2002b). Instead of using a fixed value of P (the probability of any two students choosing the same wrong distractor in a multiple choice question), these other indices used linear, quadratic regression equations, or a log-linear model, respectively to predict P.

2.2.1.3 Generalized Binomial Similarity

The previous reviewed studies only focused on identical incorrect answers. W. J. van der Linden and Sotaridona (2006) proposed a generalized binomial test method that used a compound binomial distribution for the number of identical incorrect and correct responses between any pair of examinees. The formula is as follows

$$P_{M-c} = \sum_{o=1}^{O} (P_{ico} \times P_{jco}), \tag{2.2}$$

where P_{M-c} denotes the probability of matching for the ith and jth examinees. The probabilities of selecting the oth response alternative of the cth item for examinees i and j, are P_{ico} and P_{jco} , respectively.

The probability of observing m matches on C items between two-response vectors is computed as

$$f_C^m = \sum \left(\prod_{c=1}^{C} P_{M-c}^{\mu_c} (1 - P_{M_c}^{1-\mu_c}) \right), \tag{2.3}$$

where μ_c is an indicator of whether or not a pair of examinees has the same responses to item c. The summation is across all the possible combinations of n matches on C items. The upper tail of this compound binomial distribution is used as the cut-off value to reflag the people who have some degree of agreement.

For the methods previously described, the binomial distribution is the key component to be utilized to flag people who are potentially copying from each other in the similarity analysis. Usually, Type I error rate and power are utilized to evaluate the performance of these proposed indices. The Type I error rate indicates the probability of an honest test-taker being incorrectly detected as a cheater while power represents the proba-

bility of accurately detecting pairs who colluded on the test. There some other variations, such as the ω (Wollack, 1997) and K statistics, based on the generalized binomial test method (W. J. van der Linden & Sotaridona, 2006). Based on a simulation study (Zopluoglu & Davenport, 2012), the K index yielded high power on detecting copy cheating, and is used by some large testing companies such as Educational Testing Service (ETS) and the College Board. All of the proposed similarity analysis methods could be utilized on paper-pencil, computer-based, and internet-based tests. For the computer adaptive testing, there is a lack of literature discussing the similarity analysis methods on the copy-cheating detection, which leaves opportunities for future research. In contrast to comparing item responses from examinees to detect cheating behavior, some analyses have relied on the error similarity analysis.

2.2.2 Gain Score Analysis

Cannell (1988) questioned the integrity of achievement gains being made across all states on norm-referenced tests. He pointed out that achievement gains usually occurred as a result of unethical teaching practices at the local level, which consequently obscured test security measures at the higher level (Cannell, 1988). There is some security-related research that followed Cannell's report focused on casual factors analysis (e.g., Shepard, 1990; Stonehill, 1988). An example of causal factors analysis demonstrated how teaching to the test could result in increased gains in standardized test scores (Shepard, 1990). However, it took more than a decade to consider using statistical methods for detecting unexpected score gains in the large-scale assessment (Bishop, Liassou, Bulut, & Seo,

2011).

Jacob and Levitt (2003, 2004) proposed a method for using unexpected gain scores to detect teacher cheating. They used a method that combined two indicators: (1) unexpected test score fluctuation and (2) unusual student response patterns. The authors applied a simple method by ranking the classroom-level gain scores and comparing the rankings of all the classes across two time points. If scores of some classes increased unexpectedly, the answer keys of students were checked. The presence of both indicators suggested aberrant testing behavior, such as students receiving help during the test. Consequently, both students and teacher would be flagged as cheaters. Other analyses follow a similar idea of comparing the cumulative distribution function (CDF) of total scores or scaled scores between two yearly assessment periods (Ho, 2008). If the difference between two CDFs from two years was large, or the high score range had a negative difference, or the percentage of high scores in the second year had decreased relative to the first year, then some unexpected gains between the two time points was indicated (Ho, 2008). However, no universal cut-off value to be applied for these methods currently exists and thus is rather subjectively applied.

2.2.3 Erasure (Answer Changing) Analysis

Qualls (2001) reported that students rarely erase their original responses. Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) found that erasures occur in roughly one out of every 50 items. Additionally, other studies indicated that erasures occurred with increased frequency on more difficult items and that test-takers with middle to high abilities

were more likely to change their answers (Mroch, Lu, Huang, & Harris, 2014; Primoli et al., 2011). Based on these studies, several methods were proposed for uncovering suspicious answer changing. However, the erasure analysis is more focused on the group level, such as classrooms and schools instead of individual level (Primoli et al., 2011). Individual-level behavior usually was checked by other kinds of analyses, such as similarity analysis and person-fit analysis. In erasure analysis, wrong to right (WR) change is frequently used for statistical modeling. The most straightforward analysis is the group-level *Z*-test (Bishop et al., 2011). For example, WR counts for a specific class or a school is tested against the state population mean. The group would be flagged if the test statistic was larger than a critical value with certain significant level. Bishop et al. (2011) also proposed a simple linear regression method that used the mean of the total class erases (TE) as a predictor with and the mean of the class WR as the outcome variable. The authors found this method could account for a substantial proportion of variance at the group level. However, this method suffers from heteroscedastic behavior as conditional WR variance increases as the TE sum increases (Bishop et al., 2011). In order to overcome this problem, Poisson regression was used and resulted in a better fitting model. The groups that have either linear regression or Poisson residuals greater than 1.96 were considered to be suspicious answer changing groups (Bishop et al., 2011). later this method was extended to a hierarchical linear modeling framework that used a two-level random intercepts model within schools and a two-level random slope model that regressed the number of WRs on individual-level ET counts. They concluded that both models were more appropriate and fit significantly better than the simple linear regression method. Ninety-five percent confidence intervals were constructed for each schools slope and were

flagged if the school interval did not cover the overall mean slope. Like a gain score analysis, an erasure analysis is a budding research area. Erasure analysis, however, is more focused on identifying aberrant testing behavior of groups rather than of individuals.

2.2.4 Person Fit Analysis

Many person-fit indices have been created for detecting aberrant test-taking behaviors. Person-fit indices are computed to assess different response patterns of test-takers, focusing on flagging copy-cheating and also on detecting other types of behaviors such as pre-knowledge cheating and item stealing. The use and effectiveness of person-fit indices used for copy-cheating detection is understudied compared to other analytic methods (Cizek & Wollack, 2017). However, some proposed person-fit indices have shown a high degree of power to detect certain aberrant testing behaviors based on examinees response patterns. For example, the H^T index (Sijtsma & Meijer, 1992; Sinharay, 2018) has been shown to be effective in detecting pre-knowledge cheating with high power. The central premise behind most of the person-fit indices is to check whether or not a vector of item responses is aligned with a person's latent ability that is estimated from a specific item response theory (IRT) model. Simply speaking, the probability of the reappearance of a specific item response vector given the person's latent-ability estimate from the IRT model that we used is evaluated. Representative person-fit statistics used in different testing environments are presented in the Table 2.1. There are two primary classifications of these indices: parametric and nonparametric. Some of these methods will be subsequently introduced.

Table 2.1: Representative person fit indices

Rash model	**RT based**
U (Wright & Stone, 1979)	PPMC Bayesian approach (van der Linden & Guo, 2008)
W (Wright & Masters, 1982)	l^t(Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014)
2PL and 3PL	**Guttman-based index**
l_z (Drasgow, Levine and Williams,1985)	G (Guttman, 1944)
l_z^* (snijders, 2001)	**Agreement Index**
l_s (Sinharay, 2017)	A (Kane & Brennan, 1980)
CAT	**Group-Based Index**
K (Bradlow, Weiss, & Cho, 1998)	r_{pbis} (Donlan & Fischer, 1968)
T (van Krimpen-Stoop & Meijer, 2000)	H^tindex (Sijtsma, 1988;Sijtsma & Meijer, 1992)

2.2.4.1 Representative Parametric Indices

U Index. The U index is computed from performing a residual analysis from applying the Rasch model (Rasch, 1961) to a set of examinees item responses (Wright & Stone, 1979). The Rasch model is the simplest of IRT models in that it is parameterized by a single, difficulty parameter. As a consequence of this parsimoniously parameterized model, analyses require relatively small sample sizes (i.e., the number of examinees) to produce reasonable data-model fit (Linacre, 1994). Since residuals are at the heart of the U index computation, the U statistic could alternatively be calculated using other IRT models such as 2PL or 3PL models; however, the performance of U under these extensions has not been thoroughly studied. The computation of the U person-fit index follows

$$U = \sum_{i=1}^{I} \left(\frac{[X_i - P_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]} \right), \tag{2.4}$$

where $P_i(\theta)$ is the probability of correctly answering item i given the ability estimate, θ, X_i is the dichotomous response (0, 1) of item i for a specific person. This index follows a Chi-square distribution with I degrees of freedom. Test-taking aberrances would be flagged by a critical value of the distribution at a certain significance level (α).

There are other indices based on U that have been proposed in the literature, such as the ZU, ZW and UB indices. These indices are either normalized versions of original indices with different methods such as ZU and ZW or applied weighting strategies for different items rather than treating them equally. Results from a comprehensive simulation study (Karabatsos, 2003) showed that U performed better than ZU, ZW and UB in detecting aberrant test-taking behavior of examinees.

Likelihood-Based Indices. The likelihood-based analysis of examinees response patterns results in an index called l_z^* (Snijders, 2001). The l_z^* (Snijders, 2001) statistic is the asymptotically correct standardized version of the l_z statistic (Drasgow, Levine, & Williams, 1985). Both of these statistics were based on the l_z^* statistic defined as the loglikelihood of the item level scores for a test taker

$$l = \sum_{i=1}^{I} [Y_{ij} log P_i(\theta_j) + (1 - Y_i) log(1 - P_i(\theta_j)], \tag{2.5}$$

where Y_{ij}, a random variable, denotes test taker $j's$ response (0 or 1) to item i. And, $P_i(\theta_j) = P(Y_i = 1)$ is the probability of a correct answer for test taker j on item i. Then

the expected value of the loglikelihood and its variance can be computed as

$$E(L|\theta) = \sum_{i}^{I} \left\{ P_i(\theta_j) log \frac{P_i(\theta_j)}{1-P_i(\theta_j)} + log(1-P_i(\theta_j)) \right\}, \tag{2.6}$$

and,

$$Var(L|\theta) = \sum_{i}^{I} P_i(\theta_j)(1-P_i(\theta_j) \left[log \frac{P_i(\theta_j)}{1-P_i(\theta_j)} \right]^2. \tag{2.7}$$

In general, the l statistic indicates that how likely it is to capture person j's response pattern under a fitted IRT model. Using the l index as a starting point, Drasgow and colleagues (1985) provided a standardized version of l index in the usual manner by subtracting the expected value of the right hand expression in Equation 2.6 above and dividing by its standard deviation.

$$l_z = \frac{l - E(L|\theta)}{\sigma(l|\theta)}. \tag{2.8}$$

The statistic in Equation 2.8 follows a standard normal distribution. In practice, unknown θ would be replaced by an estimate, $\hat{\theta}$. With this substitution, the value of l_z decreases with increasing degree of person misfit and large negative values of the index (i.e., those smaller than -1.645 at the 5% significance level) are indicative of aberrant response behavior. Although the sampling distribution is standard normal asymptotically, one limitation to l_z is that this index is not valid when true abilities are replaced by sample ability estimates (W. Molenaar & Hoijtink, 1990; Reise, 1990). To correct this issue, (Snijders, 2001) proposed a slight modification of the index to obtain the desired asymptotic distribution with sample estimates of ability instead of the true (unknown) values and was shown to work for different estimates of and under different IRT models (Magis, Rache, & Bland, 2012). In essence, the revised index of l_z^* (Snijders, 2001) modifies both the

expectation function and the variance function in Equation 2.8 by taking into account the sampling variability of $\hat{\theta}$.

The l_z^* index based on a 3PL model is computed as

$$l_z^* = \frac{l - E[l|\theta]}{\sqrt{\sum_i^I \left\{ ln\left(\frac{P_i(\theta_j)}{1-P_i(\theta_j}\right) - c_I(\theta_j) r_i(\theta_j)) \right\}^2 P_i(\theta_j)(1-P_i(\theta_j))}}, \tag{2.9}$$

where

$$c_I(\theta_j) = \frac{\sum_i^I P_i'(\theta_j) ln\left[\frac{P_i(\theta_j)}{1-P_i(\theta_j)}\right]}{\sum_i^I P_i'(\theta_j) r_i(\theta_j)}, \tag{2.10}$$

and,

$$r_i(\theta_j) = \frac{a_i exp[a_i(\theta_j - b_i)]}{c_i + exp[a_i(\theta_j - b_i)]}, \tag{2.11}$$

and where, $P_i(\theta_i)$ is the first derivative of $P_i(\theta_j)$ with respect to θ_j, and a_i, b_i, and c_i represent item discrimination, difficulty and pseudo-guessing parameters, respectively. The l_z^* statistic simplifies when a 1PL or 2PL IRT model is used instead. Additionally, the l_z^* statistic is still compared to a standard normal distribution to evaluate whether the test taker is aberrant or not (see, e.g., Magis et al. (2012) for a useful flow chart for practical implementation of l_z^*).

Person Response Function Analysis. Trabin and Weiss (1983) proposed the *D* index by utilizing the person response function (PRF; Weiss, 1973) to identify misfitting item-score patterns. The person response function is a non-increasing function of the item difficulty parameter. The *D* index was intended to compare the difference between the expected PRF function, based on a certain IRT model, and the observed PRF function. A significant difference between the expected and observed PRF would be indicative of a misfitting response for that examinee (Trabin & Weiss, 1983).

The k items are ordered according to their difficulty parameters. Then, the k items are assigned into S ordered subsets. Each subset contains m items, $A_1 = \{1,2,...,m\}$, $A_2 = \{m+1,...,2m\}$,...,$A_s = \{m+k-m+1,...,k\}$. The expected PRF is constructed as an estimate of the expected proportion of correct responses under a certain IRT model in each subset, and is calculated as: $m^{-1}\sum_{g\in A_s} X_g$, $s = 1,2,..,S$. The observed proportion is computed as $m^{-1}\sum_{g\in A} X_g$, $s = 1,2,..,S$. The difference between expected and observed PRFs is then computed as $D_s(\hat{\theta}) = m^{-1}\sum_{g\in A_s}[X_g - P_g(\hat{\theta})]$, $s = 1,2,\ldots S$. By taking summation across all the subsets, the D index is computed as

$$D(\hat{\theta}) = \sum_{s=1}^{S} D_s(\hat{\theta}). \tag{2.12}$$

Based on the simulation study conducted by Karabatsos (2003), when the cut-off value of the D index is equal to 0.55, the detection rate is maximized. Meanwhile, Karabatsos also indicated that the D index provides the best performance on pre-knowledge cheating among all of the parametric IRT-based indices.

2.2.4.2 Representative Non-Parametric Indices

Non-parametric person-fit statistics are defined as those that do not depend on any IRT model (implicit in this definition is the fact that an IRT model stems from a particular distribution). There are several reasons why non-parametric PFSs are popular alternatives to the more conventional parametric PFSs. First, non-parametric IRT-based indices are relatively easy to compute since they do not reply on any parametric IRT model. Computation of parametric person-fit indices involves maximum likelihood or Bayesian estimation, which can be computationally demanding compared to their non-parametric

counterparts, given characteristics of the testing situation. Second, nonparametric person-fit indices yield relatively consistent results compared with parametric person-fit indices, which often give different results depending on the kind of IRT model used. Several nonparametric indices, such as the *G* index, the norm conformity index, and the group-based index will now be presented.

Guttman Scale Index. The Guttman-based index (Guttman, 1944) or *G* index measures the degree of reasonableness of an examinee's answers to a set of test items (Karabatsos, 2003; Meijer, 1994). It was the first nonparametric person-fit index developed for detecting aberrant test-taking behavior. The way to flag test-taking aberrances is to count the number of Guttman errors. Let $P_m(m = 1, ..., I)$ denote the proportion of persons who respond correctly to item m. Assume that there are I ordered items according to P_m in a test such that $P_m \geq P_n$ $(m = 1, ..., I-1; n = m+1, ..., I)$. Then, the G index is calculated as

$$G = \sum_{m=1}^{I-1} \sum_{n=m+1}^{I} I_{mn}, \tag{2.13}$$

where I_{mn} is an indicator taking on the value of 1 if a person has a Guttman error on items m and n (i.e., I_{mn} = 1); otherwise, I_{mn} = 0.

The *G* index excludes all of the item-score combinations $(0,1)$, which are called Guttman errors. A Guttman error means the examinee answered correctly on a relatively difficult item m and answered incorrectly on an easier item n according to the Guttman scale, which orders the items from hardest to easiest. The permitted item-score patterns are $(0,1), (0,0)$, and $(1,1)$. Table 2.2 demonstrates the calculation process for six hypothesized examinees.

Table 2.2: Guttman scale and index calculation

Examinee	Item 1 (Hardest)	Item 2 (Moderate)	Item 3 (Easiest)	Responses Pairs	G
1	1	1	1	(1,1)(1,1)(1,1)	0+0+0
2	0	1	1	(0,1)(0,1)(1,1)	0+0+0
3	0	0	1	(0,0)(0,1)(1,1)	0+0+0
4	0	0	0	(0,0)(0,0)(0,0)	0+0+0
5	1	0	1	(1,0)(1,1)(0,1)	1+0+0
6	1	1	0	(1,1)(1,0)(1,0)	0+1+1

Guttman Scale based Norm Conformity Index. As its name suggests, the Norm Conformity Index (NCI; Tatsuoka & Tatsuoka, 1983) measures the extent of conformity or consistency of an individual test takers response pattern on a set of items and is defined as

$$NCI = 1 - \frac{2\sum_{i=1}^{I-1}\sum_{s=i+1}^{I} y_i(1-y_s)}{r(I-r)}, \tag{2.14}$$

where the realization of a test takers response (0 or 1) to item i is denoted as $(i = 1,..,I)$ across all items and all the items are ranked from low to high based on their difficulty levels, in which item s is more difficult than item i. The numerator in Equation 2.14 represents the total number of Guttman conformal pairs, which only allow a test taker to have two types of response pairs: (1) a relative easy item is answer correctly first and a more difficult item is answered incorrectly later, which is denoted as (1, 0), or (2) a pair of easy and hard items is answered correctly, which is denoted as (1, 1). Variable r is the unweighted total score for a test taker ($r < 1$). More specifically, NCI measures the proximity of the pattern to a baseline pattern in which all 0's precede all 1's when the items are arranged in a pre-designated order (e.g., conforming to a Guttman scale)

Non-Parametric Transposed Scalability Index. The H^T index (Sijtsma, 1986; Sijtsma & Meijer, 1992) is a nonparametric statistic which is the transposed formulation of the scalability coefficient, H^T (Loevinger, 1948) for items. The H^T index for person n is defined for a complete rectangular dataset of dichotomously scored items where the rows represent I test takers and columns denote the J items as,

$$H^T(n) = \frac{\sum_{m=1,m=n}^{J}\left(\left[\sum_{i=1}^{I} y_{ni}y_{mi}/(1-p_np_m)\right]\right)}{\sum_{m=1,m\neq n}^{J}\left(min\left[p_n(1-p_m), p_m(1-p_n)\right]\right)}. \tag{2.15}$$

where y (either 0 or 1) is the proportion of items answered correctly by test taker i, p_m denotes the proportion of items answered correctly by test taker m. The H^T index is the sum of the covariances between test taker n and the other test takers divided by the maximum possible sum of those covariances. This constricts the range of allowable values to be between -1 and 1. In scenarios in which the responses of test taker n are random, the value of $H^T(n)$ will be close to zero. In a similar manner, responses that are positively correlated with all other test takers would result in $H^T(n)$ taking on a positive value, while $H^T(n)$ would take on negative values for the situation when test takers responses are negatively correlated with other test takers. When the data are fit to the Rasch model, $H^T(n)$ is expected to be somewhat positive (Sijtsma & Meijer, 1992).

2.2.4.3 Representative Response Time based Index

Although indices based on item responses have been shown to be somewhat effective in uncovering particular types of fraudulent testing behaviors, they do have some limitations. Due to the simple structure of the item response sets, often comprised of

0s and 1s, test takers could imitate normal testing behavior, thus reducing the ability of these IRT-based methods to identify actual cheaters. This may be especially true for detection of minor anomalies (e.g., cheating occurring on certain items) for which methods that possess greater sensitivity are needed. Indices based on examinees response times represent one such class of methods that use additional testing behavior of the examinees above and beyond their response profiles. Models for response times (see, e.g., Thissen, 1983; van der Liden & Sotaridona, 2006) were introduced to examine the identification of cheating behaviors differently than using IRT-based methods. One such method devised by van der Linden & Guo (2008) proposed assessing aberrant behaviors by examining response time-based residuals, which the authors defined as the difference between actual and predicted RTs of a test takers answers using a method of crossvalidation. More recent, Marianti and colleagues (2014) suggested

$$l_t = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{log(T_{ij}) - (\zeta_i - \tau_j)}{\sigma_{e_i}}, \tag{2.16}$$

where T_{ij} is the response time for test taker j on an item i, ζ_i is the time-intensity parameter that is the averaged population time required for answering that item, τ_j is the speediness parameter for each test taker, and e_{ij} is the residual term of log response times. All the parameters are estimated based on the RT model proposed by van der Linden (2006), which is

$$log(T_{ij}) = \zeta_i - \tau_j + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2). \tag{2.17}$$

These methods—those using IRT modeling of item responses and those using response time data—have been useful in detecting various aberrant testing behaviors with varying degrees of success. However, these conventional methods of aberrant test tak-

ing behavior detection have several limitations. One limitation of traditional methods is that they are not well-equipped to integrate the vast amount of process data collected as a natural byproduct of computer-based or computer-adaptive testing environments. Data coming from log files as well as other test-taker characteristics are continually generated and recorded at regular intervals during an assessment administration and may very well communicate useful and diagnostic evidentiary information to help uncover patterns of aberrant behaviors. Also, each method is used in isolation from one another. Treating aberrant test-taking behavior detection in this manner does not exploit the potential benefit that aggregating such information across methods might reveal.

2.2.5 Use of Data Mining Methods to Detect Test Fraud

Mining response data to identify clusters of respondents (e.g., such as those who exhibit fraudulent test taking behavior) is not a new idea in assessment research. In their paper detailing the facets of data mining, Romero, Gonzlez, Ventura, Del Jess, and Herrera (2009) explained that one must use a data mining strategy that is appropriate for the type of data one wishes to identify, such as data mining to identify patterns of behaviors. Their explanation indicates that data mining can facilitate the identification of cognitive and behavior processes (Berkhin, 2006), and pertinent to the current study, aberrant test-taking behaviors. According to Kerr and Chung (2012), identification of processes within response patterns is typically done with clustering algorithms, which can be classified for the purposes of the current study as either (1) unsupervised machine learning algorithms or (2) supervised machine learning algorithms.

Clustering algorithms represent a particular class of unsupervised learning algorithms that will be the primary focus in this dissertation. Clustering algorithms are processes that use observed similarities of densities in data to identify patterns and group similar observations (Berkhin, 2006). Three unsupervised learning methods will be investigated: (1) K-Means clustering, (2) multivariate normal mixture models, and (3) self-organization mapping. A category of supervised learning methods whose primary function is accurate classification will also be investigated. The approaches to be explored are: (1) K-nearest neighbor (KNN), (2) random forests (RFs), and (3) support vector machine (SVM). A description of each method is presented followed by some advantages and disadvantages of each algorithm as they relate to aberrant test taking behavior detection.

2.2.5.1 Unsupervised Machine Learning Methods

K-Means clustering. Although there are several versions of the K-Means algorithm, the current research advocates the version defined by Hartigan and Wong (2012), which is generally accepted as the preferred K-Means algorithm (Berkhin, 2006; R Core Team, 2014). K-Means clustering attempts to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean. The algorithm begins with a set of K potential centers which can be defined by the researcher or randomly selected from the data. The choice of initial cluster centers leads to a deterministic partitioning of the space. In other words, K-Means will always return the same clustering solution given the same initial cluster centers (e.g., Steinley, 1985). Since the clustering solution relies heavily on where the algorithm launches from, especially for small

datasets (Lattin, Carroll, & Green, 2003), some have argued that the algorithm be run multiple times from different starting values to ensure the efficacy of the classification (e.g., Celebi, Kingravi, & Vela, 2013; Khan & Ahmad, 2004).

Once the centers are selected, the algorithm assigns all the test takers to their closest centers and recalculates the new centers defined by these clusters. Distance is determined by a user-specified similarity measure – often Euclidean distance or Manhattan distance (Fossey, 2017). The algorithm goes through multiple iterations of checking each test taker (e.g., response pattern) to see if it should be moved to a different cluster based on the centers updated coordinates. If so, it changes the test takers cluster membership, updates the centers coordinates, and continues to the next iteration until it converges on a solution where no points are being switched between clusters.

K-Means clustering algorithm offers several main advantages and disadvantages on aberrant behavior detection. Theses includes:

1. K-Means algorithm is easy to implement. It only requires practitioners to specify number of clusters to initiate the algorithm. Usually, in test security investigation, we are expecting to separate aberrantly behaved test takers from the normal population. Thus, two underlying clusters could be reasonably assumed as the number of initial clusters. However, it could also be a disadvantage if users have limited information to determine number of clusters underlying the data.

2. K-Means algorithm could be computationally efficient with a high dimensional dataset. The algorithm relies on a nonparametric distance measure to classify observations consuming less computational memory than other parametric methods,

which requires estimation of model parameters (Hastie, Tibshirani, & Friedman, 2009). Recently, due to large volumes of process data generated during the computer based testing, K-Means could potentially be useful to analyze high dimensional data for flagging aberrant takers in real time. However, due to the nonparametric nature, K-Means algorithm is sensitive to the initial cluster centers. Many solutions have been proposed for dealing with this issue (Li, 2011). But, these extensions could potentially sacrifice certain degree of computation efficiency.

Finite mixture modeling (FMM). Model-based clustering method that might be useful in identifying aberrant test taking behaviors is finite mixture models, specifically mixtures of multivariate distributions. Mixtures of multivariate distributions (Everitt, 1981; Titterington & Makov, 1985) have been applied to a wide range of statistical methodology and take the general form

$$f(\boldsymbol{s}_j|\varphi,\xi) = \sum_{k=1}^{K} \varphi_k f_k(\mathbf{s}_j|_{,k}), \tag{2.18}$$

where a distribution f is a mixture of K component densities $f_1,...,f_K$ and $\mathbf{s}_j$ is a p-dimensional vector containing scores for individual j $(j = 1,\ldots,n)$ on a set of p observed continuous random variables. Vector $\varphi = (\varphi_1,\ldots,\varphi_{K-1})$' contains the mixing proportions with the caveat that $0 \leq \varphi_k \leq 1$ for all $k = 1,\ldots,K$ with $\sum_{k=1}^{K} \xi_k = 1$. Vector $\xi' = (\xi_1^{'},\ldots,\xi_{K-1}^{'})$ contains all unknown parameters in all K subpopulations, where $\xi_k^{'} = (\boldsymbol{\mu}_{\boldsymbol{k}}^{'}, vech(\Sigma_k)^{'}$. Operator $vech(\Sigma_k)$ denotes a half-vectorization of a symmetric matrix Σ_k by stacking only the lower triangular part of Σ_k. Following McLachlan, Peel, and Bean (2003), the kth component density of a mixture of multivariate normal distributions

is given by

$$f_k(\boldsymbol{s}_j|\boldsymbol{\xi}_k) = (2\pi)^{-p/2}|(\Sigma_k|^{-1/2} exp\left\{\frac{1}{2}(\boldsymbol{s}_j - \boldsymbol{\mu}_k)'\Sigma_k(\boldsymbol{s}_j - \boldsymbol{\mu}_k)\right\}. \tag{2.19}$$

One of the main advantages of using a finite mixture model is that an FMM would manifest hidden clusters embedded in the streams of data by using a likelihood ratio test (LRT: Cox & Hinkley, 1974) or information based model selection criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC: Anderson & Burnham, 2002). Thus, it would be useful to explore subcategories of aberrant testing behaviors rather than simply focusing on aberrant and normally behaved groups, which could potentially provide more insights for practitioners to understand and investigate on specific behavioral groupings. Also, by assuming a multivariate Gaussian density, FMM could reflect the volume, shape and orientation of each cluster by estimating their corresponding variance-covariance structures. This piece of information could be potentially utilized for understanding characteristics of each identified cluster. For instance, if the fitted Gaussian density contours of each cluster are relatively small and separate, this could be a strong evidence of existence of different clusters. Just the opposite would occur if the fitted density contours overlapped too much with each other, the final classification would be doubtful for making final decision of number of clusters. Yet, FMM is sensitive to violations of distributional assumptions and completely exploratory. If the observations do not follow Gaussian distributions, the power of identifying underlying clusters could be decreased.

Self-Organization mapping (SOM). The SOM algorithm (also known as a Kohonen Map) is an artificial neural network algorithm where multidimensional data is mapped

to a set of k clusters (or nodes). One of the primary reasons SOM is popular is that the clusters can be mapped to a two-dimensional grid that shows which clusters are similar to each other. This is a valuable tool of visualizing data and validating clusters (Berkhin, 2006).

The SOM algorithm starts with a large learning rate coefficient, which is used to shift the cluster centers for all clusters in a large neighborhood, surrounding the winning clusters center, As time (iterations) progress, the neighborhood around each cluster shrinks to zero so that nearby clusters are not modified when a cluster is updated, and the clusters themselves are not changed as much by the presentation of new test takers because the effect of these new test takers is weighted by a decreasing learning algorithm. This is useful in situations where the researcher presents the same cases to the SOM network over and over again to achieve a more stable estimate of cluster centers. The initial cluster changes are large, with cluster centers being moved substantially by new test takers and by changes in neighboring clusters. As the algorithm runs through its iterations, the learning rate coefficient and the size of the neighborhood shrink until eventually there are only minute, fine-tuning changes to the winning clusters center (Bullinaria, 2004).

The size of the neighborhood and the rate of decrease can be set by the researcher. The rate of decrease may be linear or nonlinear, and the neighborhood may exist for all of the SOM iterations, or it may be defined so that the neighborhood radius shrinks to zero after a set number of iterations have been completed. For example, in the default settings of the som package (Yan, 2016) in R (R Core Team, 2017) statistical software, the neighborhoods radius is chosen to be larger than 2/3 of the unit-to-unit distances for all of the starting cluster centers. The som package then linearly decreases the radius of the

neighborhood over 1/3 of the iterations chosen by the researcher (Wehrens & Buydens, 2007). If 2/3 of the starting cluster centers are 100 Euclidean distance units away from each other, and the researcher specifies 300 iterations, then the radius of the neighborhood will decrease by one unit at each of the first 100 iterations, after which only the winning clusters center will be updated. Once the neighborhood radius diminishes to zero, clusters near the winning cluster are no longer updated when cases are reassigned, and the SOM algorithm solution is then identical to the logic used by the K-Means algorithm (Kohonen, 1982).

SOM has several benefits for fraudulent testing behavior detection. First, it displays complex high-dimensional topological relations of the cluster centers in a two-dimensional grid, which could be easily visualized and interpreted for test security. Second, SOM does not rely on any assumptions about the distributions of the data, and the solutions are not heavily influenced by outliers (Wehrens & Buydens, 2007). This is because, unlike K-Means, SOM never calculates a cluster centers coordinates by taking the mean coordinates of all the test takers assigned to the cluster. Instead, the cluster centers are moved incrementally depending on the case considered at each iteration.

2.2.5.2 Supervised Machine Learning Methods

K-Nearest neighbor. K-nearest neighbor (KNN) is a nonparametric clustering approach representative of supervised learning algorithms and was first proposed by Fix and Hodges (1951). KNN is a straightforward algorithm that attempts to classify new samples (unlabeled observations) by allocating them to the class of the most similar la-

beled cases by training the machine to learn a function thereby capturing the relation between the labeled outcome variable and independent variables. The algorithm starts by specifying the size of the neighborhood (K) of a data point by using a distance measure such as Euclidean distance, Manhattan distance, Murkowski distance and Hamming distance. The choice of K has a significant effect on the KNN results. When K is small, the classification decision would be less stable, and the boundary of separating the different groups would be less linear (James, Witten, Hastie, & Tibshirani, 2013). As K increases, the classification results would be more stable, and the classification boundary would be more linear, which leads to low within-group variance but high classification bias (James et al., 2013). However, this parameter could be tuned to optimize the classification results. Also, K is usually an odd number. Once K is specified, the KNN classifier would identify the K points, which are adjacent to a test observation (a new data point) in the training dataset by computing the defined distance between them by looping through the entire dataset. The conditional probability for the test observation belonging to a certain class would then be estimated. Finally, the new data point would be allocated to the class with the largest probability. The process would be continued until the last test observation is assigned. Many *R*-based KNN packages have been created for running the KNN analysis such as KernelKnn (Mouselimis, 2018), care package (Kuhn, 2017) and class package (Ripley, 2018). In the current study, the *knn* function from the class package was selected because, (1) this package is one of most well-accepted and tested packages for KNN algorithms, and (2) it is also very user-friendly with detailed instructions and documentation that appear in many data mining training websites.

KNN shares many similar advantages as K-Means algorithm, such as simplicity

and flexibility. In addition, many studies have shown that the KNN method is effectively robust to noisy training data if the training data set is large enough (e.g. Imandoust & Bolandraftar, 2013; Weinberger & Saul, 2009). Therefore, KNN has the potential to generate a stable mapping function, which could be utilized for making accurate and steady classification by limiting the influence of potential outliers. However, it suffers from its own limitations. KNN is sensitive to redundant and similar features, which could reduce the classification accuracy (Qian, Yao, & Jia, 2009). In addition, the algorithm has high computational cost if the training dataset is large due to calculating distance of each query to all other inputs in the training dataset (Imandoust & Bolandraftar, 2013).

Random forests. A random forest (RF), a representative ensemble method proposed by Breiman (2001), builds a set of classification and regression trees (CART) to make predictions by aggregating predicted results from each classification tree. CART, a nonparametric method, recursively segregates the feature space (an n-dimensional vector space associated with all the predictors) into many small rectangular areas. The CART algorithm splits predictors in a binary manner meaning each split in the tree-building process only generates two sub-nodes from a parent node. In each sub-node, subjects sharing more homogeneous properties are grouped together. This partitioning process, also called impurity reduction, minimizes the difference between the averaged impurity in the sub-nodes and the impurity in the parent node. Several entropy measures, such as the Gini index, are used to measure the impurity in each sub-node. Each node would be continually split until some stopping conditions are achieved. Commonly used stopping rules of the algorithm include (1) the minimum size of subjects left in a node, (2) a minimum change in the impurity measure after a split, and (3) information criterion such as AIC or

BIC. After a tree is built, a finalized classification of all the subjects would be predicted in each terminal node. For the RF method, a set of CARTs is built instead of using a single tree to make prediction. The rationale for this is that a classification prediction based on a single tree would be unstable. For example, if the first splitting variable were chosen differently, the predicted results would be potentially altered especially with a large number of predictors. Moreover, for the RF algorithm, a predictor at each node is randomly selected from the entire feature space for splitting the trees.

In each step of the RF algorithm, either a bootstrap sample or a subset of the entire dataset is randomly selected. Thus, by building a diverse set of tress serving as a voting committee would yield more stable and unbiased classification prediction than using a single tree. Voting here means the final prediction is achieved by averaging (weighted or unweighted) the predicted result from each tree. Many other aggregated methods have been developed such as Behavior Knowledge Space (BKS) Method (Y. S. Huang & Suen, 1995), Naive Bayes (NB) combination (Domingos & Pazzani, 1997) and Decision Templates (Kuncheva, Bezdek, & Duin, 2001). Choice of aggregation method notwithstanding, the ensemble voting method would produce more accurate predictions than using a single tree (e.g., Bauer & Kohavi, 1999; Breiman, 1998; Dietterich, 2000). The prediction accuracy could also be checked by an index known as the out-of-bag error rate (Breiman, 1996). Since each tree is built based on either a bootstrapped sample or randomly formed subset of the original dataset, the samples are retained so that tree building could be utilized for checking the prediction accuracy. The advantage of using an out-of-bag error rate is that it is a relatively more conservative and precise estimate of the error rate that is closer to the true classification in the population than the overly optimistic re-

sult from the prediction by using the original dataset (e.g., Boulesteix, Strobl, Augustin, & Daumer, 2008; Breiman, 1996). Many R packages have been created for implementing RFs algorithms such as Rpart (Therneau, 2018) and tree (Ripley, 2018) and randomForest (Breiman, 1996). In this study, randomForest is used for conducting the analysis.

RF algorithm has many advantages. Unlike other supervised learning methods, it provides tree-based data representation, which can facilitate a visual understanding about underlying characteristics of classified observations. In test security investigations, this graphical representation could be further utilized for understanding the behavioral features of aberrantly behaved test takers. Moreover, it also runs efficiently on large datasets and provides the rank of importance of all the features. This piece of information could be helpful to investigate the key factors of classifying aberrant test takers from normally behaved population with high efficiency. For instance, by applying RF algorithm, we could examine momentary responding time to each question which may indicate suspicious problem solving behaviorbehavior that may reflect a certain degree of pre-knowledge of the items. Furthermore, the RF algorithm can handle higher order variable interactions reflecting more realistic complex relations among the variables embedded in the dataset. Though RF is one of the efficient supervised learning algorithms, some studies have shown that RF can overfit its dataset if the stopping rules are not properly set (e.g. Daz-Uriarte & De Andres, 2006; Segal, 2004).

Support vector machine. Support Vector Machine (SVM; Vapnik & Lerner, 1963 has gained popularity as a supervised kernel function based classification method used in diverse scientific fields (e.g., Furey, Cristianini, Duffy, Bednarski, & Haussler, 2000; Z. Huang, Chen, Hsu, Chen, & Wu, 2004; Meyer, Leisch, & Hornik, 2003). The SVM

algorithm attempts to create an optimal separating boundary, a line, plane, or hyper-plane by using a kernel function (linear or nonlinear) that divides the feature space (an n-dimensional space for predictors) whose margins are maximal. In this regard, this boundary is the best solution out of an infinite possible number of segregating boundaries. The optimal separating boundary, also known as the maximal margin hyperplane, is formed by maximizing the distance between all the training subjects and it. The maximal margin hyperplane is defined by computing the perpendicular distance from each subject to a given separating boundary. The smallest such distance is called the margin. As its name suggests, the maximal margin boundary is that separating hyperplane for which the margin is largest. The maximal margin here is also known as the hard margin, which means all the training subjects perfectly lie on either side of the hyperplane without any misclassification. Once the maximal margin hyperplane is constructed based on a training dataset, a new test subject could be classified later based on which side of a hyperplane it is located. The hard margin plane, however, is quite sensitive to a change in a subjects data, which may be due to over-fitting the training dataset. Thus, having a hyperplane that does not perfectly separate all the cases is worthy of attention. Sometimes, this kind of classification hyperplane is also referred to as the soft margin hyperplane, which is more robust to the change of an individual subject. The general support vector classifier can be represented as:

$$f(X) = b + \sum_{i \in S} \alpha_i K(x_i, x_{i'}), \tag{2.20}$$

where $K(x_i, x_{i'})$ is a kernel function that quantifies the similarity of two observations; S is the collection of indices of these support points, α_i and b are parameters needing

to be estimated. A simple binary classification example is introduced to help clarify the hard and soft approaches. Suppose a set of n training subjects on p variables exists $\boldsymbol{x}_1,...,\boldsymbol{x}_n \in \mathbb{R}^n$ marked with the labels $y_1,...,y_n \in \{-1,1\}$, is classified into two groups by a linear high-dimensional hyper-plane defined as

$$y_i(b+\alpha_i\sum_{i=1}^{n} x_{ij}x_{i'j'}) = 0. \tag{2.21}$$

In order to find the maximal margin hyper-plane, the equation above is optimized based on the constraint to maximize M, subject to

$$\sum_{i=1}^{n} \alpha_i^2 = 1, \tag{2.22}$$

and

$$f(x) = b+\sum_{i\in S} \alpha_i K(x_i,x_{i'}) \geq M; \quad \text{for all } i = 1,\ldots,n, \tag{2.23}$$

where M represents the margin of our hyper-plane. This is an example of the hard margin case requiring each subject in the training set be on the right side of the hyper-plane with at least an M margin. The soft margin case simply allows the optimization solution to be extended by again,maximizing M, subject to

$$\sum_{i=1}^{n} \alpha_i^2 = 1,$$

$$f(X) = b+\sum_{i\in S} \alpha_i K(x_i,x_{i'}) \geq M(1-\varepsilon_i),$$

$$\varepsilon_i \geq 0 \quad \text{and} \quad \sum_{i\in n} \varepsilon_i \geq C,$$

where C is a positive tuning parameter and determines the degree of tolerance of misclassified subjects, which violates the margin. If C = 0, the softer margin would transfer to

the hard margin case. The term, ε_i $(i = 1, \ldots, n)$, could allow some subjects to be on the incorrect side of the hyper-plane. For instance, if $\varepsilon_i = 0$, then the ith observation is on the incorrect side of the hyper-plane.

Among many advantages offered by SVM, one of the main benefits of using it is that SVM has the flexibility to select different kernel functions to adequately address practical problems in different modeling scenarios. By applying a proper kernel to the specific scenario, the performance of SVM could be dramatically improved. For example, polynomial or nonlinear kernel functions may be used when the cluster labels and features are nonlinearly related. Many kernels have been created for specific cases such as natural language processing (e.g., string kernels), speech recognition (e.g., time-alignment kernels), and image processing (e.g., histogram intersection kernel). SVM is a flexible platform for identifying aberrances by incorporating more types of data into current detecting framework. For example, writing strings and speech data would be jointly modeled with other psychometric variables like item response and responding time by applying appropriate kernel functions. However, to yield accurate results based on SVM, the tuning parameter and the types of kernel function should be set properly. In this way, it is relatively harder to be implemented compared with other supervised learning methods.

All the previously mentioned clustering algorithms are processes that use observed similarities or densities in data to identity patterns or group similar observations (Berkhin, 2006). These data mining methods will be utilized to identify clusters of respondents (e.g., such as those who exhibit fraudulent test-taking behaviors). The advantages and disadvantages of applying each algorithm as they relate to aberrant test-taking detection is presented in table 2.3.

Table 2.3: Advantages and disadvantages of various data mining methods for detecting aberrant test taking behaviors

	Unsupervised Methods			Supervised Methods		
	K-Means	Gaussian finite mixture	Self-organization mapping	K-nearest neighbor	Random forest	Support vector machine
PROS	* Easy to implement	* Manifest hidden clusters	* Easy to display complex relations of clusters	* Effectively robust	* Facilitate a visual exposition	* Flexible to apply different kernel functions
	* computationally efficient with a high dimensional dataset	* Useful to explore subcategories	* No distribution assumption	* Easy to implement	* Runs efficiently on large dataset	* Computationally efficient
		* Show cluster features (volume, shape and orientation)			* Handle higher order variable interactions	
CONS	* Relies on a predefined distance	* Sensitive to violations of distributional assumptions	* Relies on a predefined distance	* Sensitive to redundant and similar features	* Over-fitting	* Hard to implement
	* Require users to determine number of clusters	* Completely exploratory	* Require users to define the various parameters (e.g., map size, learning rate)	* High computational cost	* Hard to implement	* Computationally expensive
	Sensitive to the initial cluster centers					

2.3 Incorporating Biometrics to Detect Aberrant Testing Behaviors

Many methods have been developed to detect various aberrant test-taking behaviors. Certainly, these methods have shown success flagging aberrant test takers. However, these methods are potentially limited by their input data, which is either item response or response times. To improve detection accuracy beyond the traditional modeling framework, ways of incorporating real-time bio-metric information into traditional detection methods will be introduced.

2.3.1 Insights into Problem-Solving Using Eye Tracking

Eye tracking as an essential biometric technology will provide unique insights when assessing students cognitive processes during computerized problem-solving tasks. Eye tracking also can record temporal and spatial human eye movements, which are a natural information source for proactive systems that analyze user behavior. Moreover, eye tracking could also collect information about the location and duration of an eye fixation within a specific area on a computer monitor and can be critical supplementary data used in identifying cheating behaviors.

At its core, eye tracking is the measurement of eye activity. Capturing such information in the context of large-scale assessment testing scenarios may help address and answer some interesting questions related to aberrant testing behaviors detection such as: (a) Where does an examinee look and what does this information tell us about aberrant testing behavior? (b) When does blinking occur and what information does that convey about the examinees behavior? (c) How does the pupil react to different stimuli? (d) What

are the differences between the eye-gaze patterns of the normally behaved and aberrantly behaved test takers? In contrast, item responses and response times can not provide infomation about the eye-gaze patterns.

Holmqvist et al. (2011) classified most eye tracking indicators into four groups: eye movement, gaze position, numerosity, and latency. Movement indicators (e.g., saccadic direction, saccadic length) reflect the properties of eye gaze paths such as direction, amplitude, and velocity (e.g., Lee, Badler, & Badler, 2001; Motter & Belky, 1998; Ponsoda, Scott, & Findlay, 1995; Tatler & Vincent, 2008). Position indicators (e.g., fixations, dwells) address the question such as where a test taker looks. Numerosity indicators (e.g., fixations, dwells, blink rate, and regression rate) quantify eye movement related events in absolute numbers or proportional rates (Holmqvist et al., 2011). Latency indicators are mostly related to reaction time, which catches the time from the on- or offset of a stimuli/event to a specific reaction from our eyes (e.g. Born & Kerzel, 2008b; Shepherd, Findlay, & Hockey, 1986). For example, after showing an item on the computer screen, how soon does a test taker catch the first keyword? All these eye gaze indicators play important roles in understanding and classifying different test taking behaviors.

Some important eye-tracking indicators will be discussed in details in the following subsections

2.3.1.1 Fixation

In educational assessment and testing, eye fixation reflects the degree of the test-takers attention on specific words embedded in the items. In the context of test-taking

behavior, fixation is a measure of the temporary eye stoppage at a word of an item or a part of a graphical instruction when a test taker is solving a question. Several fixation-related measures are frequently studied in the literature including fixation counts, fixation rate, fixation duration, and fixation locations.

Many of these measures are used for assessing subjects′ information perception abilities, such as reading and problem solving. For instance, in reading ability assessment studies, Born and Kerzel (2008a) found that a reader' fixation duration was expected to be longer with more difficult and less frequent words compared with commonly used words. Similar findings reported in usability studies. For instance, Born and Kerzel (1999) found long fixation could indicate difficulty in extracting information for problem solving. Furthermore, longer fixation also indicates relatively high level of content engagement, which is also a reflection of high level of interest expressed by a student (Jacob & Levitt, 2003).

2.3.1.2 Pupil Diameter

Pupil diameter could be utilized to reflect the degree of fatigue, level of interest in a particular learning content, and the amount of workload of the test takers involved a specific cognitive task. Many studies reported negative correlation between levels of fatigue and pupil size (e.g. Lowenstein, 1962; Morad, Lemberg, & Dagan, 2000; Yoss, Moyer, & Hollenhorst, 1970). For instance, Morad et al. (2000) found that measured pupillary diameters differed significantly between fatigue (24 hours sleep deprivation) and clear-headed groups, reacting to controlled visual stimulus. This difference indicates changes of pupil diameter can be an objective measure of fatigue. Moreover, some studies have shown that

emotional arousal could be an important factor modulating pupils reaction. For instance, (Zubin & Steinhauer, 1983) found that pupil diameter was enlarged when pleasant and unpleasant pictures were presented to the experimental participants. Furthermore, other studies have demonstrated that pupil diameter can be a useful event-related measure of cognitive load (Hess & Polt, 1964; van Gerven, Paas, van Merrinboer, & Schmidt, 2002). This effect has been observed for tasks such as content comprehension (Just & Carpenter, 1993), visual searching (Porter, Troscianko, & Gilchrist, 2007), and mental number calculation (Hess, 1965).

2.3.1.3 Blinking

A seemingly involuntary function of the eye, blinking, is to keep the eyeball moist. In addition, blinking is highly related to other cognitive functions, like reflex blinking. Reflex blinks are the reactions to external stimulus for various purposes, such as protecting our eye balls, or maximizing our attention on a subject. Blinking rates, a commonly used measure of blinking, is defined as the number of blinks per given amount of time. Studies have shown that blink rate is positively associated with the number of simultaneous tasks (Barbato, della Monica, Costanzo, & De Padova, 2012; Colzato, Slagter, van den Wildenberg, & Hommel, 2009). In contrast, other studies found that people are more likely to reduce their blink rates when performing visually demanding tasks. For example, Benedetto et al. (2011) found that drivers′ blink rates decreases with higher visual demand, which indicates reallocation of potential cognitive recourses. Also, Fairclough and Venables (2006) reported similar findings. Blink rate negatively correlated with task

engagement. Researchers found that this negative relationship may due to the fact that people are more likely to reduce risk of missing key information during visually engaged tasks (Drew, 1951; Kennard & Glaser, 1964).

2.3.1.4 Saccades

Saccades, an eye gaze movement measure highly is related to fixation, and reflect the motion of the eyes from one fixation to another. The amplitude of the movement could vary from small jumps from one word to another to wide-reaching searches made while looking around a stadium. This measure could help to understand some general characteristics of eye gaze paths, such as direction, length and dispersion. For instance, many reading assessment studies (e.g., Kuperman & Van Dyke, 2011; Rayner & Liversedge, 2011) have shown that deficient readers have different eye gaze paths from efficient readers. Their gaze paths are relatively shorter and more scattered compared with efficient readers (VinuelaNavarro, Erichsen, Williams, & Woodhouse, 2017).

2.3.1.5 Regression

Regression refers to events that involve the motion of the eye in the opposite direction to the text. It often reflects the events related to re-reading and answer checking. Vitu (1991) classified regression into two types: long-range regression (LRR) and short-range regression (SRR). LRR means the eye gaze moves oppositely over several words or even sentences. SRR often refer to the short and rapid backward movements. Born and Kerzel (2008a) indicated that the occurrence of long-range regressions might due to

the fact that readers have missed, forgotten, or been unclear about what they have read. For example, some studies (e.g., Blanchard & IranNejad, 1987; Booth & Weger, 2013; Inhoff, Greenberg, Solomon, & Wang, 2009; Rayner, Murphy, Henderson, & Pollatsek, 1989) have shown that when items are less familiar, ambiguous or complex, the regression rate will increase in order to reinstate or reconfirm a cognitive effort. For SRR, some studies have indicated that SRRs are highly related to how much effort or care a reader devoted into a reading task. Coff and O'regan (1987), in their study, found that subjects were more likely to have more SRRs in order to increase the accuracy of registering the content information.

The relationship between these eye-tracking measures is shown in Figure 2.1.

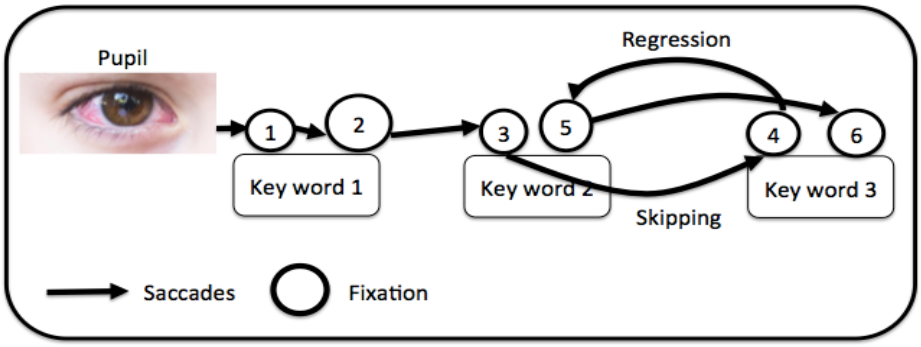

Figure 2.1: Relationship between several representative indicators

2.3.2 Representative Ways to Integrate Process Data into Psychometric Methods to Identify Aberrant Test Takers

Many testing programs have transferred from paper-pencil tests to computer-based or computer adaptive testing, which allows for simultaneously collecting multimodal data during the exam. The collected multimodel data includes three types. Product data are the

outcomes of the assessment tests such as item responses or test scores. Process data reflect the process of how a test taker form his/her final answer for an item/task such as response time and movements of mouse cursor usually recorded in the log-file. Biometric data are special cases of process data such as the eye-tracking indicator or heartbeats collected via sensors.

To understand the relationship between the processed data (e.g., RTs) and the product data (e.g., item responses), many methods have been proposed. These methods are trying to incorporate process data into the traditional psychometric models such as the Rasch model or 2PL IRT model, which are used for measuring latent abilities. In particular, RT has been used as ancillary information to better understand test takers′ performance than solely modeling the item responses. Since biometric data serve the same role as the RT, some representative methods of how to integrate RT into the traditional psychometric methods would be introduced. These methods can provide insights of how to incorporate biometric indicators later.

2.3.2.1 Incorporating RT as a Variable Into the Item Response Model

The simplest way to incorporate RT into traditional item response models is to add RT as an individual variable. Many methods have been proposed to achieve this goal (e.g., Luce, 1986; Roskam, 1997; Thissen, 1983; Verhelst, Verstralen, & Jansen, 2013)

Roskam′ model. Roskam, one of the pioneers, proposed a model to add log-transferred RT as a single term into the IRT model. His model could reflect the trade-offs between the amount of time a test taker spends and the difficulty level of a specific item.

The model is defined as

$$P_i(\theta_j) = \frac{1}{1+exp[-(\theta_j + lnT_{ij} - b_i)]}, \tag{2.24}$$

where θ_j is the person latent ability parameter, lnT_{ij} is the log-transferred response time term, and b_i represents the item specific difficulty parameter. The difference, $lnT_{ij} - b_i$, shows the trade-off between how much time a person working on a particular item and the difficulty level of that item. The interpretation of the difference $lnT_{ij} - b_i$ could be that a test taker would be more likely to spend more time on a harder question than an easy one, and vice versa.

Thissen' model. Another well-known attempt to incorporate RTs into the IRT model is a method proposed by Thissen (1983). His model treated the log-transferred RT as a dependent variable as opposite to an independent variable. The log-transferred RT is regressed on a parameter structure, which is similar to the parameterization of the two probability logistic (2PL) IRT model. The difference is that new terms are added to reflect the working speed for a person j and item i respectively. The model is defined as

$$lnT_{ij} = \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \tag{2.25}$$

where μ represents the grand mean level of the population of test takers and test item domain; β_i and τ_j are "slowness parameters" for item i and person j separately; and, ρ is a regression coefficient that indicates the degree of association between the log-transferred response times and the log odds of a correct response for the person j and the item i. The log odds of a correct response is calculated based on the 2PL model, which is parameterized as $a_i(\theta_j - b_i)$. a_i is the item discrimination parameter, b_i is the item difficulty parameter; θ_j represents the person-side latent ability parameter.

Later, Ferrando and Lorenzo-Seva (2007) extended Thissen's model in Equation 2.25 to a new version in order to accommodate the special needs of personality assessment. The updated model is given by

$$lnT_{ij} = \mu + \tau_j + \beta_i - \rho(\sqrt{a_i(\theta_j - b_i)} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.26}$$

. The difference between the two models in Equations 2.25 and 2.26 is the parameterization of the item parameter structure. Instead of using $a_i(\theta_j - b_i)$, Ferrando and Lorenzo-Seva (2007) use a slightly different item parameter structure, which is defined as $\sqrt{a_i(\theta_j - b_i)}$. However, these two models both reflect the trade-off between the working speed and responding accuracy. For example, for ρ larger than 0, the model given by Equation 2.26 shows that test takers with higher abilities would use less time than those who have relatively lower abilities, and vice vera.

2.3.2.2 Joint Modeling of Item Responses and Response Times

The previous section introduces several representative methods to directly add RT as ancillary information into the IRT models. RT is either treated as an independent variable or as a dependent variable. However, instead of treating RT as a covariate, RT itself could be modeled to either manifest the different test takers′working speed or show the corresponding characteristics of test items. The item features include responding time required by an item and the discrimination power of an item . Therefore, by jointly modeling RTs and item responses, the relationship between the working speed and the responding accuracy could be further discussed either based on the person-side or the item-side model parameters.

To jointly model item responses and RTs, it is essential to have a model to fit the item response time data properly. To achieve this purpose, many RT models have been proposed to improve the model fit by applying various distributions such as the exponential distribution, the log-normal distribution, the gamma distribution, and the Weibull distribution (e.g., Maris, 1993; Roskam, 1997; Schnipke & Scrams, 1997; Thissen, 1983; W. van der Linden, Scrams, & Schnipke, 1999). Among these proposed RT models, the log-normal response time model proposed by van der Linden (2006) draws much attention among researchers due to the easy interpretation of the model parameters, which follow the similar structure as the 2PL-IRT model. The log-normal RT model is defined as follows

$$f(t_{ij},\tau_j,\alpha_i,\beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\left(-\frac{1}{2}[\alpha_i\{\ln t_{ij}-(\beta_i-\tau_j)\}]^2\right), \tag{2.27}$$

where the latent parameter, $\tau_j \in \Re$, represents working speed for test-taker j. The item parameter $\beta_i \in \Re$ denotes time intensity, or simply, the amount of time required for answering a specific item. the parameter $\alpha_i \in \Re$ is an item time discrimination parameter. The mean value, $\ln t_{ij}$, is parameterized as $\mu_{ij} = \beta i - \tau_j$.

In the following section, methods of jointly modeling of the RTs and item response are introduced. The differences among these methods are based on the varying assumptions of how the item responses are related to the item response times.

Typically, two types of assumptions are made when jointly modeling the item responses and RTs. van der Linden (2007) assumes that the item responses and response times are independent after conditioning on their corresponding higher level random effects, which are the person latent parameters and item parameters. In other words, by

modeling the higher level random effects on the person and the item side separately, the item responses and response times are independent of each other. This is the most appealing approach to jointly model the item responses and the response times, which is defined as:

$$\begin{pmatrix} Y_{ij} \\ logT_{ij} \end{pmatrix} \sim N \left[\begin{bmatrix} \theta_j - b_i \\ \beta_i - \tau_j \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & \sigma_i^2 \end{bmatrix} \right], \tag{2.28}$$

where Y_{ij} and $logT_{ij}$ are the item responses and the log-transferred response times of person j on item i, respectively. θ_j and τ_j are the latent person parameters. b_i and β_i are the item difficulty and item time intensity parameters, respectively. σ_i^2 is the time residual variance of item i. The off-diagonals of the variance-covariance matrix are defined as 0s, which implies conditional independence of item responses and response times after modeling the patterns existing in the data covered by the mean structures. Figure 2.2 visualizes the modeling framework under the first assumption.

The second assumption assumes that the existence of the conditional dependencies (CDs) among the residuals of item responses and the log-transferred response times given the person and the item structural relationships. This may be due to two sources of variations when students take their tests: (1) between-person variability across items, and (2) within-person variability across items.

The between person variation reflects distinct test-takers' responding behaviors. Instead of assuming all the test takers are homogeneously responding on their tests. In other words, all the test takers are normally behaved test takers, we assume that some test takers are answering questions in aberrant ways such as cheating on their tests or responding

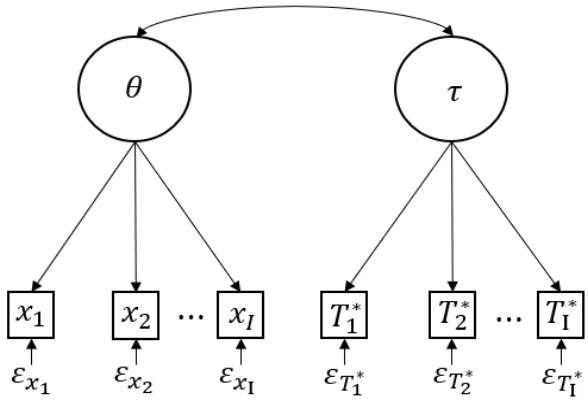

Figure 2.2: conditional independence of item responses and response times given latent ability and speediness: θ and τ are the latent ability and the speediness; T_i^* are the log-transferred response times; ε_{x_i} are the item response residuals

carelessly due to their low motivation. Nevertheless, it is still assumed that each test taker working on his/her test at a constant speed across all the items with a invariant cognitive capacity. For example, D. Molenaar, Bolsinova, Rozsa, and De Boeck (2016) proposed a mixture modeling approach to investigate the intraindividual variation in responses and response times.

$$\begin{pmatrix} Y_{ijk} \\ logT_{ijk} \end{pmatrix} \sim N \left[\begin{bmatrix} \theta_{jk} - b_{ik} \\ \beta_{ik} - \tau_{jk} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \sigma_{ik}^2 \end{bmatrix} \right], \tag{2.29}$$

where k indicates kth latent class. Y_{ijk} and $logT_{ijk}$ are the item responses and the log-transferred response times of person j on item i in latent class; and k, respectively; θ_{jk} and τ_{jk} are the latent person parameters in latent class k. b_{ik} and β_{ik} are the item difficulty and item time intensity parameters in kth latent class, separately. The σ_{ik}^2 is the time

residual variance of item i in latent class k. Figure 2.3 is a schematic of this modeling framework that depicts the relations between measured and latent variables..

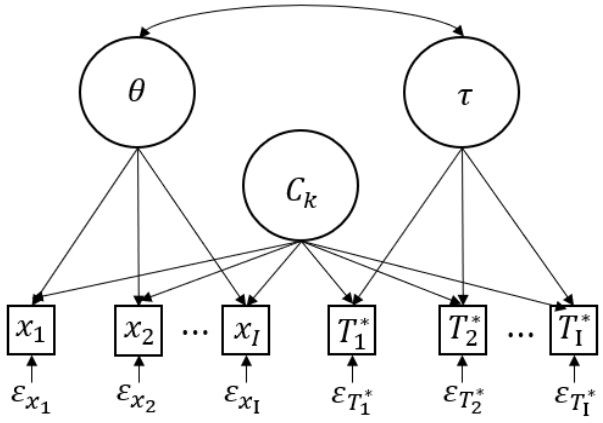

Figure 2.3: a mixture modeling approach to investigate the intraindividual variation in responses and response times: C_k indicates the kth latent class; θ and τ are the latent ability and the speediness; T_i^*, $i = 1, \ldots, I$ are the log-transferred response times; ε_{x_i}, $i = 1, \ldots, I$ are the item response residuals

In contrast, the within-person variability across items indicates that a test taker may change his working speed with variant cognitive capacity on answering different test items. Simply put, the within-person variability refers to the fact that a test takerperformance may vary from one item to another. The within-person variability could be caused by the following reasons, including (1) test fatigue (Ackerman & Kanfer, 2009; Ackerman, Kanfer, Shapiro, Newton, & Beier, 2010); (2) motivation changes (Wise & Kong, 2005); (3) guessing behaviors (Slakter, 1968); and, (4) application of various problem-solving strategies (van der Maas & Jansen, 2003).

One representative method to discuss the within-person variation would be the method proposed by Meng, Tao, and Chang (2015). They assumed that item residual and response time residual are correlated due to within-person variability. The model can

be represented as:

$$\begin{pmatrix} Y_{ij} \\ logT_{ij} \end{pmatrix} \sim N \left[\begin{bmatrix} \theta_j - b_i \\ \beta_i - \tau_j \end{bmatrix}, \begin{bmatrix} 1 & \\ \sigma_{1,2} & \sigma_i^2 \end{bmatrix} \right], \tag{2.30}$$

where $\sigma_{1,2}$ indicates the covariance between the item residual and the response time residual. Figure 2.4 represents a path diagram of this modeling framework that shows the relations between measured and latent variables.

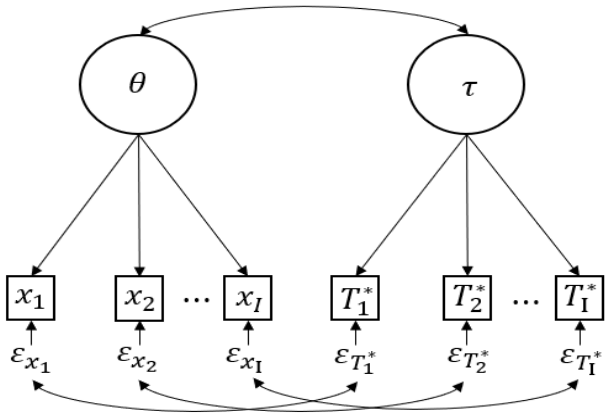

Figure 2.4: A conditional joint modeling approach for locally dependent item responses and response times

Thus far, several joint modeling approaches for item responses and response times were introduced, which provides insights about how to incorporate biometric variables into psychometric modeling frameworks. For instance, new models based on the gaze fixation counts collected via an eye tracker could be proposed to reflect the degree of test engagement when a test-taker solves a set of task questions. Also, the parametrization used for modeling RTs give a demonstration of how to model other biometric variables, which could show the individual differences and the item characteristics regarding the

interested variables. Moreover, the introduced jointly modeling approaches of item responses and response times could be further extended to add other biometric variables, which could provide comprehensive assessment about test takers′performance.

2.4 Future Directions and Challenges

Test security has been researched over the past several decades (Cizek, 1999). Cheating and other kinds of aberrant test-taking behaviors raise concerns about the validity of decisions made based on the estimated examinees scores. In order to maintain the fairness among all test-takers, it is important not only to flag improper behaviors but to take actions against them. In order to better flag the behaviors with high accuracy, statistical models would be based not only on one source of information, such as item response. In the future, all sources of information, such as bio-information technologies about test-takers, would be aggregated using highly efficient computational methods such as cloud computing. Mislevy et al. (2016), in the Maryland Assessment Research Center Conference (MARC), indicated new forms of assessments would involve psychometric models, bio-information, machine learning, and data mining methods with a high efficiency computation platform. Many methods based on ”big-data” are currently under development. Man, Harring, and Sinharay (2019) and Thomas (2016) has applied support vector machine, a data-mining method, to pre-knowledge cheating detection. He and von Davier (2015) proposed a statistical feature selection learning method by finding the features from the process data for classifying different learning patterns. More kinds of testing forms will appear, such as online testing and games-based testing. More novel

item types, such as multi-part directional dependent items, will replace currently used measures. Growth in learning will not be simply measured at several time points; it will be monitored constantly over time. How to incorporate psychometric methods to maintain the validity of the inferences from an assessment is one of the biggest challenges and directions for the future. Mueller, Zhang, and Ferrara (2016) summarized four challenges that must be overcome in the next generation of test-security research. The first is dealing with low signal-to- noise ratio (SNR). Aberrant test-taking behaviors do not always indicate those behaviors that are harmful to the inferential claims (i.e., validity) of the assessment, such as tiredness or creative responding. Finding ways to better classify with high sensitivity is an important research direction to be studied. The second challenge is in understanding effect sizes. How far can the data differ from the expectation and truly be considered cheating? How could we transfer abstract metrics into more understandable ones? The third challenge is one that still haunts psychometrics and that is how to explain and convey results to laypersons. Stakeholders have different interests and levels of understanding of how aberrant test-taking behaviors affect their decisions. How to present methods and results in a logical and comprehensible manner must be considered and taken seriously. To this end, more data visualization methods need to be developed to aide in this endeavor. The fourth challenge is how to incorporate justice into the decision-making process. This would necessarily improve by advancing methods that increase the accuracy of classification between aberrant and normal test-takers. As important in this endeavor is controlling the rates of false negatives. By having more accurate measures and more informative evidence, results from psychometric models could be incorporated into this justice system.

2.5 Conclusion

The present literature review has provided a focused overview of some aberrant test-taking behavior detection methods, including unexpected score gain analysis, erasure analysis, similarity analysis, person fit analysis, data mining methods and eye-tracking measures. There remain many different shortages and limitations associated with the proposed methods, primarily because real data analytic situations are often more complicated and chaotic than models can predict. In 1976, George Box said, all models are wrong but many may nonetheless be useful, especially when parsimonious. (p. 202). All the models or indices reviewed have their own advantages to help us understand the underlying issues. Although none of the methods are perfect, they provide the foundation for improvements in each aspect of test security related issues to be made. Future research in the field of large-scale testing will be both challenging and full of promise.

Chapter 3: Methodology

In this chapter, the experimental design and the new methods for incorporating bio-information and their usage for detecting aberrant behaviors will be introduced. First, the experimental design for data collection will be illustrated. Next, three negative binomial distribution-based visual fixation counts models will be presented. This model will be used for assessing the visual attention differences among test-takers. Furthermore, a jointly modeling approach of integrating product data, process data and biometric information will be shown. By joint modeling the three types of information, we can assess test-takers' performance in a comprehensive way. Lastly, various data mining methods will be used for classifying different types of test takers.

3.1 Experimental Design

3.1.1 Data Collection

In the proposed study, 298 students, who were over 18 years old were invited to participate in the eye-tracking lab to take one of the required exams that would mimic the taking of a high-stakes assessment. All the participants were enrolled through the UMD Psychology SONA system to avoid any selection bias based on race, gender and major,

etc. The SONA system is a specific online platform for enrolling participants for various psychological studies. Also, only those participants were recruited who did not suffer from blindness having either normal or corrected vision. All the enrolled participants took an exam, which contained 20 multiple choice items from a high-stake test offered by the ETS. Ten items were verbal reasoning questions, and ten items were quantitative methods questions.

3.1.2 Experimental Conditions

There were three conditions in the proposed research design: (1) participants in the control condition would not receive any test preparation materials, (2) participants received questions that were similar to their exam, and (3) participants in the third condition would receive similar exam questions and the answer key. Participants were randomly assigned into different experimental conditions in order to minimize targeted internal and external experimental threats. The data from all three conditions were then combined to allow the researcher to conduct blind statistical classification of examinees. Background information including motivation, test anxiety, Big Five personality traits, morality, and religiosity, for example, were also be collected. Most of the variables would be later used as the input data for classifying different test taking behaviors. This study was fully approved by the University of Maryland institutional review board (IRB).

3.1.3 Data Recording

All the test items were clearly presented as slides, which were converted to a .pdf file. Each slide only contained one item. Line space was at least doubled to accommodate the eye-tracker's accuracy level (0.5-1 degree of visual angle accuracy).

Test takers' eye movements were recorded at 60 Hz with the Gazepoint eye-tracking system. It was placed on a firm large table under a monitor (1024 by 768 resolution; 17-inch LCD). The eye-tracker has 0.5-1 degree of visual angle accuracy. The recording area was about 20-25 squares meters without windows to minimize direct and ambient sunlight. The recording room was inside a suite with limited surrounding noise.

3.2 New Test Engagement Model Based on Visual Fixation Counts

To accurately classify different type of test takers, it is important to select an eye-tracking variable indicating the degree of visual efforts a test taker puts on an item. Among all the collected eye-tracking variables, eye fixation counts could be used to understand the visual engagement when a test taker perform his/her test (e.g., Jacob & Levitt, 2003; Poole, Ball, & Phillips, 2004). For example, in the human-computer interaction and usability study, Poole et al. (2004) indicated that increased gaze fixation counts on an interested visual area show that it is more essential, more noticeable to the subject than other visual areas. Similar results were reported by Justice and Lankford (2002) and by Roy-Charland, Saint-Aubin, Klein, and Lawrence (2006). With this systematic relation between test engagement and visual fixation counts, a model measuring the cognitive connection between test takers latent visual engagement and the observed visual fixation

counts appears warranted. This model could potentially help to understand individual visual effort differences, which may prove to be an important feature in detecting aberrant test takers from the normal behaved ones.

To model the relation between gaze fixation counts and test engagement, a negative binomial fixation (NBF) model was proposed and fitted to real data gathered as part of an experiment. A Bayesian estimation approach via Markov chain Monte Carlo (MCMC) was used to estimate model parameters.

In this study, the negative binomial distribution was chosen as a link function for modeling the visual fixation counts. Unlike the Poisson distribution for count data which forces the variance to follow the mean, the negative binomial distribution allows for the data to be overdispersed—variance is unequal to the mean. In addition, the negative binomial distribution was sufficiently flexible so that, for example, the mean structure could be parameterized in useful ways that incorporate latent person as well as item parameters. Several structures for the latent person parameters was introduced that are parsimonious, and unlike other studies (Fox & Marianti, 2016) that assume constant engagement levels across items, accommodates systematic change across items reminiscent of the implied mean structure in latent growth models.

3.2.1 The Negative Binomial Fixation Model

The NBF model was designed to reflect item quality and a test taker's engagement level with a number of items, i $(i = 1, \ldots, I)$, on an assessment. The proposed NBF model follows a negative binomial distribution, which can be parameterized in various ways. A

flexible, yet conventional, parameterization is to define the negative binomial distribution as the number of failures (X) before the rth success in which the probability mass function (pmf) is defined as:

$$P(X = x|s, p) = \frac{\Gamma(x+s)}{x!\,\Gamma(s)} p^s (1-p)^x, \tag{3.1}$$

where p is the probability of success in each Bernoulli trial ($p \in [0,1]$) and $s > 0$ denotes the shape parameter. The expectation of the random variable X: $E(X) = s(1-p)/p$, and the variance of X: $Var(X) = s(1-p)/p^2$.

Instead of parameterizing the negative binomial distribution with regard to s and p, a convenient parameterization utilizes the relation between p and the expectation of the negative binomial distribution, μ. Through algebraic manipulation, parameter p can be expressed as a combination of μ and s as

$$p = \frac{s}{s+\mu}. \tag{3.2}$$

In Equation 3.2, the mean μ of the negative binomial distribution can be further decomposed into a structure that separates the latent person effect (ω_j) and item parameter effect (m_i) as

$$\mu_{ij} = exp(m_i + \omega_j), \tag{3.3}$$

where $\mu_{ij} = exp(m_i - \omega_j) > 0$; $\omega \in R$; and $m \in R$.

In summary, the NBF model is expressed as:

$$P(x_{ij}|s_i, m_i, \omega_j) = \frac{\Gamma(x_{ij}+s_i)}{x_{ij}!\,\Gamma(s_i)} \left(\frac{s_i}{exp(m_i+\omega_j)+s_i}\right)^{s_i} \left(\frac{exp(m_i+\omega_j)}{s_i+exp(m_i+\omega_j)}\right)^{x_{ij}}. \tag{3.4}$$

Parameter m_i is associated with the test and can be interpreted as the visual intensity for item i. The presumption is that this parameter represents the amount of cognitive

engagement a student tends to exert on a test. Person-specific parameter, ω_j for each of the J test takers ($j = 1, \ldots, J$), denotes the overall test engagement level test taker j, and is assumed, at least initially, to be constant across all the items. Furthermore, a discrimination parameter, α_i, for item i is defined as the inverse of σ_i, and $\sigma_i = \sqrt{\mu_{i\cdot} + \mu_{i\cdot}^2/s_i}$. Thus, α_i reflects the overall dispersion of the fixation counts on item i. Larger values of α_i lead to steeper slopes of the pmf of the negative binomial distribution while smaller values of α_i correspond to shallower slopes. Thus, given any value of the engagement intensity parameter m_i, the difference (i.e., $\Delta\omega$) between any two values of engagement level, say ω_1 and ω_2, from two test-takers would be larger indicating that the item is more discriminating than an item having a smaller α_i value.

To properly apply this model, several assumptions need to be satisfied. First, it is assumed that fixations counts for each item solely reflect the different levels of visual engagement as students perform their tests. For instance, by taking an average across all the test takers, an item with more fixation counts would indicate a higher level of visual effort required for solving that item than an item with fewer fixation counts. Also, fixation counts are assumed independent of each other by conditioning on the latent test engagement parameter (Ω). Furthermore, test engagement level (ω_j) is assumed to be constant across items. However, this assumption can be relaxed by parameterizing the mean structure incoperating the linear or quadratic changes across items.

Figure 3.1 represents the fixation count model (constant test engagement across all the tasks). Tests are solved in a sequential order from item 1 to item I. As is customary with path diagrams from structural equation modeling, circles indicate latent variables, in this case Ω, denoting latent engagement. The squares are measured indicators of the latent

variable. In the case of Figure 3.1, these represent fixation counts collected at each item point. Small arrows showing measurement errors are attached to the observed indicators. To give some idea of the distribution of fixation counts, histograms for two items (i.e., item 1 and item 3) across test takers used in the upcoming example, with superimposed normal densities are displayed in Figure 3.2.

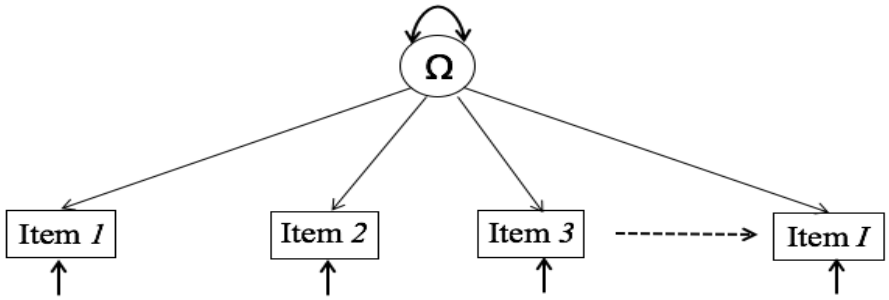

Figure 3.1: A graphical representation of the negative binomial fixation model: The circle indicates the latent variable in the model, while Ω stands for latent test engagement. The squares represent the fixation counts collected at each item point, which are the indicators to measure latent engagement. Small arrows showing measurement errors are attached to the observed indicators.

3.2.2 Negative Binomial Fixation Model with Linear Trend

The NBF model shown in the previous section assumed constant engagement level for individual test taker across all items. However, this assumption may be unrealistic because it is likely that test takers may change their responding behaviors at different stages of their tests Wise and Kong (2005). For instance, test takers may feel fatigue and be careless towards end of their tests, or, they may start guessing more at end of a test due to the time pressure to finish all the questions. Thus, a more flexible model is required that would accommodate with the changes of test engagement.

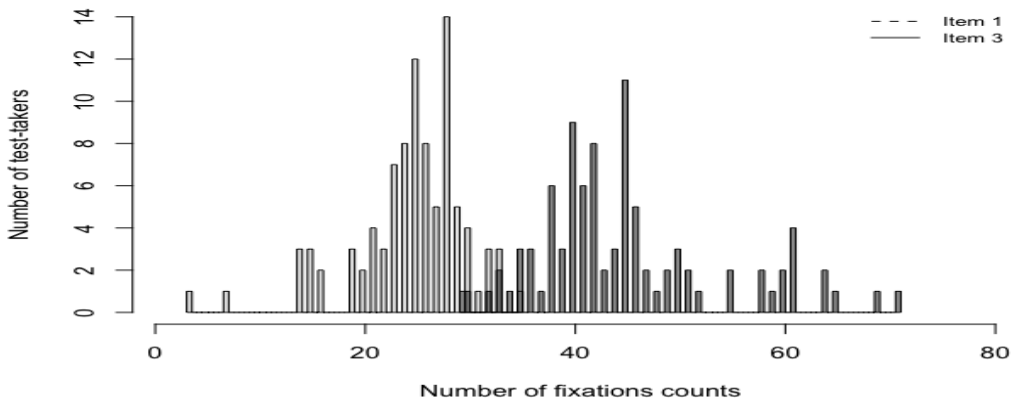

Figure 3.2: Two items with fitted fixation counts

An NBF model with flexible linear trend (NBF-LT) is proposed. The mean structure displayed in Equation 3.5 is reparameterized by adding a test-specific trend indicator, which takes on the same form as a latent growth model Bollen and Curran (2006),

$$\mu_{ij} = \exp(m_i - \mathbf{X}\Omega_j). \tag{3.5}$$

In this parameterization, intercept and slope parameters are elements of vector, Ω_j, where $\Omega_j = (\omega_{0j}, \omega_{1j})^T$ is assumed to follow a bivariate normal distribution. That is

$$\Omega_j = \begin{pmatrix} \omega_{0j} \\ \omega_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2_{\omega_0} & \sigma_{\omega_0\omega_1} \\ \sigma_{\omega_1\omega_0} & \sigma^2_{\omega_1} \end{pmatrix} \right]. \tag{3.6}$$

For test taker j, parameter ω_{0j} represents the initial engagement level at item 1 and parameter ω_{1j} is the slope parameter, which permits constant change in engagement across the items for different test takers. The means of ω_{0j} and ω_{1j} are fixed to 0. Letting the growth parameter means be zero can facilitate the interpretations of individual-specific intercepts and slopes. For example, $E[\omega_{0j}] = 0$ is the population expectation of initial

engagement level and constrains 0 to be a reference starting level. Thus, the sign of ω_{0j} indicates whether a person has greater or less initial engagement compared to the reference level. By fixing the expectation of ω_{1j} to 0, the sign of ω_{1j} indicates whether engagement is increasing or decreasing across items.

Elements of the $I \times 2$ design matrix $\mathbf{X}$ are formulated to correspond to linear growth and are defined as

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ I_1 & I_2 & I_3 & \cdots & I_I \end{bmatrix}^T, \tag{3.7}$$

where all the test takers' fixation counts are recorded for the same items, I_i. Usually, $I_1, \ldots, I_I$ take values corresponding to the sequence of answering the questions such as $1, 2, \ldots, I$. Figure 3.3 shows a graphic of the NBF model with flexible linear trend.

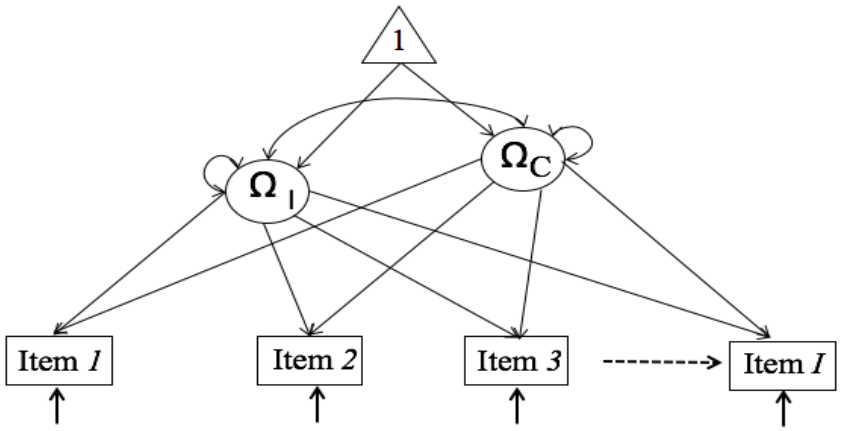

Figure 3.3: A graphical representation of the negative binomial fixation model: The circle indicates the latent variable in the model, while Ω stands for latent test engagement. The squares represent the fixation counts collected at each item point, which are the indicators to measure latent engagement. Small arrows showing measurement errors are attached to the observed indicators.

3.2.3 Negative Binomial Fixation Model with Quadratic Trend

To accommodate with curvilinear trends in engagement, the NBF-QT model was proposed. This model extension can help capture nonlinearities in engagement while test-takers perform their tests. The sign of the coefficient of the quadratic term indicates the concavity (i.e., curve's orientation) of the curve. A positive quadratic coefficient would result in trends that are convex (i.e., open up). In contrast, a negative coefficient would result in trends that are concave (i.e., opens downward). Accommodating this elaboration can be done in a straightforward way by extending the mean structure of the NBF-LT model in Equation 3.8. In the NBF-QT model, parameter vector, Ω_j, now has three elements with the inclusion of a quadratic parameter, $\Omega_j = (\omega_{0j}, \omega_{1j}, \omega_{2j})^T$. Clearly, Ω_j now follows a multivariate normal distribution as

$$\Omega_j = \begin{pmatrix} \omega_{0j} \\ \omega_{1j} \\ \omega_{2j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\omega_0} & & \\ \sigma_{\omega_1\omega_0} & \sigma^2_{\omega_1} & \\ \sigma_{\omega_2\omega_0} & \sigma_{\omega_2\omega_1} & \sigma^2_{\omega_2} \end{pmatrix} \right], \tag{3.8}$$

with $I \times 3$ design matrix $\mathbf{X}$ is now defined to accommodate the quadratic growth parameter as

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ I_1 & I_2 & I_3 & \dots & I_I \\ I_1^2 & I_2^2 & I_3^2 & \dots & I_I^2 \end{bmatrix}^T \tag{3.9}$$

3.3 A Three-way Joint Modeling Approach of Item Response, Response Time and Fixation Counts

Various ways of integrating process data have been introduced in Chapter 2. Inspired by the works cited in Chapter 2, especially the work of van der linden (2007), in this proposal, a trivariate joint modeling approach for item responses, RTs and fixation counts is proposed. This trivariate joint modeling approach delineates the trade-offs among the responding accuracy, working speediness and visual test engagement. The proposed joint modeling is an extension of the hierarchical modeling framework proposed by W. J. van der Linden (2006a). In this joint modeling approach, the one-parameter logistic (1-PL) model, the log-normal RT model and the NBF model are specified separately at level one. The variance-covariance structure of the person and item parameters are jointly estimated at level two. A Bayesian estimation approach is used to investigate the proposed hierarchical model.

3.3.1 Measurement Models at Level 1

For item responses, a one-parameter logistic (1-PL) model (Lord, 1952) is used. 1-PL model describes the relation between an item response of an examinee and one general latent ability trait, which is formulated as

$$P(u_{ij}=1|\theta_j;b_i)=\frac{1}{1+e^{-D(\theta_j-b_i)}} \tag{3.10}$$

where $P(u_{ij}=1|\theta_j;b_i)$ is the probability of a correct response to item i, $i=1,\ldots,I$ by person j, $j=1,\ldots,J$; b_i is the location parameter (the difficulty parameter) for item i, and θ_j is a general latent trait for person j. D is a scaling constant, which is fixed as 1.7.

In addition to the 1-PL model, the log-normal RT model (W. J. van der Linden, 2006b) formulated in Equation 2.27 will be utilized to reflect the relationship between the response times and latent working speed. Moreover, the NBF model formulated in Equation 3.5 will be used to cover the association between the visual fixation counts and the latent test visual engagement.

3.3.2 Modeling Item Domain and Person Domain models at Level 2

The second-level models incorporates two correlational structures to account for the dependencies on both the item and person parameters, respectively.

3.3.2.1 Modeling Person Domain Parameters

In this joint modeling approach, the person domain covers three latent person-side variables, which are latent ability θ, working speed τ, and visual engagement ω. The relation among these three person-side latent variables for the population of test takers is assumed to follow a multivariate normal distribution such that

$$\Theta_p = (\theta, \tau, \omega)^T \sim MVN(\mu_p, \Sigma_p), \tag{3.11}$$

with mean vector, $\mu_p = (\mu_\theta, \mu_\tau, \mu_\omega)'$, and covariance matrix

$$\Sigma_p = \begin{pmatrix} \sigma_\theta^2 & & \\ \sigma_{\theta\tau} & \sigma_\tau^2 & \\ \sigma_{\theta\omega} & \sigma_{\tau\omega} & \sigma_\omega^2 \end{pmatrix}. \tag{3.12}$$

The parameters, $\sigma_{\theta,\tau}$ represent the linear dependencies between the latent ability and the speediness of the test-taker. $\sigma_{\theta,\omega}$ represent the relation between the ability and the

test visual engagement of the test-taker. $\sigma_{\tau,\omega}$ represent the association between speediness and the test visual engagement of the test-taker. The sign of the parameters estimates indicate the trade-offs among all these latent variables. For instance, A negative value of $\sigma_{\theta,\tau}$ indicates that test-takers who solve a task more quickly also have lower latent ability (Bolsinova, De Boeck, & Tijmstra, 2017; De Boeck, Chen, & Davison, 2017; W. J. van der Linden, 2006a).

3.3.2.2 Modeling Item Domain Parameters

To account for the item parameter dependencies in this joint modeling approach, a multivariate normal distribution is defined for the item parameters, $\Xi_I = (b_i, \beta_i, m_i)'$, such that

$$\Xi_I \sim MVN(\mu_I, \Sigma_I), \tag{3.13}$$

where the mean vector and symmetric covariance matrix, μ_I and Σ_I, are defined respectively as $\mu_I = (\mu_b, \mu_\beta, \mu_m)'$ and

$$\Sigma_I = \begin{pmatrix} \sigma_b^2 & & \\ \sigma_{b\beta} & \sigma_\beta^2 & \\ \sigma_{b\omega} & \sigma_{\beta m} & \sigma_m^2 \end{pmatrix}. \tag{3.14}$$

These moments are a restrictive version of parameter vector $\Xi_{GI} = (b_i, \beta_i, m_i, \alpha_i, \zeta_i)'$, in which $\mu_{GI} = (\mu_{bi}, \mu_{\beta_i}, \mu_{m_i}, \mu_{\alpha_i}, \mu_{\zeta_i})'$ and

$$\Sigma_{GI}\begin{pmatrix} \sigma_b^2 & & & & \\ \sigma_{b\beta} & \sigma_\beta^2 & & & \\ \sigma_{bm} & \sigma_{\beta m} & \sigma_m^2 & & \\ \sigma_{b\alpha} & \sigma_{\beta\alpha} & \sigma_{m\alpha} & \sigma_\alpha^2 & \\ \sigma_{b\zeta} & \sigma_{\beta\zeta} & \sigma_{m\zeta} & \sigma_{\alpha\zeta} & \sigma_\zeta^2 \end{pmatrix}$$

respectively. Restrictions are put on these item parameters such that the only parameters to be estimated will be item location, time intensity, and item visual engagement intensity. For instance, studies by Bolt and Lall (2003), Fox, Entink, and Avetisyan (2014), and Wang and Nydick (2015) show that estimating the correlation among the item slopes, item time discrimination, and item visual engagement discrimination could potentially lead to model over-fitting. The estimation precision of person-side parameters would be reduced due to the lower degrees of freedom induced by needlessly estimating these correlations. Figure 3.3 displays the graphical representation of the trivariate jointing modeling of item response, response time, and visual fixation counts.

The proposed trivariate joint model can help to integrate eye-tracking indicator (visual fixation) into traditional psychometric modeling framework. With this joint modeling framework, we could have a comprehensive picture of test takers' cognitive processes essential to understanding the underlying problem-solving process that is impossible to assess from item responses alone. In addition, the current joint modeling approach can be extended to incorporate other essential biometric indicators. Thus, the current joint modeling approach serves as a elementary foundation for bridging biometric and psychometric information.

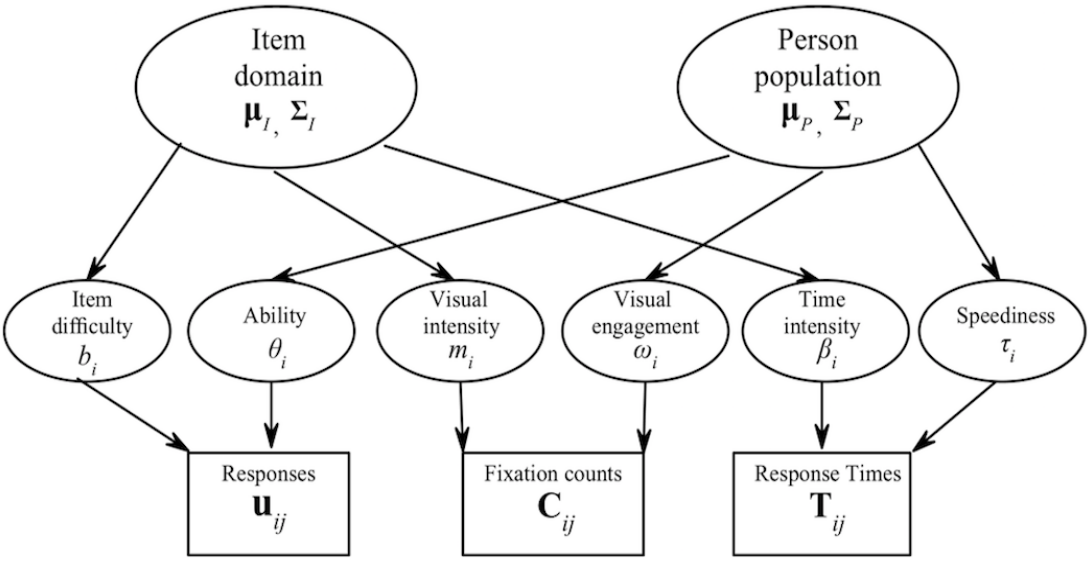

Figure 3.4: Trivariate joint model approach of item response, response time, and visual fixation counts

3.3.3 Model Parameter Estimation

In this study, Bayesian estimation of model parameters is implemented in Just Another Gibbs Sampler (JAGS; Plummer, 2015), which is housed in the `R2jags` package (Su & Yajima, 2015). Convergence is assessed via the `coda` package. Two chains using 96,000 total iterations with thinning of 2 to alleviate auto-correlation among draws, were executed. Model parameter estimates and standard deviations were summarized based on the posterior densities using the final 4,000 iterations after burning-in 92,000. The potential scale reduction factor (PSRF) was used for evaluating convergence for all model parameters (Gelman, Carlin, Stern, & Rubin, 2003). For the current study, a PSRF value of 1.1 or less for each model parameter was used as the arbiter indicating convergence.

Constraints for Modeling Identification

To properly identify the scales of the latent variables, model constraints are needed either on the item side (fixing the summation of item thresholds to zero) or the person-side (fixing the expectation of the latent ability parameter to zero). In this study, the model identification scales were fixed on the person-side by following the convention used for IRT model estimation (Volodin & Adams, 1995; Wu, Adams, Wilson, & Haldane, 1998).

For the 1-PL model, the population mean of the latent ability, θ, was set to 0 (Lord, 1952), and, the item discrimination parameter for each item was fixed to unity. For the log-normal RT model, the population mean of latent speediness, τ, was constrained to 0 as well (W. J. van der Linden, 2006c). For the NBF model, the mean of the latent person-side visual engagement parameter Ω is also set to zero (Man & Harring, 2019).

$$\mu_\omega = \mu_\theta = \mu_\tau = 0. \tag{3.15}$$

Prior Distributions

The prior distribution of item parameters, Ξ_I referring to Equation 3.13, for the proposed model is assumed to be trivariate normal. A Gamma distribution is assumed for the time discrimination parameter [i.e., $\nu_i \sim Gamma(1,1)$]. This is the inverse of the variances of the log-times on different items (σ_ε^2) based on the RT model: $\log(T_{ij}) \sim N(\beta_i - \tau_j, \sigma_\varepsilon^2)$. In addition, the fixation dispersion parameter for each item [i.e., $s_i \sim IG(1,1), i = 1,...,I$] is assumed to follow an inverse Gamma distribution as well. Hyper-priors are defined as

$$\mu_d \sim N(0,2), \quad \mu_\beta \sim N(4.0,2), \quad \mu_m \sim N(0,1) \quad \Sigma_I \sim IW\left(\mathbf{I}_I, \nu\right),$$

where $\mathbf{I}_I$ is an 3 by 3 identity matrix, and ν is the degree of freedom, which in this case is equal to 3.

Similarly, the prior specification for the person parameters, Θ_p refering to Equation 3.11, of the three-way joint model follows a trivariate normal distribution. And, the μ_I fixed as 0s. And,

$$\Sigma_p = \begin{pmatrix} \sigma_\theta^2 & & \\ \sigma_{\theta\tau} & \sigma_\tau^2 & \\ \sigma_{\theta\omega} & \sigma_{\tau\omega} & \sigma_\omega^2 \end{pmatrix} \sim IW\left(\mathbf{I}_P, \nu\right).$$

The joint posterior probability for the proposed model can be represented as

$$p(\Theta_p, \Xi_I|\mathbf{u}, \log(\mathbf{T}), \mathbf{c}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J} p(\mathbf{u}_{ij}, \log(\mathbf{T}_{ij}), \mathbf{c}_{ij}|\Theta_j, \Xi_i) p(\Theta_j|\mu_p, \Sigma_p) p(\Xi_i|\mu_I, \Sigma_I)$$

$$p(\mu_d) p(\mu_\beta) p(\mu_m) p(\Sigma_I|\nu) p(\mu_p|\mathbf{0}, \Sigma_p) p(\Sigma_p|\nu),$$

3.3.4 Evaluating Model-data Fit: Posterior Predictive Model Checking

In this study, posterior predictive model checking (PPMC) was used for evaluating whether the proposed model adequately accounted for the variability existing in the data. Specifically, PPMC was used to check our model-data fit (see, e.g., Gelman, Meng, & Stern, 1996; Levy, 2009; Rubin, 1996; Sinharay, Johnson, & Stern, 2006).

Introduction of the Method

Let ψ=$(\Theta_p^T, \Xi_I^T)^T$ be the vector of parameters, we are interested in estimating, and let $\mathbf{y}$ be the set of observed data (e,g., item responses, response times, and visual fixation counts). Thus, the likelihood based on the conditional distribution of the data given model parameters could be expressed as $p(y|\psi)$, and the prior distributions of all the model parameters could be denoted as $p(\psi)$. By applying Bayes' rule, the posterior distribution for a given set of parameters could be expressed as

$$p(\psi|\mathbf{y}) \equiv \frac{p(\mathbf{y}|\psi)p(\psi)}{\int_\psi p(\mathbf{y}|\psi)p(\psi)d\psi}. \tag{3.16}$$

To check the model-data fit by PPMC, predicted data are generated from the joint posterior distribution. The generated replicated dataset is denoted as y_r^{pred} for r = 1, 2,...,R; where R indicates the number of draws from the joint posterior distribution. The distribution of predicted data, named as the posterior predictive distribution of predicted data (see, Equation 3.16), could be utilized for checking the data model fit.

$$p(\mathbf{y}^{pred}|\mathbf{y}) = \int p(\mathbf{y}^{pred}|\boldsymbol{\psi})p(\boldsymbol{\psi}|y)d\boldsymbol{\psi}. \tag{3.17}$$

Model fit is evaluated by comparing the differences between the predicted data y_r^{pred} for r = 1, 2,...,R, and the observed data, y. A small difference would be indicative of satisfactory data-model fit. Instead of directly comparing the predicted data and the observed data, a discrepancy measure, $T(\cdot)$, a function of data and model parameters, is usually computed, which summarizes the data and the corresponding model parameters (Gelman et al., 1996).

The model-data fit can be evaluated by comparing the difference between the $T(y^{pred}, \boldsymbol{\psi})$ and $T(y, \boldsymbol{\psi})$, which are calculated based on predicted and realized data, respectively. In practice, a *posterior predictive* p-value (PPP-value) is defined as the probability of obtaining the predicted data that is more extreme than the observed data. The estimated PPP-value is the proportion of $T(y^{pred}, \boldsymbol{\psi})$ equal to or larger than $T(y, \boldsymbol{\psi})$ over the R draws. A PPP-value close to 0 or 1 is indicative of poor model-data fit since the predicted data y_r^{pred} is more extreme than the observed data, y. The posterior predictive p-value (PPP-value) is defined as

$$p = p(T(y^{pred}, \boldsymbol{\psi}) \geq T(y, \boldsymbol{\psi})) = \int\int I_{T(y^{pred}, \boldsymbol{\psi}) \geq T(y, \boldsymbol{\psi})} p(y^{pred}|\boldsymbol{\psi})p(\boldsymbol{\psi}|y)dy^{pred}d\boldsymbol{\psi}, \tag{3.18}$$

where $I_{\cdot}$ is the indicator function. To compute the data-model fit for the proposed model by applying the PPMC method, Sinharay et al. (2006) suggested the following three-step procedure outlined in Patz and Junker (1999):

1. Draw the item parameter and person parameter estimates for the proposed model from the posterior distribution (see, Equation 3.16).

2. Draw y^{pred} from the proposed model given by Equation 3.17 based on the drawn item parameter and person parameter estimates in step 1.

3. Compute the values of observed and predictive discrepancy measures (e.g., item-fit statistics or descriptive statistics only based on data) from the above draws of parameters and data set.

The data-model fit can be evaluated based on the computed PPP-values, which are given by the Equation 3.18. Figure 3.4, a modification of a schematic presented by Sinharay et al. (2006), graphically demonstrates the detailed procedure of using the PPMC method to evaluate the data-model fit.

Discrepancy Measures for the Proposed Models

Three statistics will be introduced in this section. Those three statistics will be used as different discrepancy measures, $T(\cdot)$, to evaluate the item by person-level date-model fit for item responses, response times, and visual fixation counts, separately. Specifically, the values of $T(y^{pred}, \psi)$ and $T(y, \psi)$ will be calculated based on the predicted dataset and observed dataset based on three statistics. Then, PPP-values will be calculated preferably based on the discrepancies between $T(y^{pred}, \psi)$ and $T(y, \psi)$ as Figure 2 demonstrated. The three item-fit statistics are: (1) the W index (Wright & Stone, 1979); (2) the L index

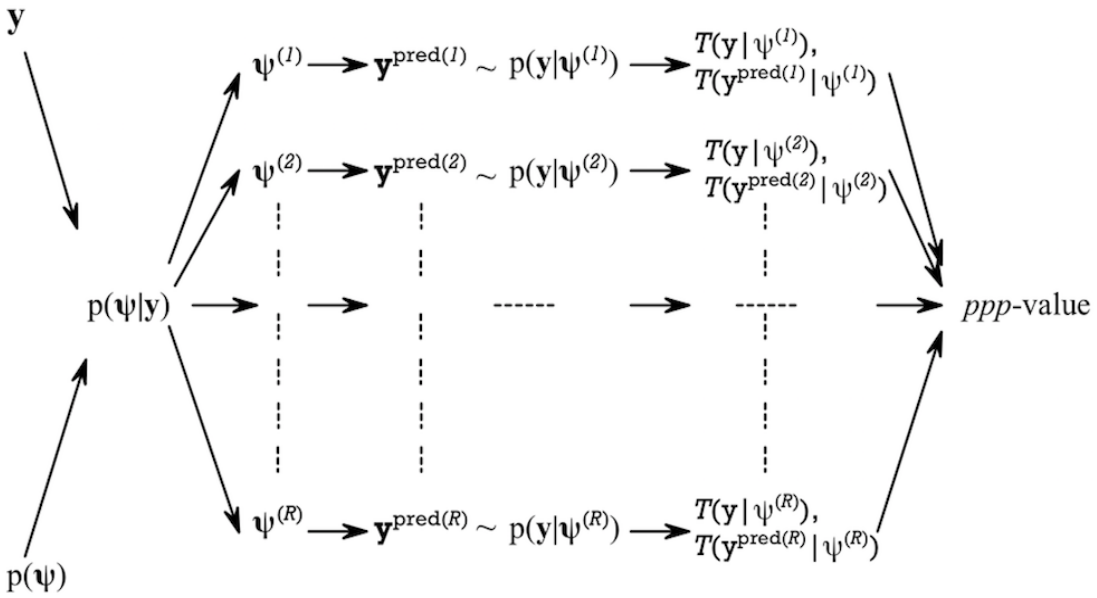

Figure 3.5: Graphical demonstration of posterior predictive model checking (PPMC) Method.$\mathbf{y}$, observed data; $\mathbf{y}^{\mathbf{pred}}$, predicted data; ψ, model parameters; $p(\psi)$, prior distributions of model parameters; $p(\psi|\mathbf{y})$, posterior distributions of model parameters; $T(.)$ discrepancy measures.

(Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014); and, newly proposed (3) M index, which will be discussed in detail subsequently.

Item response based W statistic. The W index is computed from performing a residual analysis from applying the Rasch model (Rasch, 1960) to a set of examinees' item responses (Wright & Stone, 1979). As a consequence of this parsimoniously parameterized model, analyses require relatively small sample sizes (i.e., the number of examinees) to produce reasonable data-model fit (Linacre & Wright, 1994). The computation of the W index follows

$$W_{ij} = \frac{[Y_{ij} - P_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{3.19}$$

where $P_i(\theta)$ is the probability of correctly answering item i given the ability estimate, θ, Y_{ij} is the dichotomous response (0, 1) of item i for a specific person j.

RT-based L statistic. Marianti et al. (2014) suggested an RT-based item-fit statistic, named the L statistic. Parameters used for calculating the L statistic are estimated based on the RT model proposed by W. J. van der Linden (2006c). The L statistic is formulated as

$$L_{ij} = \frac{[ln(t_{ij}) - \beta_i + \tau_j]^2}{\sigma_{e_i}^2}, \tag{3.20}$$

where t_{ij} is the response time for test taker j on an item i, β_i is the time-intensity parameter that is the averaged population time required for answering that item, τ_j is the speediness parameter for each test taker, and σ_{e_i} is defined as $1/\nu_i$.

Visual fixations based M statistic. To evaluate the data model fit based on the visual fixation counts, a visual fixation counts based item-fit M statistic is proposed. The M statistic is a residual-based model-fit measure, which is constructed from a summation of the variance weighted squared residuals defined as the differences between the observed outcome, c_{ij}, and predicted value, $E(c_{ij})$. (Cochran, 1952; Fox & Marianti, 2017b; **?**). The M statistic is formulated as

$$M_{ij} = \frac{[c_{ij} - exp(m_i - \omega_j)]^2}{\sigma_{ij}^2}, \tag{3.21}$$

where c_{ij} is the visual fixation counts for test taker j on an item i, m_i is the visual-intensity parameter that is the averaged population visual efforts required for answering that item,

m_i is the individualized visual engagement parameter, and σ_{ij}^2 is the variance of the visual fixation counts, which is defined as $\sigma_{ij}^2 = exp(m_i - \omega_j) + exp(2(m_i - \omega_j))/s_i$.

Those three statistics were utilized as different discrepancy measures (see Figure 3.5), to calculate the PPP-values evaluating the item by person-level date-model fit for item responses, response times, and visual fixation counts, respectively. Having a PPP-values close to 0 based on a discrepancy measure would indicate problematic data-model fit, and implies the proposed model fails to sufficiently regenerate the data (Sinharay et al., 2006). In the results section, the item-wise data-model fit for the item responses, response times, and visual fixation counts will be calculated by averaging over all the persons' PPP-values for each item, and the results will be reported.

3.4 Integration of Bio- and Psychometrical Infomation into Machine Learning Methods for Detecting Aberrant Behaviors

In this study, data mining algorithms introduced in Chapter 2, a class of methods for clustering observations, will be a useful platform for combining various information that can detect different types of aberrant behaviors. The sensitivity will be tested to detect aberrant behavior will be potentially increased by incorporating not only process and biometrical data as inputs into these algorithms, but also indices based on traditional approaches. Additionally, in contrast to applications involving traditional IRT-based and RT methods, data mining algorithms will be used to examine both linear and nonlinear relations among variables, thereby increasing its flexibility to benefit from modeling interactions between background, psychometric, and biometric data. To classify three types of

test-takers mentioned in the section 3.1, different detection methods will be applied. Two representative data mining methods: unsupervised (K-means clustering) and supervised (random forest) learning methods will be investigated. In addition, an item response-based person-fit index H^t and a response time based person-fit index l^t will be calculated. A real dataset will be analyzed to compare the various detection methods.

3.4.1 Data Normalization

Data normalization, also called feature scaling, is a process to transfer the range of different independent variables or features onto a common scale. This process will be performed for the supervised learning methods in this study (see, Vapnik 1963, for a discussion of the advantages for supervised learning methods). For traditional and unsupervised learning methods, however, data normalization will not be performed because they are either invariant to monotonic transformations of individual features or because they can change the original characteristics of the data. Also, many traditional methods are parameter model-based clustering methods, which do not rely on the geometric distance measures for classification. Thus, it is not necessary to implement feature normalization (e.g., Dubes & Jain, 1988; Hastie et al., 2009; Strobl, Malley, & Tutz, 2009).

Four types of scaling methods have drawn much attention in practical usage. These are scaling by variance, mean normalization, scaling by minimum and maximum values, and scaling to unit length (Fukunaga, 2013). In this study, variables were scaled by its minimum and maximum values, implying that all independent variables will be scaled to the range [0, 1]. An advantage of this type of scaling is that it can accommodate binary item responses.

3.4.2 Feature Selection

The purpose of the feature selection is to choose the essential featuers to improve the overal classification accuracy. After data normalization, to best capture the hidden insights from the dataset, and make inferences from the model; a set of features $\{x_{\{1\}}, x_{\{2\}} \ldots, x_{\{m\}}\}$, also called independent variables or attributes, will be selected from the total number of potential input features $\{x_{\{1\}}, x_{\{2\}} \ldots, x_{\{M\}}\}$, where $m < M$. This is essentially a filtering process (see discussion in John, 1994; Koller, 1996; Miller, 1990). Implementing this selection process increases the interpretability of the model making it becomes less complex and more parsimonious. In this study, two filtering methods will be used as feature selection methods. These methods are: (1) Pearson correlation between any pair of input variables, and (2) the variable importance index (VII). VII works by randomly permuting the values of a feature (input) variable, which brakes the original relation between the variable and other variables. Then, the permuted feature is used again with other unpermuted features for making predictions. If the prediction accuracy decreases, the gap before and after the permutation on a specific variable, averaged across all the trees, will be used as a measure of variable importance. Usually, the permutation importance is calculated based on the out-of-bag (OOB) subjects, which are the samples left behind for training the classification trees. The OOB samples can be utilized as a test dataset for evaluating the prediction accuracy.

The VII for each tree t is defined as

$$VI(X_j) = \frac{1}{n_t} \sum_t^{n_t} (\widehat{errOOB}_t - errOOB_t)$$

where the $\widehat{errOOB}_t$ represents the classification errors based on the permuted OBB sam-

ples for a specific tree t. $errOOB_t$ denotes the classification errors based on the OBB sample without any permutations for the tree t. The VII for a feature is averaged importance score across all the trees built in the forest.

Based on the methods mentioned above, a set of variables that will be considered in the analyses are listed in the Table 3.1.

Table 3.1: Input psychological and biological variables for data mining methods

Psychological variables	**Biological variables**
Item responses	Fixation duration (sec)
Item response times for each item	Number of fixations
Total response time for the entire test	Average time to 1st review (sec)
Self-reported motivation indicators	Averaged revisits (average number of revisits made to the AOI)
Latent speediness	Latent visual engagement levels
Efforts of test preparation indicators	Revisit indicator to the AOI (0/1)
Ten-item personality inventory	

3.4.3 Outcome Measures and Expected Results

Based on the methods mentioned above, a final set of variables will be selected from results of the two feature selection methods. For each of the proposed methods, a method-based classification of aberrant and non-aberrant test takers will be obtained. Sensitivity and specificity will be used as outcome measures for evaluating the performance of different methods. Sensitivity is defined here as the percentage of test takers that are identified as aberrant and that are classified as aberrant by the particular method. It will be calculated as $100 \times [TP/(TP+FP)]$. TP stands for the true positive (i.e., TP are those test takers correctly classified as aberrant) and FP stands for the false positive (i.e., FP are those test takers incorrectly classified as non-aberrant). Specificity on the other hand, is defined here as percentage of test takers that are identified as non-aberrant

(normal) and that are classified as non-aberrant by the particular method. Again, specificity will be computed as $100 \times [TN/(TN+FN)]$. TN stands for the true negative, and FN stands for the false negative. As a mean of comparison of the performance of aberrant test taking behavior detection, the sensitivity and specificity rates will be reported and compared across the proposed methods.

3.5 Research Significance

First of all, this study will explore the ways of incorporating biological information into traditional psychological methods by developing new models and utilizing data-mining algorithms to better understand test takers behaviors. This work extends methodological work has the potential not only to aid administrators of large scale assessments to ferret out aberrant behaving examinees, but also can lead to future research in the area of test security. Second, this study has the potential to create and develop methods that may very well flag aberrances with increased accuracy. An advantage of the newly proposed statistical methods is that they would not be based solely on one source of information, such as item responses, but rather on multiple sources of information about test-takers including data stemming from bio-information technologies and integrating log file information. Ideally, all of this information will be the inputs to be aggregated using highly efficient computational methods such as cloud computing in the data-mining framework. Third, through the methodological investigations and analyses using empirical data, the signal-to-noise ratio (SNR) could be increased, which means greater and more accurate classification with high sensitivity to detect different types of aberrant testing behaviors.

Fourth, the performances of the different methods (e.g., item response based person-fit analysis, response time based fraud detection methods, K-means clustering, random forest) in detecting different types of aberrant behaviors such as pre-knowledge cheating and copy cheating in terms of classification accuracy will be further manifested by this study.

Chapter 4: Results

In this chapter, the results of the study presented in Chapter 3 will be elaborated. First, data visualization and exploratory data analysis (EDA) will be conducted. Then, the results of three innovative eye-gaze fixation models are presented. In addition, a proposed three-way factor model – jointly modeling item responses, RTs, and visual fixation counts – will be fitted to the data in different experimental conditions. Therefore, the behavioral pattern differences across various experimental conditions are assessed. Lastly, results from implementing both unsupervised and supervised learning methods that classify types of test-takers will be presented.

4.1 Summary Statistics of the Collected Data across conditions

A total of $N = 335$ university students who had normal or corrected vision were recruited for the study. Students were asked to take a test consisting of $I = 10$ questions related to verbal reasoning. The test material used for the current study followed the structure of a high-stakes credentialing exam. Data from subjects who did not complete the designed tasks were excluded from the following analysis, leaving $N = 298$ participants in the study. Table 4.1 lists the numbers of subjects in each condition.

Table 4.1: Number of subjects in each condition

	Condition 1	Condition 2	Condition 3
,Number of subjects	93	98	107

Note: Condition 1: participants in the control condition who did not receive any test preparation materials. Condition 2: participants received items that were similar to their exam. Condition 3: participants in the third condition would receive similar exam questions and the answer key.

The collected dataset includes 103 variables, which measure visual engagement, working speed, responding accuracy, content revisits, test anxiety, and personality. The variable names are listed in Table A.1 (see Appendix A). Table B.1 (see Appendix B) shows the summary statistics of all the variables across the different experimental conditions.

4.2 Data Visualization Across Different Experimental Conditions

Based on the descriptive statistics of the collected data, it is not hard to gain insights about the group differences by comparing the means for each variable across three experimental conditions. In order to have better understanding about the data and to properly model it for accurate inferences, eventually, the collected data will be explored by showing the bivariate scatterplots of the major variables, which are quite useful and straightforward for interpreting tends and the associations among the key variables. All the scatterplots were created based on the total scores for each individual, see Figure 4.1. For instance, on the top left of Figure 4.1, the total scores were calculating by summing up the 10 item scores. By visualizing the key variables, it is helpful to figure out the most appropriate means for answering our research questions.

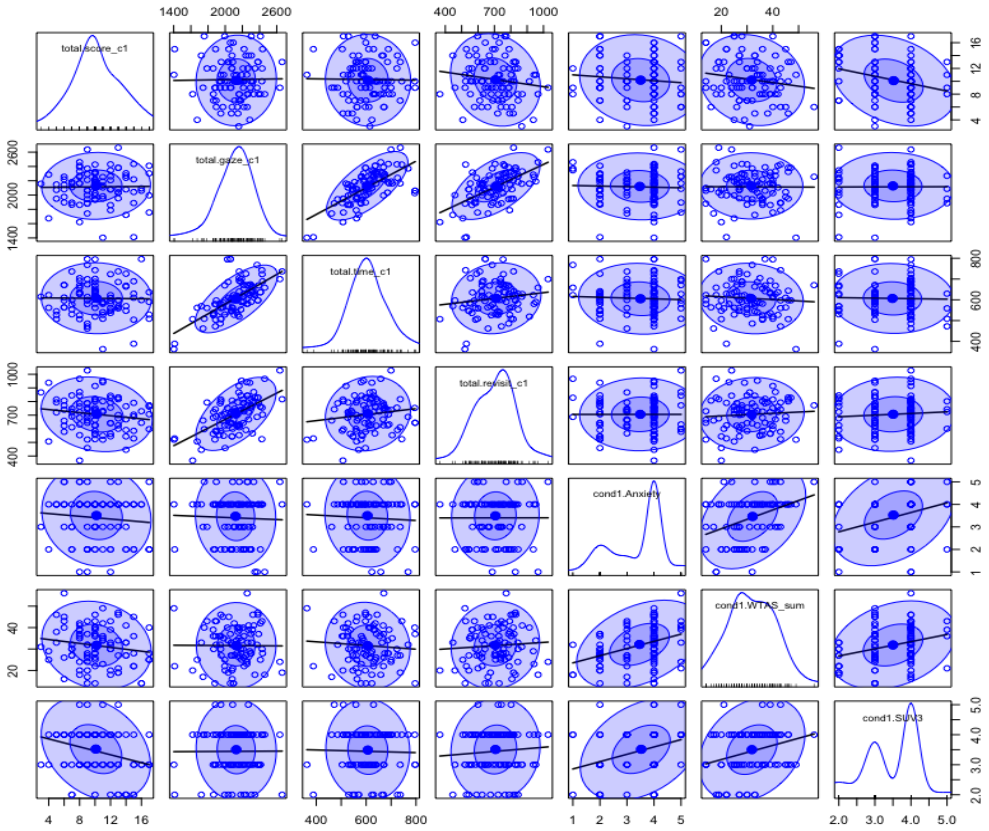

Figure 4.1: Scatterplots of essential variables under condition 1. The variable names showing in the matrix from the top left to the bottom right are: total.score, total.gaze, total.time, total.revisits, total.anxiety.score, WTAS.score, personality score. The distribution of each variable is listed on the diagonal of the plot matrix. The bivariate scatterplots are listed on the off-diagonal.

Figure 4.1 displays the scatterplots for the essential variables for assessing the behavioral patterns of test-takers in condition 1. For the first three panels showing the distributions, listed from the top left on the diagonal of the matrix, the total scores, total gaze counts, and total response times are essentially normally distributed. These factors will be jointly modeled to uncover associations among the latent constructs endorsed by these

indicators. For other Likert-scaled indicators measuring test-anxiety and personality, although the distributions look erratic, they will be used to as the input features for each of the data mining methods. Because the data mining methods are non-parametric, they depend less on the underlying distribution of the input features.

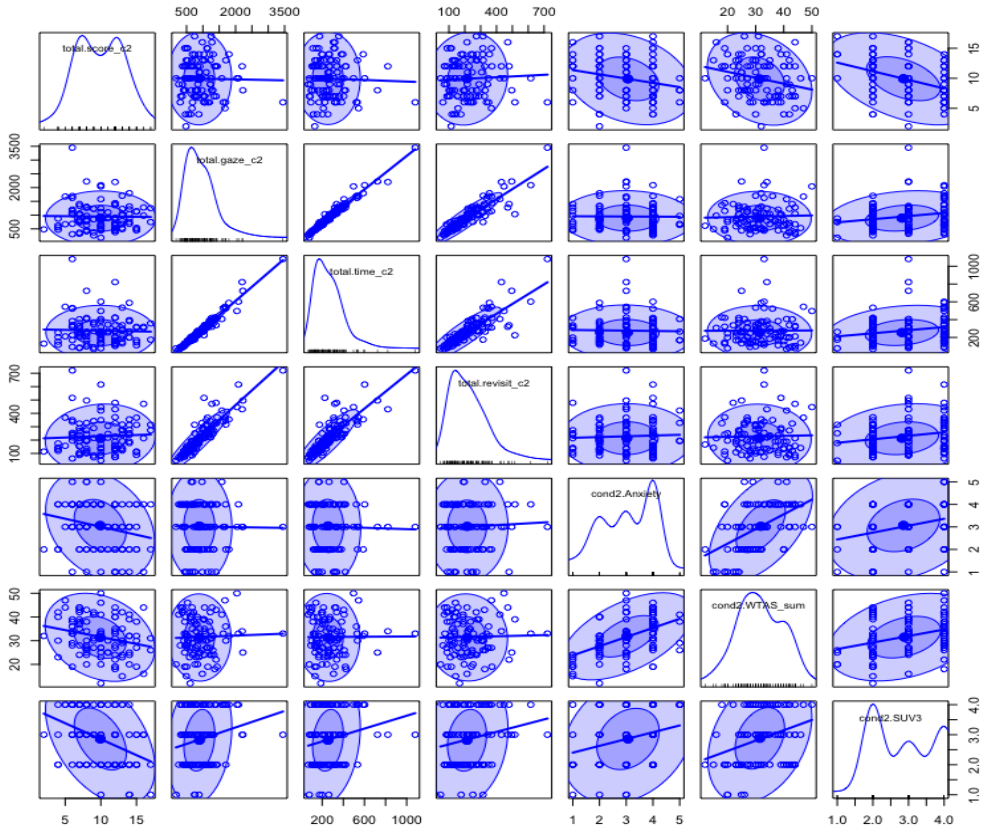

Figure 4.2: Scatterplots of essential variables under condition 2.

Figure 4.2 shows the scatterplots for the same set of variables demonstrated in Figure 4.1 for assessing test-takers behavioral patterns under condition 2. Focusing on the distributions in the first three panels listed from the top left on the diagonal (total scores, total gaze counts, and total responding times) their panel plots show bimodal and skewed

distributions, which are different from the ones shown in condition 1. The bimodal distribution may indicate a mix of two groups of test-takers with different test-taking strategies, responding to the items in different ways. In addition, the total gaze and total response time are skewed to the right, which means, on average, test-takers tend to spend a shorter time finishing the items on their tests. For other Likert-scaled indicators measuring test-anxiety and personality, the distributions became more skewed and peaked compared than the ones demonstrated in Figure 4.1.

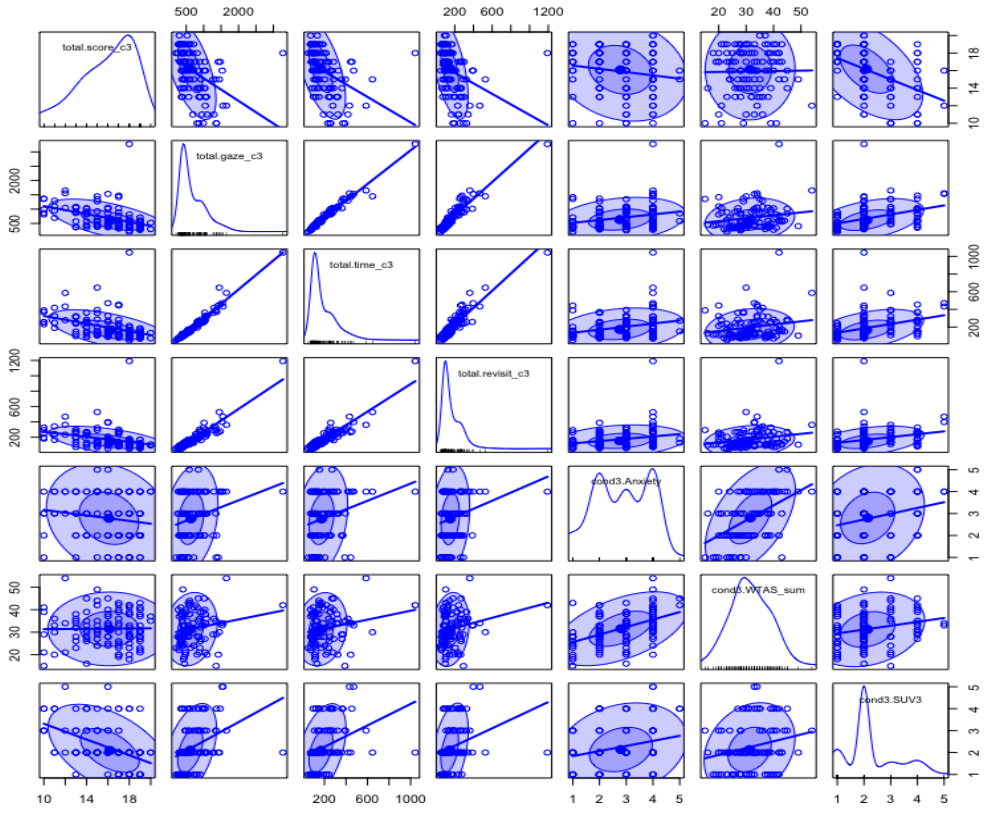

Figure 4.3: Scatterplots of essential variables under condition 3.

Figure 4.3 demonstrates test-takers behavioral patterns under condition 3. In general, all the distributions listed on the diagonal are relatively more skewed with less variability. Looking at the first three panels listed from the top left, their distributions are very skewed with high peaks, which indicate the responding behavioral patterns of test-takers under condition 3 are dramatically different from test-takers in the other conditions. The results show that test-takers in this group correctly answered the items more rapidly with less visual attention. Also, all the test-takers in condition 3 behaved more alike.

4.3 Negative Binomial Visual Fixation Models

In this section, visual fixation, an essential eye-tracking indicator, is modeled to reflect the degree of test engagement when a test taker solves a set of questions. Three negative binomial models demonstrated in Chapter 3 were evaluated for modeling visual fixation counts produced by test takers answering questions. The three models are: 1) negative binomial fixation model (NBFM); 2) negative binomial fixation model with linear trend (NBFM-LT); and, negative binomial fixation model with quadratic trend (NBFM-QT).

4.3.1 Item Parameter Estimates

Table 4.2 illustrates the parameter estimates of 10 items over the three proposed models. The item parameter estimates reflecting the visual intensity, $\hat{m}$, varied from 3.113 to 5.419. Item 1 was the least engagement-intensive item ($\hat{m} = 3.194$). In contrast, item 10 demanded the most visual effort from test-takers ($\hat{m} = 5.407$). One thing to note is

that the item visual intensities were higher for the last three items. This confirmed our presumptions in light of the fact that the last three questions were reading comprehension questions, which required more visual effort for test-takers. Also, the results indicated that item 1 was the most discriminating item on the test. The results were consistent across the three proposed models.

Table 4.2: Item parameter estimates

Model	NBFM		NBFM-LT		NBFM-QT	
Item	m	α	m	α	m	α
1	3.194(.029)	0.187(.010)	3.113(.053)	0.187(.010)	3.209(.023)	0.186(.010)
2	3.829(.024)	0.137(.007)	3.745(.051)	0.136(.008)	3.841(.017)	0.138(.007)
3	3.792(.027)	0.113(.009)	3.708(.052)	0.114(.010)	3.803(.022)	0.111(.009)
4	4.264(.044)	0.036(.003)	4.180(.062)	0.037(.003)	4.276(.040)	0.036(.003)
5	4.744(.022)	0.068(.007)	4.660(.050)	0.066(.007)	4.755(.015)	0.069(.006)
6	4.417(.032)	0.045(.004)	4.334(.055)	0.046(.004)	4.428(.028)	0.044(.004)
7	3.990(.035)	0.064(.005)	3.907(.056)	0.063(.005)	4.001(.029)	0.065(.005)
8	5.255(.020)	0.054(.006)	5.172(.049)	0.052(.006)	5.267(.115)	0.057(.005)
9	5.248(.025)	0.030(.003)	5.165(.051)	0.031(.003)	5.259(.018)	0.030(.002)
10	5.407(.020)	0.050(.005)	5.324(.049)	0.050(.006)	5.419(.012)	0.046(.004)

Note: NBFM: negative binomial fixations model; NBFM-LT: negative binomial fixations model with linear trend; NBFM-QT: negative binomial fixations model with quadratic trend. m: visual intensity parameter showing how much visual effort required for answering an item. α: visual discrimination parameter.

4.3.2 Person Parameter Estimates

Table 4.3 presents the estimated variance-covariance matrix of person parameters of the three proposed models. In general, overall small variability was seen in random test takers engagement. The variance of the test engagement fitted with the NBF model was 0.029 (SD = 0.004), which indicates a noteworthy contrast in individuals degrees of test engagement. Figure 4.4 displays the constant test engagement for each test-taker across the 10 items. For the NBF-LT model, the variability of the initial test engagement was 0.086 (SD = 0.03), and the variance of the slopes of test engagement was 0.017 (SD = 0.003). Moreover, the estimate of covariance between the initial test engagement and

the slope parameters based on the NBF-LT model was negative, which was -0.016 (SD = 0.006, Cor.= -0.418). This result shows that test takers who were highly engaged initially also exhibited increase in engagement with a lower growth rate than those whose initial engagement levels were lower as demonstrated in Figure 4.5. For the NBF-QT model, the variance of the initial test engagement was 0.053 (SD = 0.08), the variance of the slope was 0.003 (SD = 0.002), and the quadratic term variance was 0.0003 (SD = 0.0002). However, the fitted quadratic term was negligible. The individual engagement non-linear trajectories were presented in Figure 4.6.

Table 4.3: Variance covariance estimates

Model	NBFM	NBFM-LT		NBFM-QT		
Par.	Means(SD)	Means(SD)	Means(SD)	Means(SD)	Means(SD)	Means(SD)
ω_0	.029(.004)	.086(.03)		.053(.008)		
ω_1		-.016 (.006)	.017(.003)	0	.003(.002)	
ω_2				0	0	.0003(.0002)

In summary, three negative binomial distribution-based fixations models were proposed. The first model named as NBF model was defined by assuming constant engagement levels across all the items. A slope term and a quadratic term were added to the first model as two extensions. The NBF-LT model and the NBF-QR model used a parsimonious parameterization of the mean structure to capture changes in engagement exhibiting either linear or nonlinear trends. Results revealed measurement quantities and individual differences in their test engagement during problem-solving. Two item engagement parameters: engagement intensity and discrimination parameters, were designed for reflecting the visual efforts associated with an item. The estimated person parameter revealed individual test engagement differences across items. In the following section, the NBFM model will be utilized to jointly model visual fixation counts, item responses, and RTs,

which might help comprehensively assess the test-takers' performance.

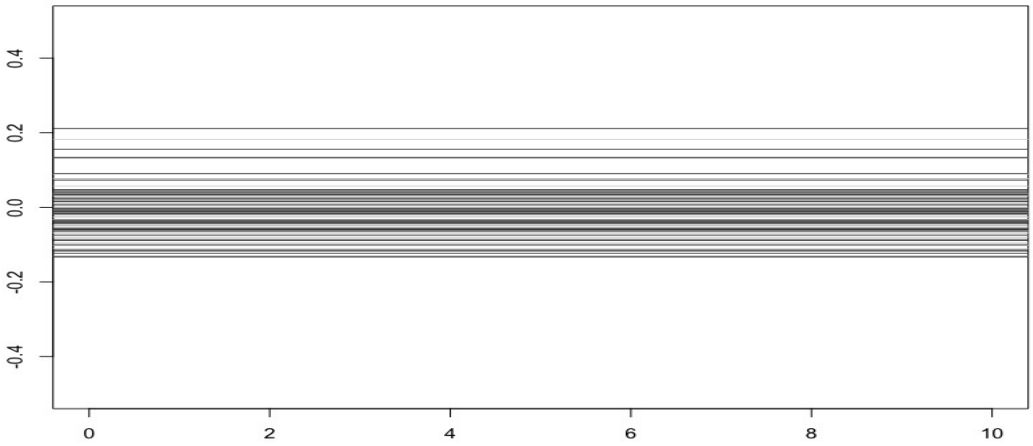

Figure 4.4: Individual test engagement estimates based on the negative binomial fixation model.

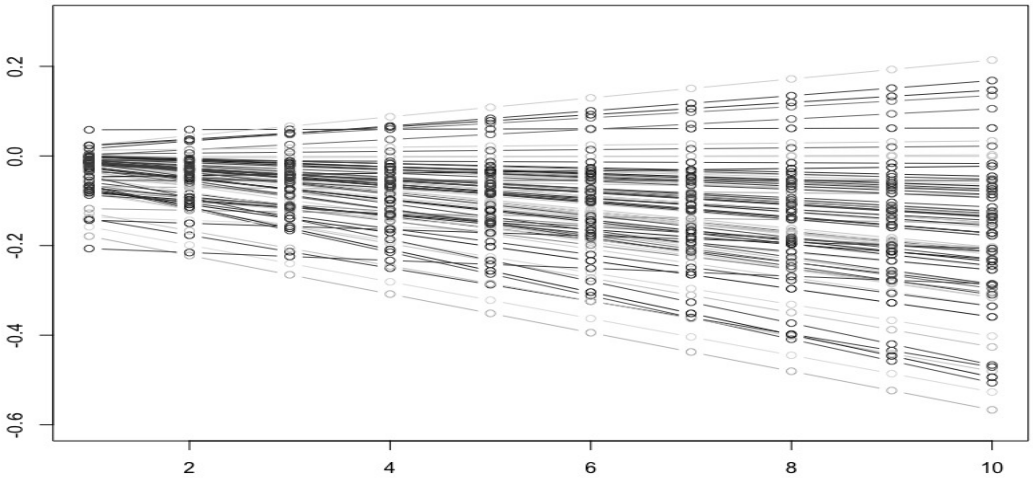

Figure 4.5: Individual test engagement estimates based on the negative binomial fixation model with linear trend.

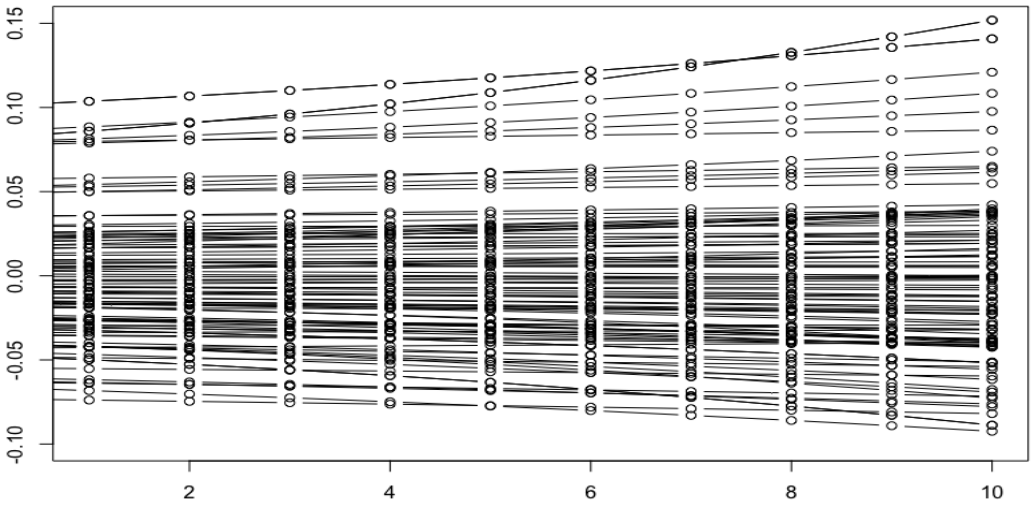

Figure 4.6: Individual test engagement estimates based on the negative binomial fixation model with quadratic linear trend.

4.4 Three-way Factor Model Parameter Estimates

In this section, as an example, hypothesized normally behaved test-takers from condition 1 were analyzed to gain insights on test-takers behavioral characteristics and item features based on the newly proposed three-way factor model, which jointly analyze item responses, RTs, and visual fixation counts simultaneously. Item parameter estimates based on the three measurement models were reported, which include item features such as item difficulty, time intensity, and visual intensity. Additionally, the associations between all the person-side latent constructs are discussed, demonstrating the test-takers behavioral characteristics. Additionally, the trade-offs among all the item-side parameters are discussed in a subsequent section.

4.4.1 Item Parameter Estimates

Item parameter estimates for condition 1 data, of N = 93 subjects who did not receive any test preparation materials, are presented below. The summarized item features across the three measurement models are reported in Table 4.4.

Table 4.4: Item parameter estimates of three-way factor model

Model	1-PL	RT		NBFM	
Item	b	β	ν	m	α
1	-1.05 (.241)	1.92 (.046)	0.43 (.032)	3.20 (.026)	0.19 (.010)
2	-0.52 (.235)	2.65 (.027)	0.21 (.017)	3.86 (.022)	0.14 (.007)
3	-0.56 (.228)	2.59 (.029)	0.23 (.018)	3.81 (.023)	0.11 (.009)
4	0.31 (.227)	3.02 (.043)	0.36 (.028)	4.28 (.043)	0.04 (.003)
5	0.97 (.237)	3.58 (.027)	0.21 (.016)	4.77 (.020)	0.07 (.007)
6	-0.14 (.215)	3.09 (.036)	0.32 (.024)	4.42 (.030)	0.05 (.004)
7	0.50 (.222)	2.67 (.039)	0.36 (.027)	4.03 (.033)	0.06 (.005)
8	0.99 (.240)	3.98 (.025)	0.18 (.014)	5.28 (.017)	0.05 (.006)
9	0.61 (.217)	3.97 (.025)	0.20 (.016)	5.26 (.022)	0.03 (.003)
10	0.09 (.215)	4.14 (.024)	0.18 (.014)	5.42 (.018)	0.05 (.005)

1-PL IRT Model. The 1-PL IRT model was fit to the data to estimate the item difficulty. Table 4.4 shows that the item difficulty parameter estimates, $\hat{b}$, varied from -1.03 to 0.98, which based on comparing the estimate to its standard deviation of posterior distribution (or looking at the 95% credible interval), were statistically significant from zeros. Among all the items, item 1 was the easiest item while item 8 was the most difficult item, which was expected since item 8 was a reading comprehension question while item 1 was a sentence equivalence question, which only consists of a single sentence and one blank.

Lognormal RT Model. In terms of the RT model, the time intensity parameter estimates, $\hat{\beta}$, ranged from 1.92 to 4.41, which were all statistically significant from zeros demonstrated in Table 4.4. On average, test-takers spent the least time responding to item

1 ($\hat{\beta}_1$ = 1.92). In contrast, item 10 required the most amount of time on average for test-takers to answer. In terms of visual discrimination parameter α varied from 0.03 to 0.19, the results indicated that item 1 was the most discriminating item on the test while item 10 was the least one.

NBF Model. For the NBF model, Table 4.4 outlines the parameter estimates of 10 items. The visual intensity parameter, $\hat{m}$, varied from 3.113 to 5.419. Item 1 was the least engagement-intensive item ($\hat{m} = 3.20$). In contrast, item 10 required the most visual effort from test-takers ($\hat{m} = 5.407$). which matched our expectations since the item 10 was a reading comprehension question requiring a lot of visual effort for test-takers to answer. Also, the results indicated that item 1 was the most discriminating item on the test.

4.4.2 Variance-Covariance Estimates

Table 4.5 exhibits the results of the parameter estimates of the variance-covariance of person- and item-domain at the structural level of the three-factor model (see Figure 3.4). This is of interest due to the structure-level item domain variance-covariance matrix between all the item parameters indicating pair-wise associations between item responses, RTs, and visual fixation counts.

4.4.3 Item-Side Variance-Covariance Structure

The estimated covariance between item difficulties and item visual intensities was 0.43 (Cor. = 0.59) with a 95% credible interval of 0.02 to 0.56 indicating that item difficulties were positively correlated with item visual intensities for the current test (see Figure 4.7). The estimated covariance between item difficulties and item time intensities was 0.42 (Cor. = 0.59) with a 95% credible interval of 0.29 to 0.87, which shows sig-

Table 4.5: Variance-covariance estimates of three-way model

Item Item Parameters			Person Parameters		
Variance-CovarianceParameters			Variance-CovarianceParameters		
	Mean	CI		Mean	CI
σ_b^2	0.701	(0.022,0.839)	σ_θ^2	0.488	(0.233,0.850)
σ_m^2	0.737	(0.292,0.875)	σ_ω^2	0.017	(0.013,0.007)
σ_β^2	0.731	(0.280,0.874)	σ_τ^2	0.021	(0.015,0.029)
$\sigma_{b,\beta}$	0.421	(0.292,0.875)	$\sigma_{\theta\omega}$	0.001	(-0.020,0.023)
$\sigma_{b,m}$	0.426	(0.022,0.555)	$\sigma_{\theta\tau}$	0.000	(-0.026,0.027)
$\sigma_{\beta,m}$	0.606	(0.189,0.733)	$\sigma_{\omega\tau}$	0.003	(-0.001,0.007)

nificant association between the item difficulties and time intensities for the given test. Moreover, the estimated covariance between item visual intensities and item time intensities was 0.61 (Cor. = 0.83) with a 95% credible interval of 0.19 to 0.73, which is also showing significant result that the item time intensities were positively correlated with item visual intensities (see Figure 4.7).

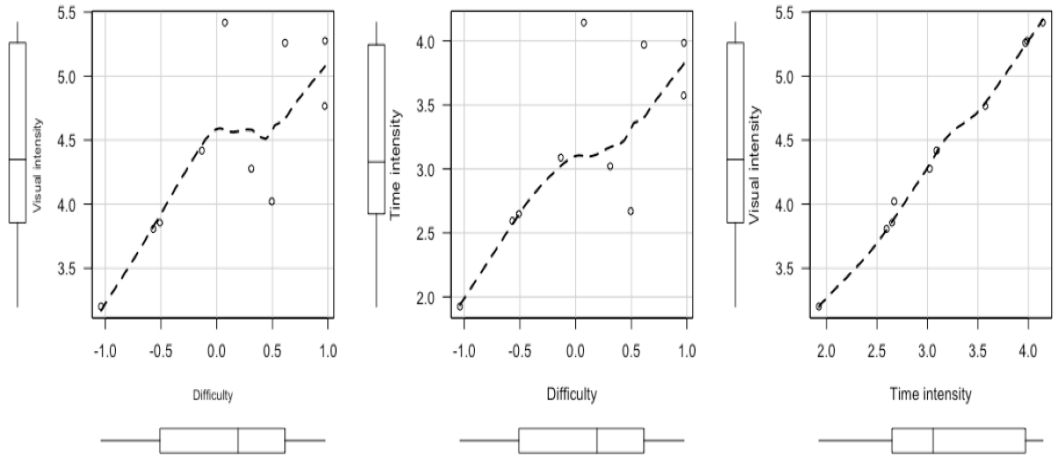

Figure 4.7: Scatter-plots for item parameter estimates. A loess non-parametric smoothed curve is plotted for each scatter-plot

4.4.4 Person-Side Variance-Covariance Structure

The person-side covariance $\sigma_{\theta,\tau}$, $\sigma_{\theta,\omega}$, and $\sigma_{\omega,\tau}$, illustrated in Table 4.5, were estimated to be 0.001 (95% credible interval: 0.026 to 0.027; Cor. = 0.005); -0.001 (95% credible interval: -0.023 to 0.020; Cor. = -0.011); and -0.003 (95% credible interval: -0.007 to 0.001; Cor. = -0.159) respectively, see Figure 4.8. Remarkably, all the person-side covariance estimates were not statistically significant from 0. The non-significant correlations could be a result of the subjects lacking motivation required to finish the designed assessment (Wise & Kong, 2005)

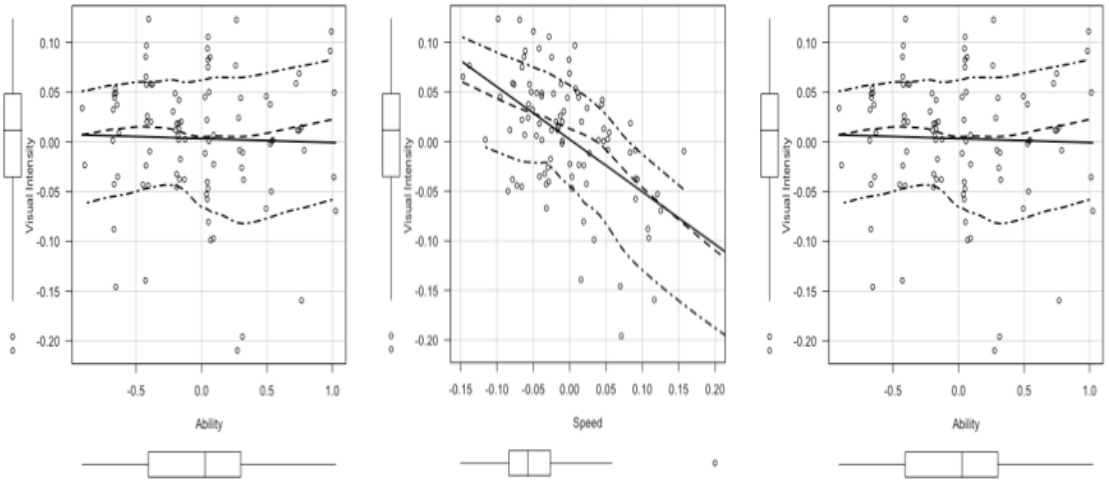

Figure 4.8: Scatter-plots for person parameter estimates. A loess non-parametric smoothed curve is plotted for each scatter-plot

4.4.5 Accessing the Item-Wise Data Model Fit

Posterior predictive model checking (PPMC; Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014) was used to evaluate model-data fit. Specifically, three item-fit statistics based discrepancy measures were used to calculate the item-wise data model fit for item responses, RTs, and visual fixation counts separately. Recall that the three item

fit statistics introduced in Chapter 3 are: 1) W index (Wright Stone, 1979); 2) l_t index (Marianti, Fox, Avetisyan, Veldkamp, Tijmstra, 2014); and, newly proposed 3) M_c index.

Figure 4.9 shows the PPMC-values for W, l_t, as well as M_c. In general, comparison of the PPMC-values for the three models across 10 items shows satisfactory data model fit. All the PPMC-values were above the 0.05 level, which is the cut-off of evaluating the model data fit. A PPMC value greater than 0.05 indicates that there are no systematic differences between the realized and predictive values, and thus an adequate data model fit. However, the IRT model was the least satisfactory because its PPMC-values over 10 items were systematically lower than the ones calculated based on the RT model and NBF model. Whereas, the PPMC-values calculated based on the W were still above the 0.05 threshold indicating agreeable fit. In addition, Figure 4.9 further demonstrates the details of item-wise fits for the given data set, which summarize the PPMC-values calculated over 2000 iterations. The three dash horizontal lines denote the percentiles of PPMC-values ranging from 0.05, 0.5 and 0.95, respectively. PPMC-values lying below the dash line at 0.05 levels would indicate the non-satisfactory data-model fit. All the PPMC-values in Figure 1 were above 0.05, which confirms that the proposed three-way joint model fits the data set well.

In all, the results are suggestive and demonstrated several interesting findings. First, given the condition 1 dataset, the fitted three-factor model reveals that the three latent dimensions were not statistically significantly correlated to each other, which demonstrates weak trade-offs among the accuracy, working speed, and visual engagement of test-takers when their eyes were being tracked. Second, the estimated structure of the measurement features of the proposed model was instructive for the practitioners in the testing industry.

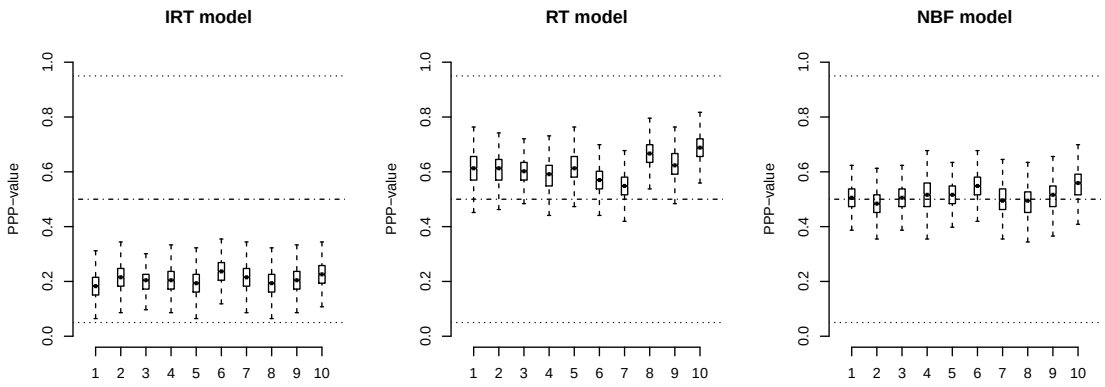

Figure 4.9: Posterior predictive p-values for 1-PL IRT model, log-normal response time model, and negative binomial visual fixation counts model over 10 items. The three dash horizontal lines denote 0.05, 0.5 and 0.95, respectively. The box-plots represent the item by person-level PPP values. The whiskers indicate the minimum and maximum PPP values for each item.

The results show that item difficulty, time intensity, and visual intensity were positively related to each other, which indicates that difficult items require more time and visual efforts for the test takers to answer. By running the three-way factor model, we could potentially comprehensively evaluate the test-takers performance in technology-enhanced environment such as game base testing or scenario-based virtual reality learning tasks.

4.5 Assessing Test-Taking Behaviors Across Different Experimental Conditions

To understand and evaluate the pattern differences in test-taking behaviors across distinct experimental conditions, a multiple-group joint three-way factor model of item responses, RTs, and visual fixation counts were fitted separately to the data in different conditions. Parameter estimates of the level-1 measurement models across the three conditions were reported. Moreover, the distinctions of the associations of the person-side

and the item-side parameters were reported by showing the corresponding covariance estimates across the contrasting experimental conditions.

4.5.1 Impact of Having Pre-knowledge of Test Items on Item Characteristics

To evaluate the impact of having pre-knowledge of test questions on the properties of test items (see Figure 4.10), Table 4.6 displays a comparison of item parameter estimates of the proposed model with respect to the three experimental conditions. In general, item difficulties ($\hat{b}$), time intensities ($\hat{\beta}$), and visual intensities ($\hat{m}$), on average, tend to show lower values in the condition 3 than the other two conditions (see Table 4.6). This is potentially attributable to the fact that test-takers tend to spend less time, and less visual effort on a test with which they were more familiar by practicing the similar items in advance.

Item difficulty estimates across conditions. In general, items, on average, appeared to be much easier in condition 3 than the other two conditions. Across item difficulties, $\hat{b}$ ranged from -1.06 to 0.99 in the condition 1, varied from -1.55 to 0.79 in the condition 2, and fluctuated from -4.09 to -0.24. Intriguingly, the difference ($b_{dff(1,2)}$) in item difficulties between the condition 1 and condition 2 is not as large as the difference ($b_{dff(1,3)}$) between condition 1 and condition 3 (see Figure 4.10), which means practicing items beforehand without knowing the answer keys has limited impact on item difficulties. In contrast, the item difficulties would decrease greatly if the test-takers practice the equivalent items with keys.

Time intensity estimates across conditions. Similarly, test-takers who practiced

the items or knew the answer keys beforehand tend to take less time to finish their tests. By averaging the time intensities across the 10 items, $\bar{\hat{\beta}}$ (the averaged time intensity) is 3.21 in the condition 1; 2.367 in the condition 2, and 2.102 in the condition 3 (see Table 4.6). By taking the exponential of each averaged time intensity estimate, the unit of $\bar{\hat{\beta}}$ were converted into seconds. On average, test-takers in condition 1 took about 25 sec to finish an item, ones in condition 2 used about 11 sec, and ones in condition 3 took about 8 sec. The results show that, on average, that test-takers in condition 3 who were practicing items beforehand with answer keys worked three times faster than the ones in condition 1 who did not receive any test preparation materials on answering an item.

Visual intensity estimates across conditions. A trend of visual intensities similar to the summarized response patterns in the previous session was observed, which indicates test-takers familiar with the items tend to put less visual effort on searching for information to answer the questions (see Figure 4.10). By averaging the visual intensities across the 10 items, $\bar{\hat{m}}$ (the averaged visual intensity) is 4.427 in the condition 1; 3.707 in the condition 2, and 3.505 in the condition 3 (see Table 4.6). By taking the exponential of each averaged visual intensity estimate, the unit of $\bar{\hat{m}}$ were converted into counts. In general, test-takers in condition 1 generated about 84 fixation counts to finish an item, ones in condition 2 produced about 40 fixation counts, and ones in condition 3 created about 33 fixations. The results show that, on average, that test-takers in condition 3 put much less visual effort than the ones in the other two conditions on solving questions.

Table 4.6: Item parameter estimates across different experimental conditions

Condition	Model	1-PL		RT				NBFM			
	Item	b	sd	β	sd	ν	sd	m	sd	α	sd
C1	1	-1.05	0.24	1.92	0.05	0.43	0.03	3.20	0.03	0.03	0.012
	2	-0.52	0.24	2.65	0.03	0.21	0.02	3.85	0.02	0.03	0.006
	3	-0.6	0.23	2.59	0.03	0.24	0.02	3.80	0.02	0.04	0.011
	4	0.31	0.22	3.02	0.04	0.36	0.03	4.27	0.04	0.05	0.003
	5	0.97	0.24	3.57	0.03	0.21	0.02	4.76	0.02	0.03	0.007
	6	-0.14	0.22	3.08	0.04	0.33	0.02	4.42	0.03	0.03	0.004
	7	0.50	0.22	2,67	0.04	0.36	0.03	4.02	0.03	0.03	0.005
	8	0.99	0.24	3.98	0.03	0.18	0.01	5.27	0.01	0.04	0.007
	9	0.6	0.22	3.97	0.02	0.21	0.02	5.26	0.02	0.04	0.003
	10	0.09	0.23	4.14	0.02	0.18	0.01	5.42	0.01	0.04	0.005
C2	1	-0.67	0.24	1.81	0.06	0.44	0.03	3.07	0.06	0.12	0.011
	2	-1.55	0.29	1.81	0.06	0.39	0.03	3.03	0.05	0.13	0.012
	3	-0.48	0.22	1.99	0.05	0.34	0.03	3.22	0.05	0.12	0.010
	4	0.13	0.23	2.43	0.06	0.45	0.03	3.73	0.06	0.05	0.005
	5	0.64	0.24	2.74	0.06	0.42	0.03	4.01	0.05	0.05	0.004
	6	-0.21	0.23	2.34	0.06	0.42	0.03	3.66	0.05	0.06	0.005
	7	0.78	0.24	2.21	0.06	0.42	0.03	3.48	0.06	0.07	0.006
	8	0.66	0.23	2.90	0.07	0.56	0.04	4.32	0.06	0.03	0.002
	9	0.79	0.24	2.78	0.08	0.63	0.05	4.38	0.07	0.02	0.002
	10	-0.17	0.23	2.66	0.08	0.62	0.05	4.17	0.07	0.02	0.002
C3	1	-4.09	0.55	1.69	0.06	0.4	0.03	2.98	0.06	0.13	0.012
	2	-3.76	0.51	1.51	0.06	0.41	0.03	2.80	0.06	0.15	0.014
	3	-1.91	0.31	1.92	0.08	0.59	0.04	3.36	0.07	0.05	0.005
	4	-1.72	0.28	2.29	0.08	0.73	0.05	3.82	0.08	0.03	0.003
	5	-1.92	0.29	2.46	0.06	0.41	0.03	3.77	0.06	0.05	0.004
	6	-1.81	0.30	2.16	0.06	0.43	0.03	3.52	0.06	0.07	0.006
	7	-1.23	0.28	2.11	0.06	0.41	0.03	3.40	0.06	0.08	0.007
	8	-2.39	0.33	2.32	0.07	0.52	0.04	3.80	0.06	0.04	0.004
	9	-0.24	0.25	2.23	0.07	0.56	0.04	3.76	0.07	0.04	0.004
	10	-0.81	0.28	2.33	0.07	0.55	0.04	3.84	0.07	0.04	0.004

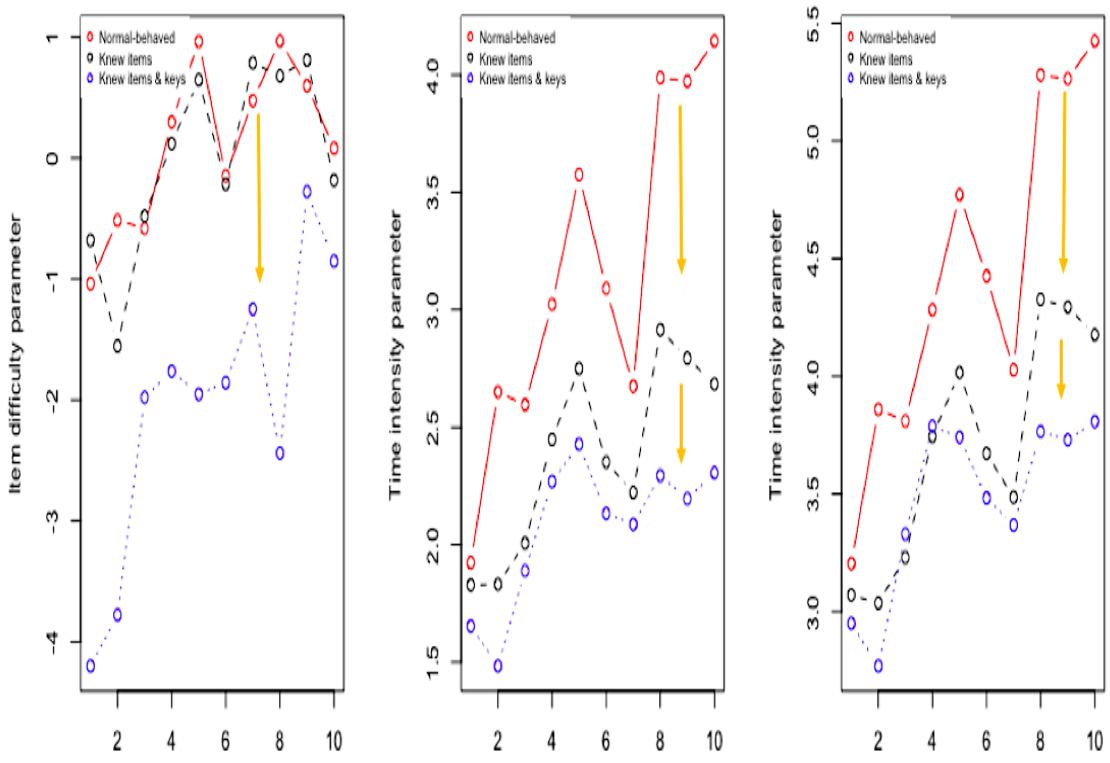

Figure 4.10: Item parameter estimates across distinct experimental conditions, and negative binomial visual fixation counts model. Red: Condition 1; Black: condition 2, and Blue: condition 3.

4.5.2 Impact of Having Pre-knowledge of Test Items on Test-Takers Behavior

Table 4.7 shows the impact of having pre-knowledge of test items on the test-takers behaviors. The behavioral pattern differences were demonstrated via comparison of the three person-side covariances, indicating association among the interested latent constructs (ability, working speed, and visual engagement) across the three experimental conditions. As a trend, as students gain more pre-knowledge of the test items the correlation between latent ability and working speed increased from 0.005 in condition 1 (95% credible interval: -0.023 to 0.020) to 0.672 (95% credible interval: 0.496 to 0.621) in

condition 3. The increased correlation between latent ability and working speed might be caused by test-takers in condition 3 receiving practice items with answer keys. Therefore, they answered more items correctly than the ones who did not receive any test preparation materials.

Table 4.7: Person-side correlation matrix estimates

Condtions	C1		C2		C3	
Paramter	Mean	CI	Mean	CI	Mean	CI
$Cor_{\theta,\omega}$	-0.011	(-0.244,0.227)	-0.193	(-0.437,-0.108)	-0.678	(-0.812,-0.505)
$Cor_{\theta,\tau}$	0.005	(-0.239,0.251)	0.24	(-0.020,0.327)	0.672	(0.496,0.810)
$Cor_{\omega,\tau}$	-0.152	(-0.359,-0.080)	-0.899	(-0.935,-0.886)	-0.91	(-0.942,-0.867)

Note: C1: condition 1; C2: condition 2; C3: condition 3; CI: credible interval; Cor.: Correlation.

In terms of changes in the trade-offs between the latent ability and visual engagement across conditions, Figure 4.11 shows that test-takers who were familiar with the test items tended to put less visual efforts on answering items. The correlation between those two latent constructs dropped from -0.011 in condition 1 (95% credible interval: -0.244 to 0.227) to -0.678 in the condition 3 (95% credible interval: -0.812 to -0.505). Similarly, negative trade-offs between the working speed and visual engagement was observed. The correlation ($r_{\theta,\omega}$) decreased from -0.152 in the condition 1 (95% credible interval: -0.359 to 0.062) to -0.910 in the condition 3 (95% credible interval: -0.942 to -0.867). This result infers that as test-takers knew the answer keys of practice items, they favored quickly answering the questions without elaborately paying attention to the content (See. Figure 4.11).

To summarize, the item parameters at the measurement level were significantly affected by the amount of pre-knowledge test-takers had, especially when the test-takers practiced equivalent items with answer keys. Correspondingly, the associations among

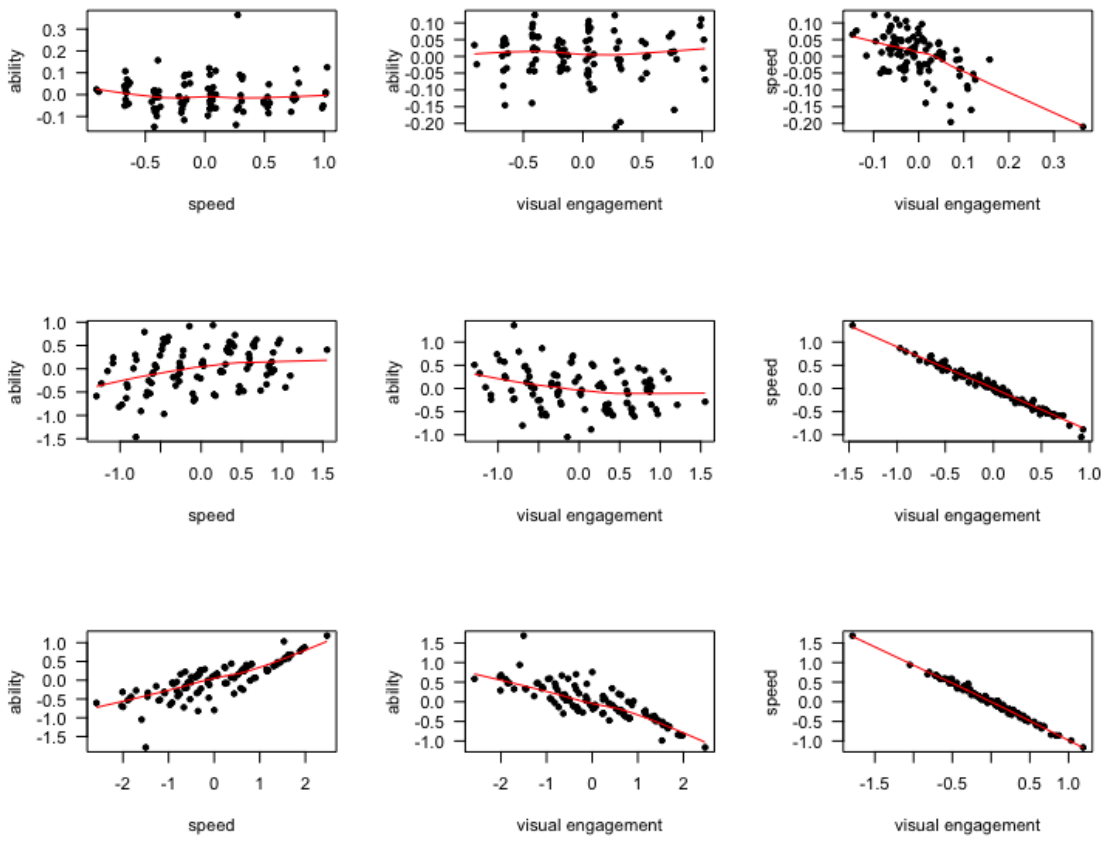

Figure 4.11: Scatterplots for person-side parameter estimates. A loess non-parametric smoothed curve is plotted for each scatterplot

person-side latent constructs (e.g., latent ability, working speed, and visual engagement) were greatly affected by the pre-knowledge, as well. The ability estimates of test-takers with pre-knowledge on test-items were positively correlated with their working speed. Their abilities were negatively associated with their visual engagement levels, and their working speed was negatively correlated with their visual engagement levels. In other words, test-takers might be inclined to finish their tests quickly by paying less attention to the content of items, and answering most items correctly.

In contrast, when the test-takers had access to test-preparation materials without

keys, their ability estimates were not correlated significantly with their working speed or visual engagement. However, a strong negative correlation between the working speed and visual engagement could still be expected. For instance, a high ability test taker statistically was likely to work either quickly or slowly on their test. In addition, the ones finishing their tests quickly paid less visual attention to the content compared to the ones who worked slowly. This is of interest because testing companies could potentially tackle identification of suspicious aberrant test-takers by matching their behavioral characteristics with our findings mentioned above.

4.6 Use of Person-fit Statistics to Classify Different Responding Behaviors

PFS's are widely utilized in the industry as approaches to identify how aberrant test takers behaved during their tests. In order to show differences in the performance between the PSF's and the data mining methods on separating aberrant cases from the normally behaved ones, the results of two representative PFS's studied previously were evaluated: an item response based PFS called l_z^* statistic,and a RT based PFS, called l^t statistic. The l_z^* statistic was calculated by using the *lzstar* function in the R package PerFit (Tendeiro, Meijer, & Niessen, 2016). Commonly, PFSs are computed when a small portion of aberrances exists in a dataset. To show the shortages of using PFSs to classify cases showing more than two types of aberrant-responding behaviors , the current dataset mixing of cases from three conditions were directly fitted to the *lzstar* function.

In this study, the results reported based on the PFS's would be questionable because there are significantly fewer hypothesized normally behaved cases in condition one than in

the other two aberrant conditions, which violates the basic assumption of using PFS's. As a result, it is hard to come up trustworthy cut-offs values since they were highly dependent on the underlying ability distribution of the test-takers, which was heavily contaminated by the large number of aberrant cases. Table 4.8 shows the sensitivity and specificity rates for using the two representative PFS's. It can be seen in Table 4.8 that having a large number of aberrant cases resulted in substantially low values in both sensitivity and specificity rates. Figure 4.12 indicates that PFS-based methods failed to separate the normally behaved subjects from the aberrances (l_z^* PFS: left panel; l^t statistic: right panel).

Table 4.8: Sensitivity and specificity for PFS IRT- and RT-based methods

% Consistent Decision	l_z^*PFS	l^tPFS
Sensitivity	0.04	0.00
Specificity	0.88	0.89
Overall accuracy	0.30	0.28

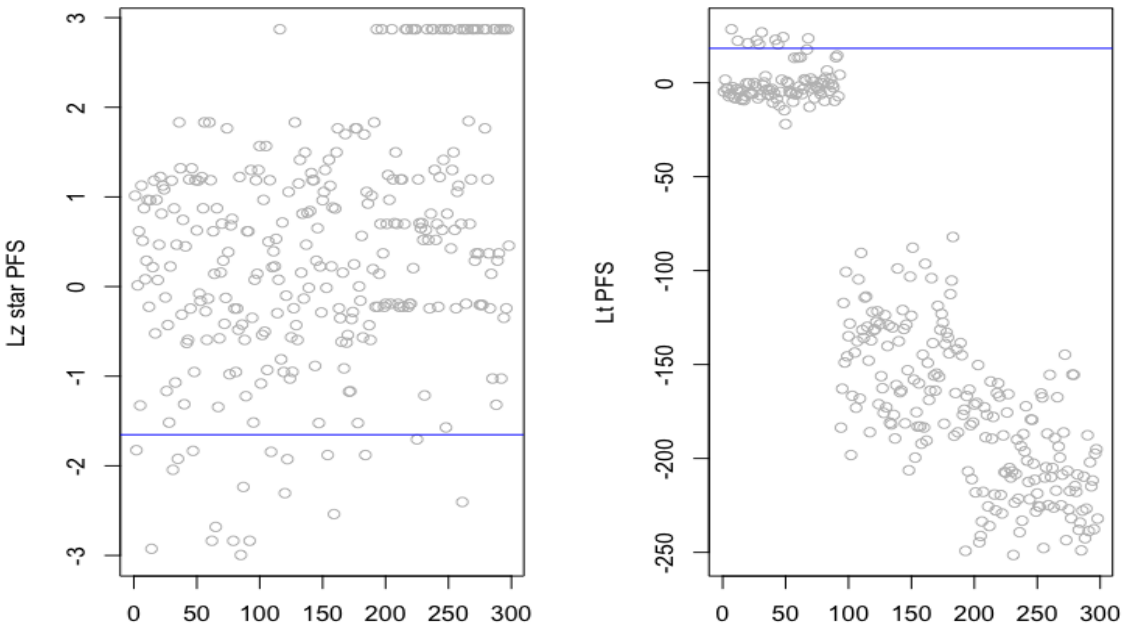

Figure 4.12: PFS's performance of classifying different type of responding behaviors. l_Z^* PFS is on the left, l_t PFS is on the right. The blue line indicates the cut-off

4.7 Use of Data Mining Methods to Classify Different Responding Behaviors

In this section, the use of representative data mining methods as an alternative group of methods to the PFS's was examined with a focus on classifying different types of test-takers who belong to distinct experimental conditions. Although the previously investigated item response and RT-based methods are popular in the industry as approaches to identifying aberrant test-taking behavior (e.g., pre-knowledge and copy cheating), the data mining-based methods have yet to be fully investigated. To show the benefits of using data mining methods over the traditional PFS's, two groups of data mining methods introduced in Chapter 3 were used to classify different responding behavior, which are: 1) unsupervised learning methods, and 2) supervised learning methods.

To properly use the unsupervised and supervised learning methods, data normalization was performed to put all the input variables onto the same scale by using the maximum-minimum method mentioned in Chapter 3. The final set of features was selected based on two methods discussed in Chapter 3: 1) Pearson correlation between any pair of input variables and 2) the variable importance index (VII).

To achieve the optimal classification accuracy, the feature selection was conducted. Among 60 total features, 13 features were highly correlated ($r \geq 0.9$), shown in Figure 4.13. Additionally, the rank of importance of features calculated based on the VII method was demonstrated in Figure 4.14. Among all the features, the top ranked features weighed heavily on classifying different responding behaviors were related to: 1) total scores, 2) revisits, 3) latent visual engagement, 4) fixation counts, and 5) RTs. In contrast, the per-

sonality measures and Westside test anxiety measures were less important for separating aberrances from normally behaved test-takers. Based on the VII values, the last 8 features were removed from the feature set to achieve optimal results: 1) WTAS1, 2) WTAS9, 3) WTAS8, 4) WTAS5, 5) Conscientious, 6) WTAS 10, 7) Agreeableness, and 8) WTAS6. In summary, 52 total features were selected for the rest of the analysis.

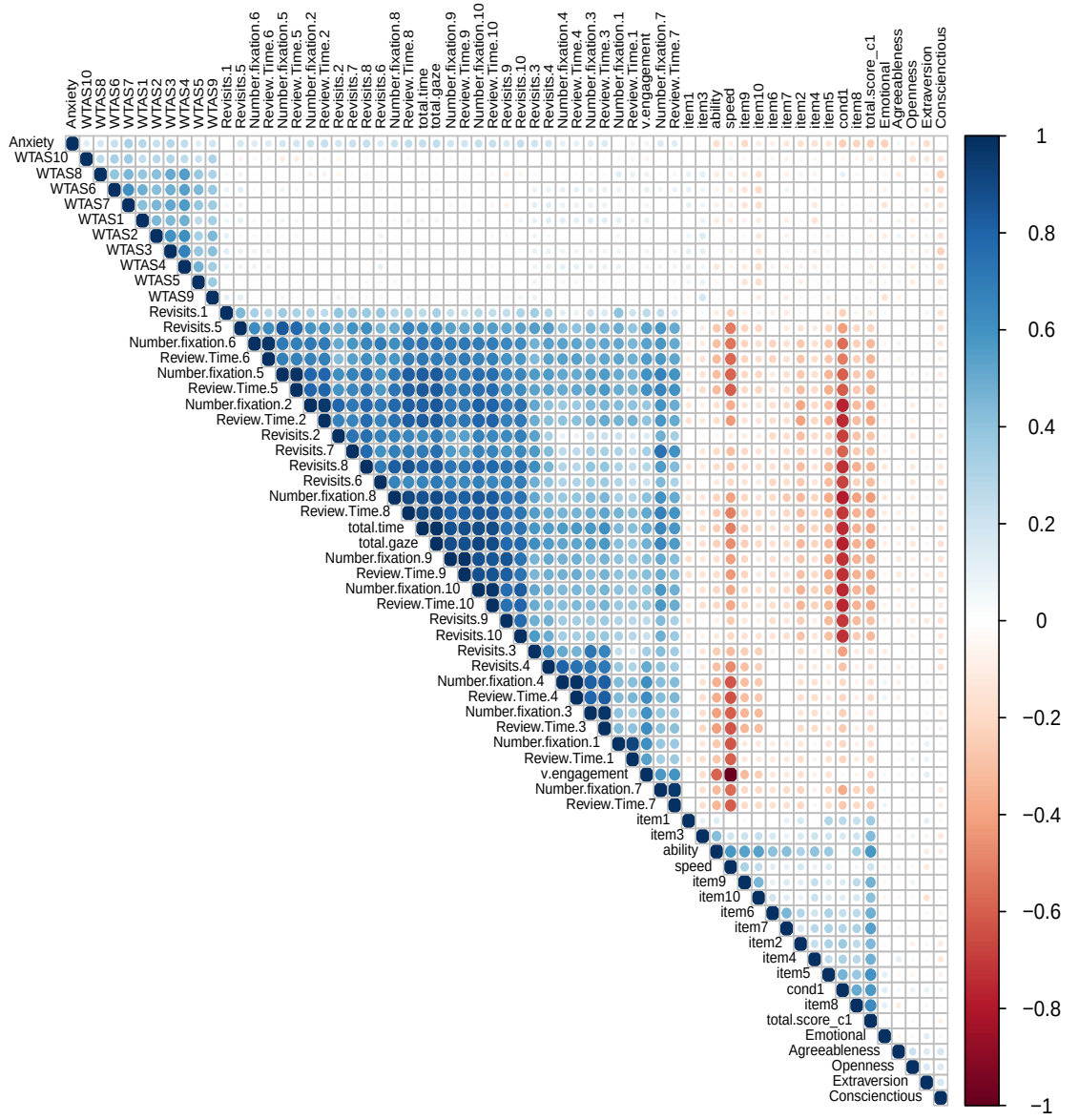

Figure 4.13: Pair-wise correlations between features

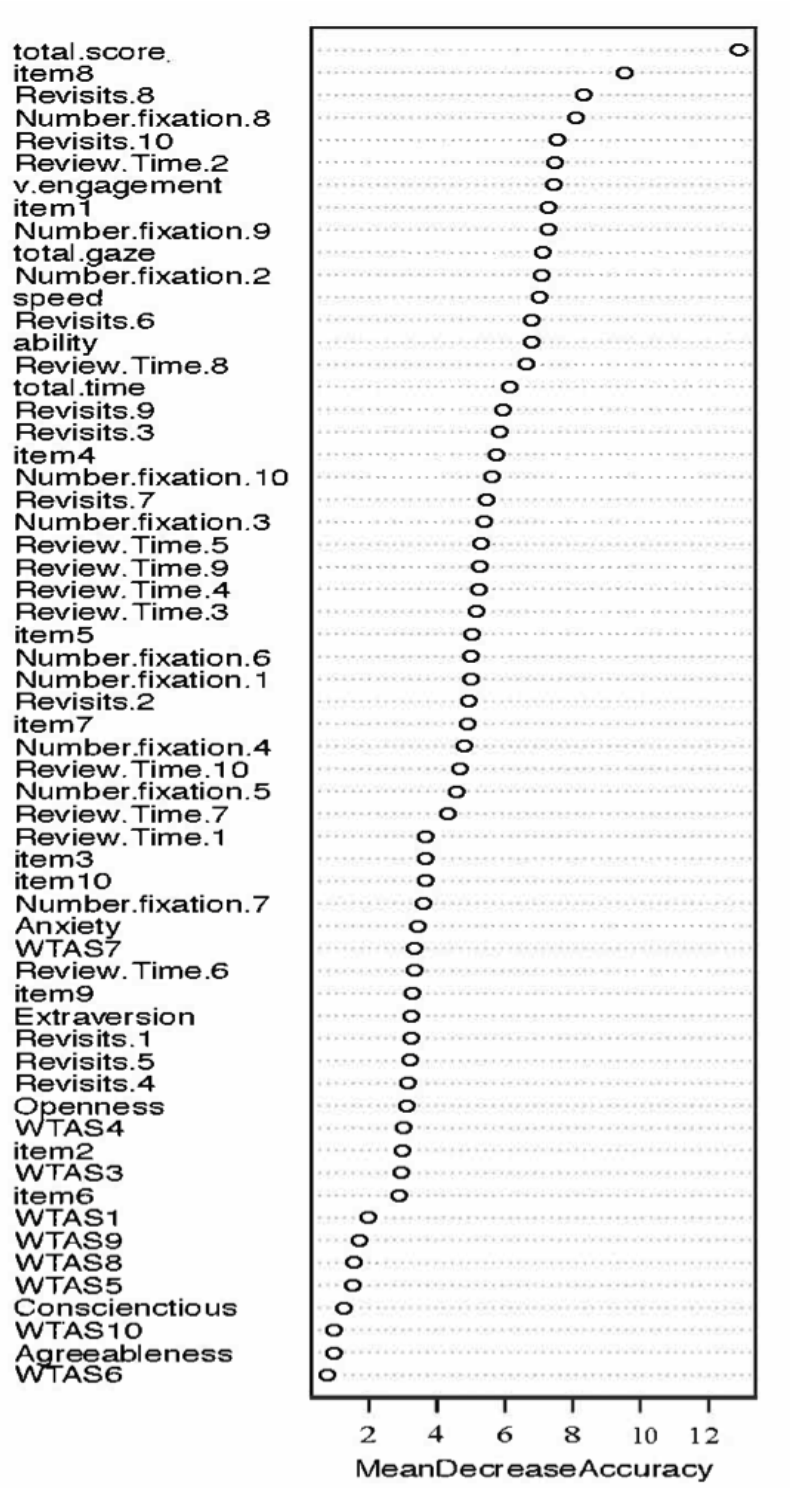

Figure 4.14: Feature importance

4.7.1 Representative Unsupervised Learning Methods

K-Means. After normalizing data and selecting essential features, a representative unsupervised learning method was performed to cluster subjects who belong to different test-taking experimental conditions. The focus was twofold here: (1) to evaluate the improvement in classification accuracy in using unsupervised learning methods, and (2) to reduce the error rates of misidentifying aberrances. The specificity and sensitivity were calculated along with the accuracy rate across different conditions, and results were summarized based on 10-folds cross-validation. After applying the K-means method, three clusters of test-takers were identified based on their responding behavioral characteristics. This is because the elbow point of the dashed line in Figure 4.15, showing that the total within-cluster sum of squares, drifted much less after the elbow point, which is $K = 3$ (the optimal number of clusters). Within each cluster, the sums of squares were 293.9 (cluster 1), 191.55 (cluster 2), and 342.24 (cluster 3) separately. The ratio of the within-cluster sum of squares to the total sum of squares is 0.73.

Table 4.9: Classification accuracy for K-Means methods with three groups

	True label		
	1	2	3
Sensitivity	0.989	0.575	0.877
Specificity	0.946	0.945	0.823
Overall accuracy			0.812

An evaluation of the classification of all the test-takers based on the K-Means method is demonstrated in Table 4.9. Table 4.9 indicates that subjects who belong to condition 1 and 3 were well classified with sensitivity rates of 98.9% and 87.7%. However, 33 subjects with the correct label of condition 2 were classified incorrectly into

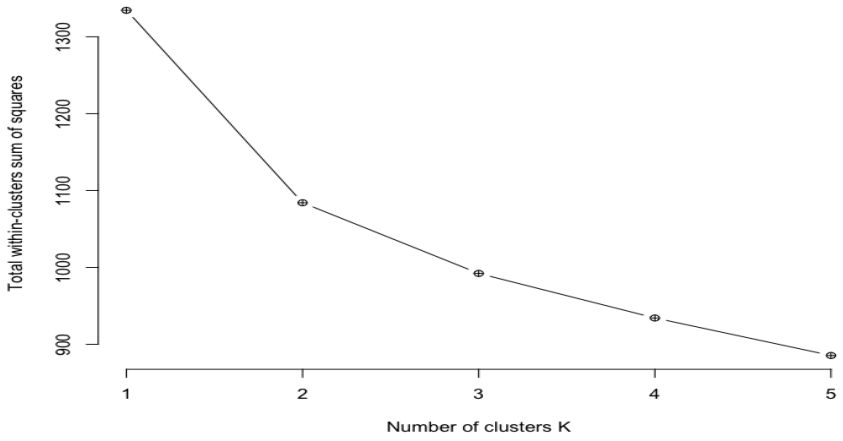

Figure 4.15: Number of optimal clusters based on K-Means method

condition 3, which indicates the K-means method is limited to differentiate the condition-2 and 3 subjects with the selected input features. However, Table 4.10 shows that both sensitivity (99.4%) and specificity (89.3%) were high after combining conditions 2 and 3 into one class, all of which have certain amount of pre-knowledge on test items.

Table 4.10: Sensitivity and specificity for K-means methods with two groups

% Consistent Decision	K-Means
Sensitivity	0.994
Specificity	0.893
Overall accuracy	0.96

As an example, Figure 4.16 visualizes segregations among the three clusters as a result of applying K-Means method to the data based on two features: 1) the number of fixations, and 2) the number of revisits across ten items. From Figure 4.16, item 9 shows a clear boundary between the subjects who belong to condition 1 (circles) and the others (condition 2 - crosses and condition 3 - triangles).

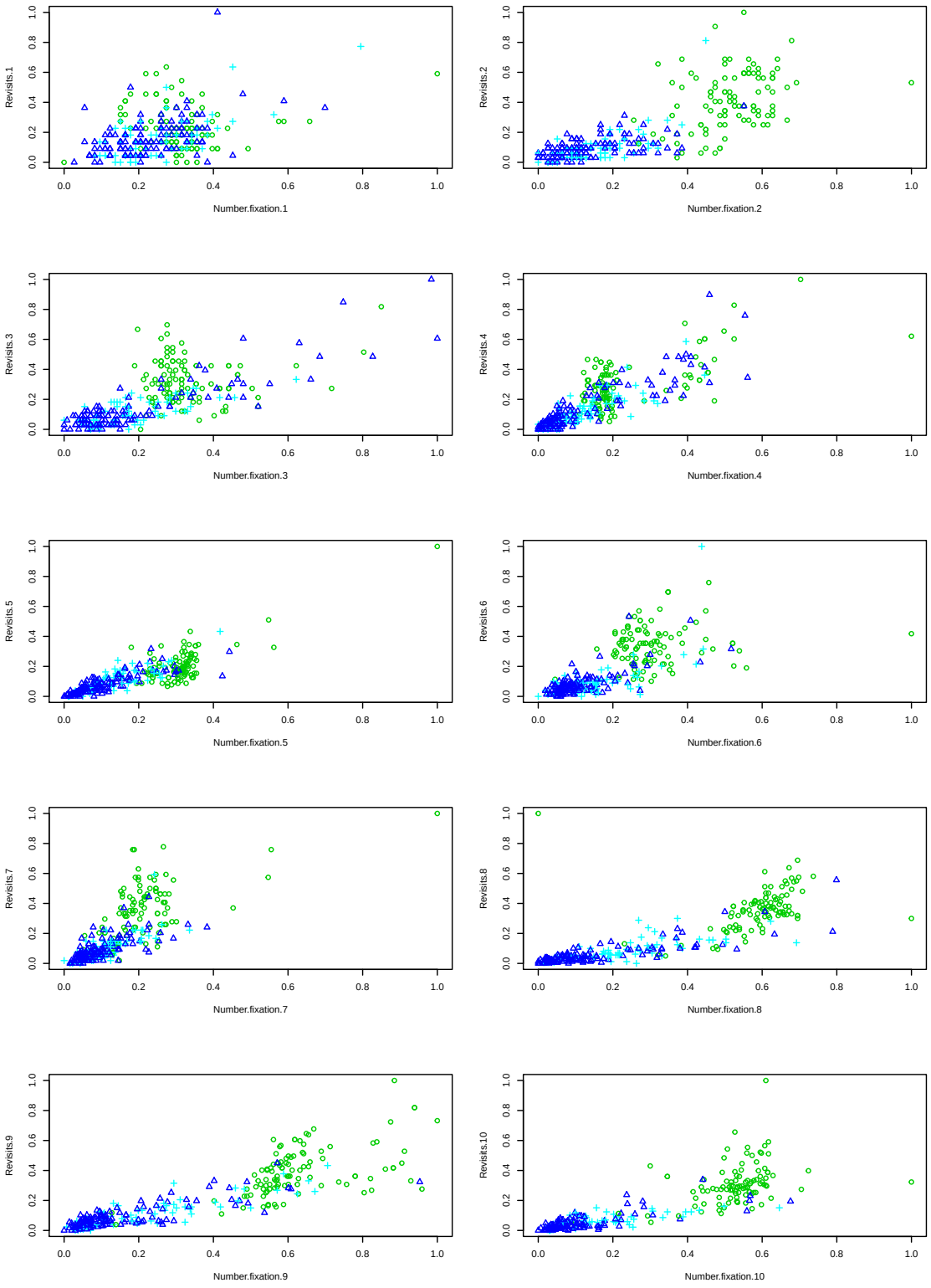

Figure 4.16: Segregations among the three groups based on K-means method. Condition 1 marked as circles; condition 2 marked as crosses, and condition 3 marked as triangles.

These results provide some insight on how the K-means method performs more precisely compared to traditional PSFs. First, the K-means method has much higher power to separate the three clusters with high balanced accuracy (0.994) compared to the PFSs (l_z^*: 0.32, and l_t: 0.28). Moreover, after combining all the subjects in condition 2 and 3, the overall predictive accuracy rate of the K-means method was 0.96. Second, K-means could be used to identify more than two clusters of subjects who had different behavioral characters. Whereas, the PFSs could only be used to separate the aberrant cases from the normally behaved test-takers. This means that the K-means method has an invaluable advantage to be used to identify aberrances when multiple aberrant behaviors are mixed together. Importantly, real world data is even more complex than the experimental data, and requires the unsupervised learning method like K-means to help the practitioners to make fine decisions about who behaved aberrantly or not.

4.7.2 Representative Supervised Learning Methods

In this section, two selected supervised learning methods, K-nearest neighbors (KNN) and random forest (RF), were performed to classify subjects in various experimental conditions in order to see whether it is possible to obtain higher classification accuracy rates compared to the unsupervised learning method, like K-means, as well as further reduce error rates. The specificity, sensitivity, and accuracy rates are of interest here.

KNN. The KNN algorithm was used to predict the class membership of a subject by identifying other cases closest to it that show similar behavioral pattern. Starting the algorithm requires specifying the number of neighborhoods (K) as a tuning parameter.

Figure 4.17 shows that by calculating error rate across different neighborhood values (K) given the dataset, three clusters should be expected since it yields the lowest classification error rate (error rate = 0.157).

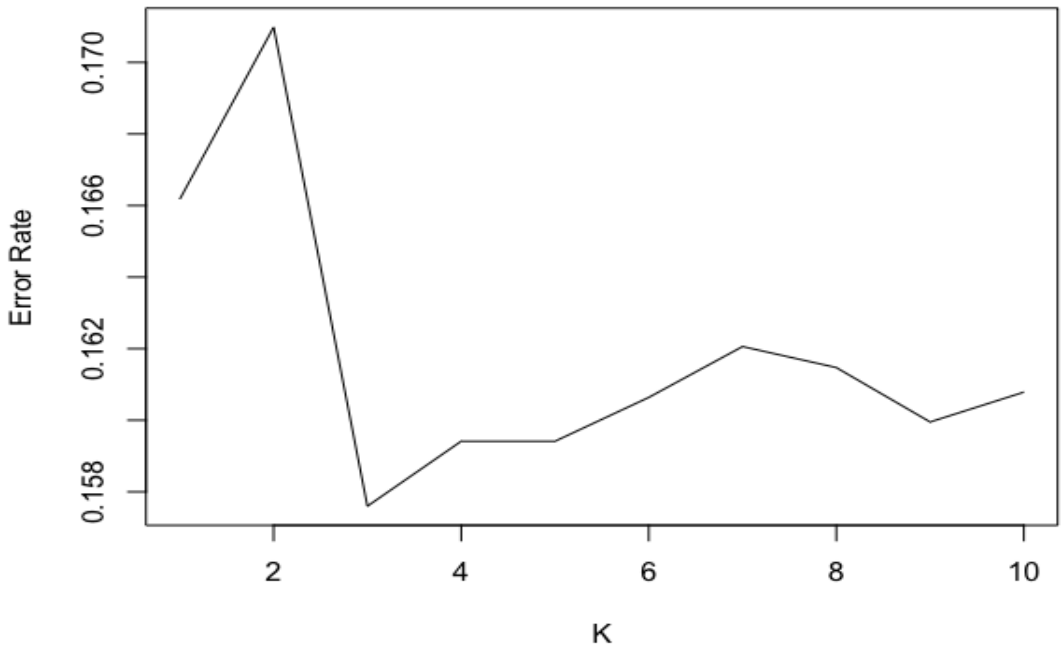

Figure 4.17: The optimal number of neighborhood based on the KNN algorithm.

The performance of the KNN method is summarized in Table 4.11. Subjects who were in condition 1 and 3 were accurately classified with high sensitivity rates of 99.9% and 90.6% respectively. However, the sensitivity rate for condition 2 was relatively lower, about 64%, which indicates it is challenging to use the KNN method to separate the subjects in condition 2 from condition 3. This can be attributed to the fact that subjects in condition 2 and 3 both practiced the items before taking their real tests. As a result, they could possibly behave similarly due to having pre-knowledge of the items, which made it

more difficult for KNN method to differentiate between these two conditions. However, if we combine the condition- two and three as one group, Table 4.12 shows that there are substantial increases in both sensitivity (99.4%) and specificity (93.3%).

Table 4.11: Classification accuracy for KNN methods with three groups

	True label		
	1	2	3
Sensitivity	0.99	0.643	0.906
Specificity	0.933	0.944	0.88
Overall accuracy			0.841

Table 4.12: Sensitivity and specificity for KNN Methods with two groups

% Consistent Decision	K-Means
Sensitivity	0.99
Specificity	0.933
Overall accuracy	0.951

As demonstrated in Figure 4.18, the grouping result of using the KNN method was demonstrated based on two features: 1) the number of fixations, and 2) the review time across ten items. It can be seen from Figure 4.17, for instance, that item 2 shows a clear separation between the subjects who belong to condition 1 (circles) and the others (condition 2 - crosses and condition 3 - triangles).

In general, the results for the KNN method were similar to the results of for K-means, with a high accuracy rate of classifying different responding behaviors. However, one thing to note is that the specificities were relatively higher with the KNN method as compared to the K-means method. This means that the supervised learning methods have promising power to differentiate various types of responding behaviors, while at the same time protecting normally behaved ones from being misidentified as aberrant cases

in practice.

Random Forest. To make results easy to explain to practitioners, another supervised learning method was used to classify different types of responding behaviors. Using the RF method, decision trees can be displayed graphically. This graphical display easily shows the critical features for yielding the final clusters.

To generate valid results based on the RF method, two parameters need to be tuned or defined. One is the number of trees in the forest, and another is the number of variables (mtry) that need to be randomly considered at each splitting node. Usually, the default setting for mtry is calculated as the square root of the total number of features. Therefore, in this study, mtry is equal to 7 (rounded down). The number of trees is tuned by computing the classification error, which is demonstrated in Figure 4.19. It can be shown from Figure 4.19 that having 304 trees yields the lowest classification error.

The performance of the RF method is summarized in Table 4.13. It can be seen that subjects who are in conditions 1 and 3 were accurately classified by the RF algorithm at high sensitivity rates, 99% and 87% respectively. The sensitivity rate for condition 2 is 87%, which is about 23% higher than the sensitivity rate calculated based on the KNN method. Also, after combining conditions 2 and 3 as one group, Table 4.14 shows that both sensitivity and specificity are high, which indicates that the RF method successfully identified the subjects who had pre-knowledge of test items. The RF method yields the highest overall accuracy rate compared to other methods, approximately 98.4%.

A classification tree built based on the training dataset is plotted in Figure 4.20. As shown in Figure 4.20, the tree splits from the top node, which is the number of fixations for item 10. Under that node, we can see two options: either yes marked as Y

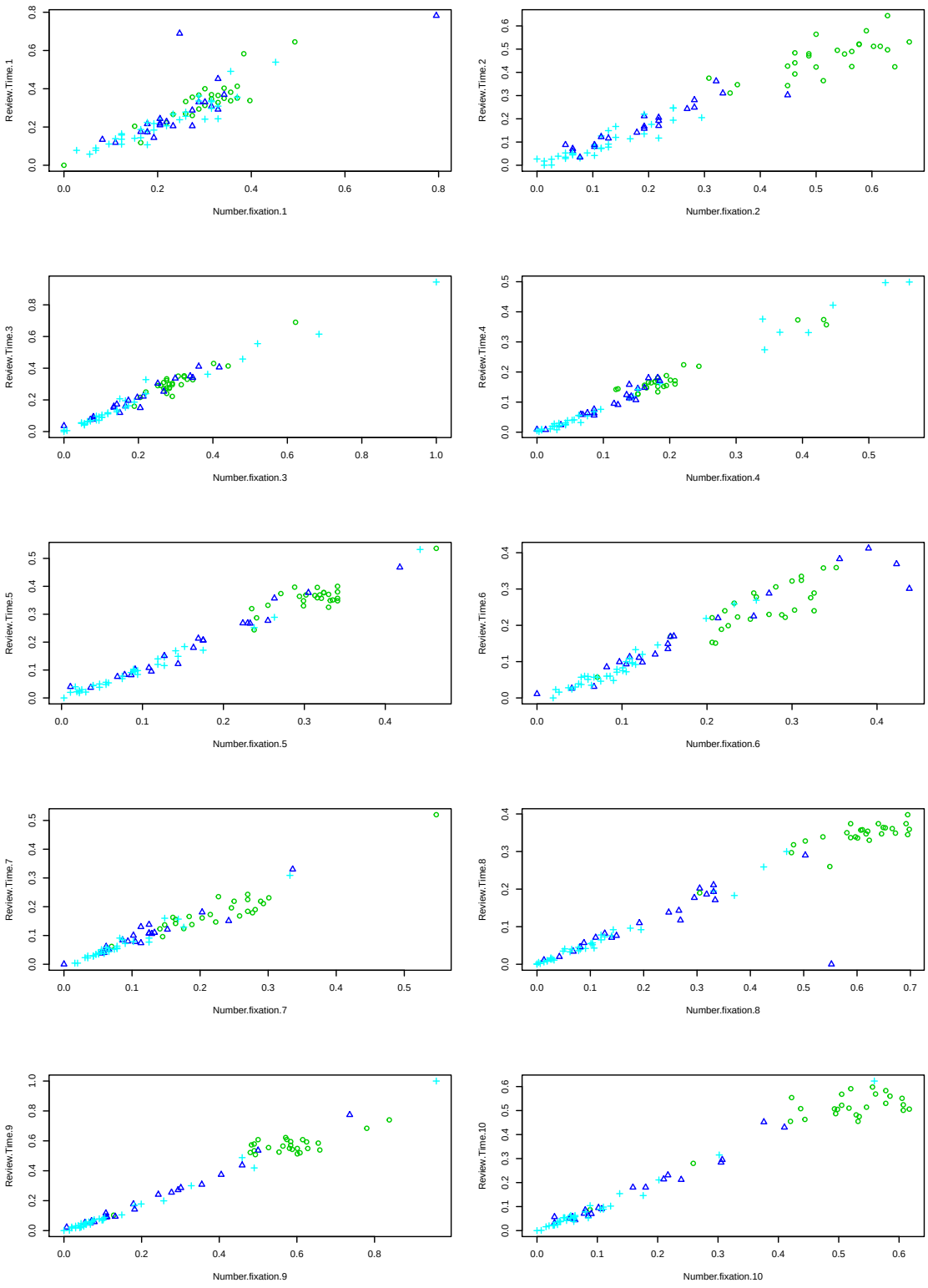

Figure 4.18: Segregations among the three groups based on KNN method. Condition 1 marked as circles; condition 2 marked as crosses, and condition 3 marked as triangles.

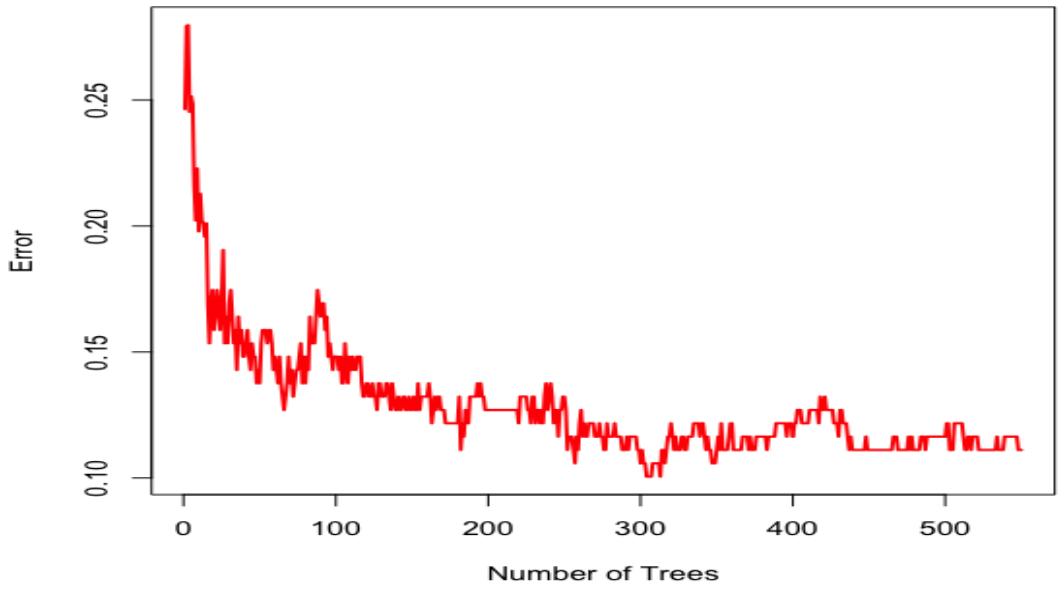

Figure 4.19: Number of trees in random forest.

Table 4.13: Classification accuracy for RF Methods with three groups

	True label		
	1	2	3
Sensitivity	0.99	0.862	0.866
Specificity	0.977	0.936	0.943
Overall accuracy			0.905

or no labeled as N. If the cases satisfied the condition listed at the node, then they were assigned to the left side of the node, and if not, they were assigned to the right side. The splitting process continues until some stopping condition was met as mentioned in Chapter 3. At the bottom level, the final predicted cluster membership was assigned to each case. The membership was assigned a value of 0, 0.5 and 1, indicating condition 1, 2 or 3 respectively. By following all of the conditions (nodes) from the top to the bottom, practitioners could gain insights about the behavioral characteristics of a group who behaved

Table 4.14: Sensitivity and specificity for RF methods with two groups

% Consistent Decision	K-Means
Sensitivity	0.99
Specificity	0.977
Overall accuracy	0.984

aberrantly in their tests. For example, as shown in Figure 4.20, when a test-taker put low visual effort into answering item 10 (Number.fixation.10 $<$ 0.34), performing carelessly over the entire test, but answered item 8 correctly, this person would most likely be classified as an aberrant test-taker who had a lot of pre-knowledge about the items. This is significant as one seeks for a method to accurately flag the aberrantly behaved test-takers with interpretable graphs.

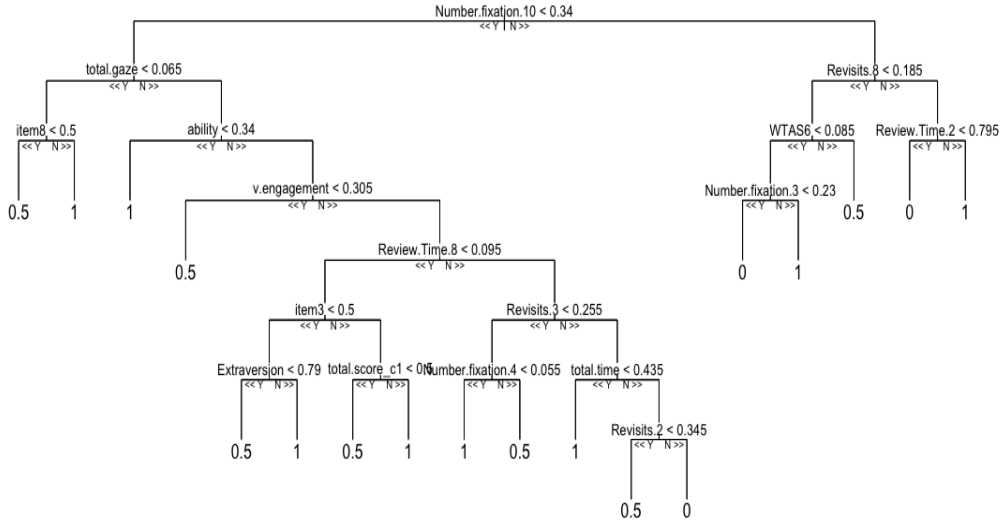

Figure 4.20: Classification Tree as a demonstration of classifying different types of responding behaviors. A value of 0, 0.5 and 1, indicating condition 1, 2 or 3 respectively.

In summary, a comparison of the capacities on clustering different types of responding behaviors across different statistical methods was presented. The results suggest that the overall sensitivity, specificity, and overall accuracy rates based on the supervised learning methods were relatively higher than the ones based on the traditional IRT- and RT-based methods as well as the unsupervised learning methods. In addition, it is challenging to use IRT- and RT-based methods to accommodate complex datasets with large numbers of aberrances. In contrast, data mining methods are able to overcome that limitation and accurately classify different types of test-taking behaviors. Particularly, the specificity base of the RF method was much higher than other methods, which implies that the RF method could potentially protect normally behaved test takers from incorrect classified as wrongdoers.

Chapter 5: Discussion

In this study, many methods were created, developed, and investigated that incorporate bio-information technology, namely eye tracking, in the classification of different types of responding behaviors in computer-based testing scenarios. This study explores the potential to combine psychometric and biometric information to assess test-takers behaviors.

First, the collected experimental data was visualized and summarized. Next, three innovative gaze fixation based models were proposed. The first model, named the NBF model, was defined by assuming constant engagement levels across all the items. A slope term and a quadratic term were added to the first model as two extensions to the base model. The NBF-LT model and the NBF-QR model used a parsimonious parameterization of the mean structure to capture changes in engagement exhibiting either linear or nonlinear trends. To properly identify the scale of the latent variables, the expectations of the person-side latent variables Ω were fixed as 0s, which were aligned with the previous research from Fox and Marianti (2016). The proposed model helped to understand individualized differences in test engagement levels and show item characters, including how much visual effort was required for answering an item and its discriminating power.

Second, A three-way a hierarchical joint model was proposed to jointly model item responses, RTs, and visual fixation counts. A 1-PL IRT model was used to model the

item responses, a log-normal response time model was utilized for modeling RTs without truncation, and a negative binomial visual fixation counts model was selected to model the gaze fixations. These three measurement models were jointly modeled at the lower level of the hierarchy. At the higher level of the hierarchical model, the mean vectors and variance-covariance structures were estimated for both person and item parameters, respectively. This modeling approach permits the evaluation of trade-offs among responding accuracy, working speed, and visual engagement as reflected by the person-domain model.

Third, a three-way joint model was fit to the data across different experimental conditions. Thus, behavioral pattern differences across different experimental conditions were uncovered along with the gaps of item parameters estimates. The results show that pre-knowledge had large effects on the item characteristics. In addition, the associations among person-side behavioral constructs (e.g., latent ability, working speed, and visual engagement) were greatly affected by pre-knowledge. With pre-knowledge on test-items, the ability estimates of test-takers were positively correlated with their working speed. In contrast, ability estimates were negatively associated with test-takers visual engagement levels, and their working speeds were negatively correlated with their visual engagement levels. One thing to note is that when the test-takers had no access to test-preparation materials, their ability estimates were not correlated with their working speed nor with their visual engagement levels. This is of interest because testing companies could potentially tackle identification of suspicious aberrant test-takers by matching their behavioral characters against these findings.

Lastly, representative data mining methods were utilized to classify different types

of test-takers using multimodal data including estimates from the joint modeling previously mentioned. These newer methods are well suited to tackle identification of various test-taking behaviors as they can incorporate vast amounts of data coming from numerous rich sources (e.g., process data, biometrical data, and psychometric data). This point is particularly salient as many testing administrations are moving away from pencil-and-paper assessments toward computer-based environments. The findings from this study showed that the data mining methods investigated here gave relatively high detection rates (sensitivity) compared with traditional methodsespecially the supervised methods. Traditional methods were able to flag aberrant test takers who have pre-knowledge of the test items without incorrectly classifying normally-behaved test-takers as aberrances.

The current study successfully, as an attempt, integrated biometric and psychometric information with various machine learning methods to classify different types of responding behaviors. To better understand the individual differences in their test engagement during problem-solving, three models were proposed to manifest some unique insights about test takers problem-solving patterns. Then, the proposed joint-modeling approach marries various measurement models from the field of psychometrics with data mining methods from the field of machine learning. Additionally, this method could be used for any scenario in which researcher believes that biometric and psychometric information may be essential in the classifications of various behaviors, not limited to educational testing. Yet, there were limitations in the application of this method in the current study, which are discussed in more detail below.

5.0.1 Limitations for the Current Work

All the proposed models were estimated within a Bayesian framework using an MCMC algorithm. This estimation procedure appeared to be effective at recovering item and person parameters. However, a more comprehensive simulation study would need to be conducted to investigate conditions found in practice.

Also, as with the majority of studies, the findings of this study have to be seen in light of some limitations. Conclusions about the manifested test-takers' behavioral patterns across different experimental conditions was merely based on the current research design, which is subject to the several limiting factors such as sample size and number of test items. Therefore, practitioners or applied researchers should be cautious as to what findings could be properly generalized to industry practice.

An additional limitation of applying the results of this study is that the proposed models require collecting gaze data via eye tracking devices, which are expensive and difficult to use. However, the proposed models could be widely applied to analyze multimodal data, including eye-tracking data, to evaluate students task performance in a technology-enhanced simulation based testing system.

To take advantage of supervised methods for classifying different types of responding behaviors, a true class of membership labels need to be created in the first place. For practical use, test companies need to know the true labels of different types of responding behaviors based on a serious investigation in order to build a blacklist to train the models. To overcome this shortage, traditional and unsupervised learning methods could be used to build preliminary labels. In the end, all the results could be taken into account to make

final decisions on test behaviors.

5.0.2 Recommendations for Future Directions

Several model extensions could be further considered beyond those presented here. An interesting next elaboration might be to extend the three-way joint model by incorporating finite mixtures. This is vital for educational and psychological testing, where test takers show different problem-solving behaviors such as carelessness and copy cheating, or implement unintentionally distinguished strategies in different learning groups (e.g, male / female, low- / high-achievement students, etc). The latent groups that would be uncovered from such an analysis could promote a better clustering of different types of test takers with different behavioral patterns. A second elaboration is to develop response models for polytomous or graded responses. This is important for psychological testing where items are Likert-scaled. Another methodological extension would be to carry out a sensitivity analysis to measure the impact of various prior distributions on parameter estimation for the proposed models.

In addition, data mining methods are able to work with complex datasets with large sample sizes and high dimensional features, but also can be surprisingly useful in the case where sample size is small compared to the number of features. In this latter case, traditional parametric methods may be useless or may yield unstable parameter estimates due to the limited sample size. With the development of TELS, data mining methods can play an important role to analyze such data with high efficiency, as shown in this study.

To classify various test-taking behaviors with high accuracy, numerous sources of information regarding test-takers, including information originating from biometric tech-

nologies, need to be aggregated utilizing profoundly proficient computational methods like cloud computing. Also, other types of biometric information could be integrated into the current modeling framework such as Electroencephalogram (EEG), facial expression recognition, and body and body part movement. While writing this thesis, the COVID-19 virus is spreading rapidly over the world. As a consequence, many schools decided to switch to online instruction and learning, which results in an increscent need of having an online protocol platform to security the online delivered exams. Thus, it is essential to continuously develop new measures and tools integrating all kinds of psychometric and biometric information to secure the exams rendered in differ forms.

Mislevy (2016) showed that new types of educational assessments would include psychometric models, biometrics, machine learning, and data mining methods utilizing a high-effectiveness computation system. Through methodological examinations and investigations utilizing empirical data, the signal-to-noise ratio (SNR) could be improved. This implies that grouping accuracy could be improved, gaining sensitivity to nuanced behaviorial patterns.

Obviously, reality is significantly more unpredicatable than any model could sufficiently catch. A high-efficiency modeling structure like that used by many machine-learning methods could be utilized to uncover unpredicted, concealed patterns in a large amounts of informationletting the information represent itself. This might be the best route for classifying different test-taker behaviors, which can be very challenging to recognize from typical test-taking behaviors. However, test companies also need to be alert to the possibilities of overextending the statistical results to judge a test-taker's suspicious testing behavior without a follow-up panoramic investigation. Put the matter another way,

test companies need to seek specific guidance from experts on how to prevent potential epical false-positives.

Appendix A: List of Variable Names

Table A.1: Variable names

Variable (abbreviation)	Full Terminology in the Data and Description
No.fixation	The number of visual fixation counts.
Res.Time	Response time: time used by a test-taker to answer an item.
P-value	P-value: a measure of item difficulty, which the proportion of test-takers who answered an item correctly. A high P-value indicates high easiness.
No.Revisits	The number of revisits: the frequency of saccades move back to the previously viewed area of interest (AOI).
Total.Score	Total scores: the number of items answered correctly.
Total.Time	Total time: the total time spent by a test-taker to finish the test.
Total.gaze	Total time: the total number of visual fixation counts generated by a test-taker while performing the test.
V.Engmt	Latent visual engagement: an individualized parameter showing the visual engagement level for each test-taker performing a test, which is estimated from the negative binomial visual engagement model.
Ability	Latent ability: an individualized ability parameter indicating cognitive ability of decoding questions, which is estimated from the one-parameter logistic model.
Speed	Latent speediness: a parameter representing how fast a test-taker work on his/her test, which is estimated based on the lognormal response time model.
WTAS	WTAS: Westside test anxiety scale.
Extraversion	Extraversion: one of the big five personality traits. The scores are calculated by averaging item 1 and item 6 in the ten-item personality inventory - (TIPI).
Agreeableness	Agreeableness: one of the big five personality traits. The scores are calculated by averaging item 2 and item 7 in the ten-item personality inventory - (TIPI).
Conscientious	Conscientious: one of the big five personality traits. The scores are calculated by averaging item 3 and item 8 in the ten-item personality inventory - (TIPI).
Emotional	Emotional: one of the big five personality traits. The scores are calculated by averaging item 4 and item 9 in the ten-item personality inventory - (TIPI).
Openness	Openness: one of the big five personality traits. The scores are calculated by averaging item 5 and item 10 in the ten-item personality inventory - (TIPI).

Appendix B: Summary Statistics of All the Variables

Table B.1: Summary Statistics of All the Variables

	Experimental Conditions								
	Condition 1(N=93)			Condition 2(N=98)			Condition 3(N=107)		
	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med
No.fixation.1	24.6	5.4	25	22.3	10	20	20.2	9	18
No.fixation.2	47.4	6.7	47	21.9	10.4	20.5	17	9.3	15
No.fixation.3	45.1	8.9	43	27.3	16.1	24	32.3	28.8	18
No.fixation.4	72.4	31.3	62	48.9	42.5	37.5	51.9	47.5	28
No.fixation.5	117.7	15	122	61.1	38.4	51	49.9	45	39
No.fixation.6	83.4	22.1	80	43.3	29	35	36.8	31.6	28
No.fixation.7	55.8	14.4	56	35.3	29.7	28	32.3	22.7	26
No.fixation.8	195.9	18.8	197	84.5	56.1	75	55.8	68.9	35
No.fixation.9	193	34.6	182	83.4	66.2	64.5	48.6	45.8	33
No.fixation.10	226.4	25.9	228	75.7	71.2	51.5	53.5	50.5	35
Res.Time.1	7.3	2	7.9	6.8	3.7	5.8	5.8	2.7	5.2
Res.Time.2	14.3	2.2	14.3	6.9	3.7	5.7	5	3.2	4.1
Res.Time.3	13.7	3.1	12.9	8.5	5.4	7.1	9.3	8.6	5.2
Res.Time.4	22.1	9.8	18.6	15.1	14	10.6	15.1	14.4	7.9
Res.Time.5	36	4.6	37	18.8	12.4	14.6	14.7	13.1	10.9
Res.Time.6	23	7.1	21.3	12.7	8.7	10.2	10.5	9.4	7.6
Res.Time.7	15.1	4.2	14.6	10.9	9.9	8.5	9.8	7.6	7.4
Res.Time.8	54	4.2	55	23.6	16.7	20.1	14.9	18.9	8.9
Res.Time.9	53.7	9.4	51	22.9	19.4	17	12.6	12.7	7.8
Res.Time.10	63.4	6.8	64.3	20.8	20.6	12	14.2	14.8	8.8
P-value.1	0.7	0.5	1	0.6	0.5	1	1	0.1	1
P-value.2	0.6	0.5	1	0.8	0.4	1	1	0.2	1
P-value.3	0.6	0.5	1	0.6	0.5	1	0.8	0.4	1
P-value.4	0.4	0.5	0	0.5	0.5	0	0.8	0.4	1
P-value.5	0.3	0.5	0	0.4	0.5	0	0.8	0.4	1
P-value.6	0.5	0.5	1	0.6	0.5	1	0.8	0.4	1
P-value.7	0.4	0.5	0	0.3	0.5	0	0.7	0.5	1
P-value.8	0.3	0.5	0	0.4	0.5	0	0.9	0.3	1
P-value.9	0.4	0.5	0	0.3	0.5	0	0.5	0.5	1
P-value.10	0.5	0.5	1	0.6	0.5	1	0.6	0.5	1
No.Revisits.1	5.2	3.1	5	3.7	3.4	3	3.5	2.4	3
No.Revisits.2	13.9	6	14	3.6	3.4	3	2.9	2.2	3
No.Revisits.3	10.9	4.3	10	3.5	2.9	3	5.2	6.1	3
No.Revisits.4	16.6	7.9	16	8.1	8.1	5.5	9.7	10.8	5
No.Revisits.5	20.3	8	19	11.2	8.4	9	9.1	12.2	6
No.Revisits.6	28.5	10.3	28	7.6	9.5	6	6.4	5.4	5
No.Revisits.7	19	8.8	19	6.2	7.4	4	5.2	5.2	4
No.Revisits.8	60.2	17	58	14.1	13.2	10	9.8	17.7	5
No.Revisits.9	52	19.5	51	16.8	13.4	14.5	10.9	11.9	8
No.Revisits.10	60.8	24	55	11.8	11.1	8.5	9.2	10.7	6
Total.Score	10.2	3.3	10	9.9	3.4	10	15.9	2.6	16
Total.Time	608	81.7	599.5	276.6	166.1	250	197.6	143.8	147.3
Total.gaze	2117.5	229.1	2142	949.4	514.4	841.5	697.2	430.2	560
V.Engmt	0	0.1	0	0	0.4	0	0	0.5	0.1
Ability	0	0.5	0	0	0.7	0	0	1.1	0
Speed	0	0.1	0	0	0.5	0.1	0	0.5	0.1
Anxiety	3.4	1	4	3	1.1	3	2.8	1.1	3
WTAS1	2.7	1.2	3	2.7	1.1	3	2.7	1.1	3
WTAS2	3.4	1.2	4	3.4	1.2	3	3.5	1.2	4
WTAS3	2.9	1.1	3	2.8	1.2	3	2.9	1.1	3
WTAS4	2.6	1.2	2	2.5	1.1	2	2.7	1	3
WTAS5	2.8	1.1	3	2.8	1	3	2.8	1	3
WTAS6	2.3	1.2	2	2.3	1.1	2	2.3	1.1	2
WTAS7	2.4	1.3	2	2.5	1.2	2	2.5	1.1	2
WTAS8	2.9	1.4	3	3	1.1	3	3.2	1.2	3
WTAS9	3.9	1.2	4	3.8	1.2	4	3.9	1.1	4
WTAS10	2.6	1.2	3	2.8	1.2	3	2.4	1.2	2
Extraversion	4.2	1.5	4	4.2	1.5	4	4.5	1.5	4.5
Agreeableness	4.7	1.3	5	5	1.2	5	4.9	1.1	5
Conscientious	5.1	1.3	5.5	5.4	1.2	5.5	5.4	1.2	5.5
Emotional	3.6	1	3.5	3.7	1	3.5	3.9	0.9	4
Openness	5	1.3	5	5.5	1	5.5	5.2	1.1	5.5

Note: SD represents standard deviation; Med represents median.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*, 163-168.

Ackerman, P. L., Kanfer, R., Shapiro, S. W., Newton, S., & Beier, M. E. (2010). Cognitive fatigue during testing: An examination of trait, time-on-task, and strategy influences. *Human Performance*, *23*, 381-402.

Anderson, D., & Burnham, K. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, *66*, 912-918.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*(345), 44-49.

Barbato, G., della Monica, C., Costanzo, A., & De Padova, V. (2012). Dopamine activation in neuroticism as measured by spontaneous eye blink rate. *Physiology Behavior*, *5*(2), 332-336.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*(2), 105-139.

Bay, L. (1995). Detection of cheating on multiple-choice examinations. In *Annual Meeting of the American Educational Research Association.* San Francisco, CA.

Belleza, F. S., & Belleza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, *16*(3), 151-155.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via kullback-leibler divergence and k-index. *Applied Psychological Measurement*, *34*(6), 379-392.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(3), 199-208.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan & T. M. Nicholas C. (Eds.), *Grouping multidimensional data* (p. 25-71). Berlin, Germany: Springer.

Bishop, S., & Egan, K. (2016). Detecting erasures and unusual gain scores. *Handbook of Quantitative Methods for Detecting Cheating on Tests*, 193.

Bishop, S., Liassou, D., Bulut, O., & Seo, D. G. (2011). Modeling erasure behavior. In *Annual meeting of the National Council on Measurement in Education.* New Orleans, LA.

Blanchard, H. E., & IranNejad, A. (1987). Comprehension processes and eye movement patterns in the reading of surpriseending stories. *Discourse Processes*, *10*, 127-138.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*, 1126-1148.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, *27*, 395-414.

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory Cognition*, *41*, 82-97.

Born, S., & Kerzel, D. (1999). Computer interface evaluation using eye movements:

methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Born, S., & Kerzel, D. (2008a). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-382.

Born, S., & Kerzel, D. (2008b). Influence of target and distractor contrast on the remote distractor effect. *Vision Research*, *48*(28), 2805-2816.

Boulesteix, A. L., Strobl, C., Augustin, T., & Daumer, M. (2008). Bagging predictors. *Evaluating Microarray-Based Classifiers: An Overview*, *6*, 77-97.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, *26*(1), 801-849.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Bullinaria, J. A. (2004). Introduction to neural networks. In M. N. Jos & M. F. Santos (Eds.), *School of computer science* (p. 512-523). Birmingham, UK: Springer.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in america's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice*, *7*(2), 5-9.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200-210.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Routledge.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. New Yor, NY: Routledge.

Cochran, W. G. (1952). The $\chi 2$ test of goodness of fit. *The Annals of Mathematical Statistics*, 315–345.

Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Academic Medicine*, *60*(2), 136-7.

Colzato, L. S., Slagter, H. A., van den Wildenberg, W. P., & Hommel, B. (2009). Closing ones eyes to reality: evidence for a dopaminergic basis of psychoticism from spontaneous eye blink rates. *Personality and Individual Differences*, *46*(3), 377-380.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman Hall.

Coff, C., & O'regan, J. K. (1987). Reducing the influence of non-target stimuli on saccade accuracy: Predictability and latency effects. *Vision Research*, *27*, 227-240.

Crawford, C. C. (1930). Dishonesty in objective tests. *The School Review*, *38*(10), 776-781.

Dai, Y. (2013). A mixture rasch model with a covariate: A simulation study via bayesian markov chain monte carlo estimation. *Applied Psychological Measurement*, *37*(5), 375396.

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*, 225-237.

Dickenson, H. F. (1945). Identical errors and deception. *The Journal of Educational Research*, *38*(7), 534-542.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In B. Dash S. Subudhi (Ed.), *International workshop on multiple classifier systems* (p. 1-15). Berlin, Germany: Springer.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Pattern Recognition*, *19*(2-3), 103-130.

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*(1), 67-86.

Drew, G. C. (1951). Variations in reflex blink-rate during visual-motor tasks. *Quarterly Journal of Experimental Psychology*, *3*(2), 73-88.

Dubes, R. C., & Jain, A. K. (1988). *Algorithms for clustering data*. Upper Saddle River, NJ: Prentice-Hall.

Daz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 1471-2105.

Everitt, B. S. (1981). A monte carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, *16*, 171-180.

Fairclough, S. H., & Venables, L. (2006). Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology*, *71*(1), 100-110.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525-543.

Fix, E., & Hodges, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. *International Statistical Review*, *3*, 238-247.

Fossey, W. A. (2017). *An evaluation of clustering algorithms for modeling game-based assessment work processes* (PhD thesis).

Fox, J. P., Entink, R. K., & Avetisyan, M. (2014). Compensatory and noncompensatory multidimensional randomized item response models. *British Journal of Mathematical and Statistical Psychology*, *67*, 133-152.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*, 540-553.

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

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

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *2*(4), 235-256.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Cambridge, MA: Academic Press.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*, 906-914.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York: Chapman & Hall.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Journal of Educational and Behavioral Statistics*, *6*, 733-760.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139-150.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. American College Testing Program Iowa City.

Hartigan, J. A., & Wong, M. A. (2012). A k-means clustering algorithm. *Applied Statistics*, *28*, 100-108.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York. NY: Springer.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In *Quantitative psychology research* (p. 173-190). Springer, Cham.

Hendrawan, I., Glas, C. A., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement*, *29*(1), 26-44.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*(4), 46-55.

Hess, E. H., & Polt, J. M. (1964). upil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190-1192.

Ho, A. D. (2008). The problem with proficiency: Limitations of statistics and policy under no child left behind. *Educational Researcher*, *37*(6), 351-360.

Holland, P. W. (1996a). Assessing unusual agreement between the incorrect answers of two examinees using the k-index: Statistical theory and empirical support. *ETS Research Report Series*, *1*, 141.

Holland, P. W. (1996b). Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support. *ETS Research Report Series*, *1*, 1-41.

Holmqvist, K., Nystrm, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures.* OUP Oxford.

Huang, Y. S., & Suen, C. Y. (1995). A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*, 90-94.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, *37*, 543-558.

Huber, P. J. (1991). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, *1*, 799-821.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, *3*, 605-610.

Inhoff, A. W., Greenberg, S. N., Solomon, M., & Wang, C. A. (2009). Word integration and regression programming during reading: A test of the ez reader 10 model. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1571-1584.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843-877.

Jacob, B. A., & Levitt, S. D. (2004). To catch a cheat: the pressures of accountability may encourage school personnel to doctor the results from high-stakes tests. here's

how to stop them. *Education Next*, *1*(4), 68.

James, G., Witten, D., Hastie, T., & Tibshirani. (2013). *An introduction to statistical learning*. New York, NY: Springer.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Exprimentale*, *47*(2), 310-333.

Justice, M., & Lankford, C. (2002). Pilot findings. *Communication Disorders Quarterly*, *1*, 11-21.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Kennard, D. W., & Glaser, G. H. (1964). An analysis of eyelid movements. *Journal of Nervous and Mental Disease*, *139*(1), 31-48.

Kerr, D., & Chung, G. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining*, *4*, 144-182.

Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, *25*(11), 1293-1302.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59-69.

Kuhn, M. (2017). caret: Classification and regression training. R package version 6.0-71. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=som

Kuncheva, L. I., Bezdek, J. C., & Duin, R. P. (2001). Decision templates for multiple

classifier fusion: an experimental comparison. *Pattern Recognition*, *34*(2), 299-314.

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42-73.

Lattin, J. M., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Thomson Brooks/Cole.

Lee, S. P., Badler, J. B., & Badler, N. I. (2001). Eyes alive. *ACM Transactions on Graphics (TOG)*, *21*(3), 637-644.

Levy, M. R. J. . S. S., R. (2009). Posterior predictive model checking for multidimensionality in item response theory. applied psychological measurement. *Applied Psychological Measurement*, *33*, 519-537.

Li, C. S. (2011). Cluster center initialization method for k-means algorithm over data sets with two clusters. *Procedia Engineering*, *24*, 214.

Linacre, J., & Wright, B. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507.

Lord, F. M. (1952). *A theory of test scores*. Richmond, VA: Psychometric Corporation.

Lowenstein, L. I., O. (1962). Models for speed and time-limit tests. In H. Dawson (Ed.), *The eye* (p. 187-208). Academic Press: New York.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization (no. 8)*. New York: Oxford University Press.

Magis, D., Rache, G., & Bland, S. (2012). A didactic presentation of snijderss lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57-81.

Man, K., Harring, J., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, *56*, 251-279.

Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, *79*(4), 617635. Retrieved from https://doi.org/10.1177/0013164418824148 doi: 10.1177/0013164418824148

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

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445-469.

McLachlan, G. J., Peel, D., & Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics Data Analysis*, *41*, 379-388.

Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311-314.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extention of the schmitt, cortina, and whitney study. *Applied Psychological Measurement*, *21*, 99-113.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item response patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261-272.

Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Applied Psychological Measurement*, *52*, 1-27.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*, 169-186.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. *Technology and Testing: Improving Educational and Psychological Measurement*, 23-48.

Molenaar, D., Bolsinova, M., Rozsa, S., & De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the hungarian wisc-iv block design test. *Journal of Intelligence*, *4*, 10-25.

Molenaar, W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75-106.

Morad, Y., Lemberg, H., & Dagan, Y. (2000). Pupillography as an objective indicator of fatigue. *Current Eye Research*, *21*, 535-542.

Motter, B. C., & Belky, E. J. (1998). The zone of focal attention during active visual search. *Vision Research*, *38*(7), 1007-1022.

Mouselimis, L. (2018). Kernelknn: Kernel k nearest neighbors. R package version 1.0.8. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=KernelKnn

Mroch, A. A., Lu, Y., Huang, C. Y., & Harris, D. J. (2014). Patterns of examinee erasure behavior for a large-scale assessment. test fraud: Statistical detection and methodology. *Test fraud: Statistical Detection and Methodology*, *1*, 137-148.

Mueller, L., Zhang, Y., & Ferrara, S. (2016). What have we learned? In *Handbook of quantitative methods for detecting cheating on tests* (p. 373-390). Routledge, New York: NY.

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.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366. Retrieved from https://doi.org/10.3102/10769986024004342 doi: 10.3102/10769986024004342

Plummer, M. (2015). *Jags: Just another gibbs sampler.* Retrieved from http://mcmc-jags.sourceforge.net/ (version 4.0.0)

Ponsoda, V., Scott, D., & Findlay, J. M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica*, *88*(2), 167-185.

Poole, A., Ball, L. J., & Phillips, P. (2004). In search of salience: A response-time and eye-movement analysis of bookmark recognition. In S. Fincher, P. Markopoulos, D. Moore, & R. Ruddle (Eds.), *People and computers xviii design for life.* London:Springer.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, *60*(2), 211-229.

Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011). Erasure descriptive statistics and covariates. In *Annual Meeting of the National Council on*

Measurement in Education. New Orleans, LA.

Qian, Y., Yao, F., & Jia, S. (2009). Band selection for hyperspectral imagery using affinity propagation. *IET Computer Vision*, *3*, 213-222.

Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, *20*(1), 9-16.

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen Lydiche.

Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (p. 751-766). New York, NY, US: Oxford University Press.

Rayner, K., Murphy, L. A., Henderson, J. M., & Pollatsek, A. (1989). Selective attentional dyslexia. *Cognitive Neuropsychology*, *6*, 357-378.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127-137.

Ripley, B. (2018). tree: Classification and regression trees. R package version 1.0-37. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tree

Romero, C., Gonzlez, P., Ventura, S., Del Jess, M. J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using moodle data. *Expert Systems with Applications*, *36*, 632-1644.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 187-208).

New York: Springer.

Roy-Charland, A., Saint-Aubin, J., Klein, R. M., & Lawrence, M. (2006). Eye movements as direct tests of the go model for the missing-letter effect. *Perception Psychophysics*, *3*, 324337.

Rubin, D. B. (1996). Comment: On posterior predictive p-values. *Statistica Sinica*, *6*, 787-792.

Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, *20*(3), 475-489.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213-232.

Schoenig, R., Geraets, J., & Mulkey, J. (2016). *The test security framework: Why different tests need different test security requirements.* Retrieved from http://mcmc-jags.sourceforge.net/

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression.* Division of Biostatistics, University of California, San Francisco, CA.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology Section A*, *38*(3), 475-491.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Metho-*

den, *7*(22), 131-145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*(2), 149-157.

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*, 46-68.

Sinharay, S. (2018). Are the nonparametric person-fit statistics more powerful than their parametric counterparts? revisiting the simulations in karabatsos (2003). *Applied Measurement in Education*, *31*, 98-98.

Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. *Educational and Psychological Measurement*, *77*(1), 54-81.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298-321. Retrieved from https://doi.org/10.1177/0146621605285517 doi: 10.1177/0146621605285517

Skorupski, W. P., & Egan, K. (2011). Detecting cheating through the use of hierarchical growth models. In *Annual Meeting of the National Council on Measurement in Education.* New Orleans, LA.

Slakter, M. J. (1968). The effect of guessing strategy on objective test scores. *Journal of Educational Measurement*, *5*, 217-222.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342.

Sotaridona, L. S., & Meijer, R. R. (2002a). Statistical properties of the K-index for

detecting answer copying. *Journal of Educational Measurement*, *39*(2), 115-132.

Sotaridona, L. S., & Meijer, R. R. (2002b). Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, *39*(2), 115-132.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*(1), 53-69.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006a). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, *30*, 412431.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006b). Statistical methods for the detection of answer copying on achievement tests. *Applied Psychological Measurement*, *41*(5), 361-37.

Steinley, D. (1985). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1-34.

Stonehill, R. M. (1988). Norm-referenced test gains may be real: A response to john jacob cannell. *Educational Measurement: Issues and Practice*, *7*(2), 23-24.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323-333.

Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to run JAGS.* (version 0.5)

Tatler, B. W., & Vincent, B. T. (2008). Eyes alive. *Journal of Eye Movement Research*, *2*(2), 37-44.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, *20*, 221-230.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). Perfit: An r package for

person-fit analysis in IRT. *Journal of Statistical Software*, *74*, 1-27.

Therneau, . A. B., M.T. (2018). *Rpart: Recursive partitioning and regression trees.* (version 4.1-13)

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (p. 179-203). New York: Academic Press.

Thomas, S. L. (2016). The use of data mining techniques to detect cheating. In *The sixteenth annual maryland conference: Data analytics and psychometrics: Informing assessment practices.* College Park, MD.

Titterington, D. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Hoboken, NJ: John Wiley Sons.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In *New horizons in testing* (p. 83-108). Elsevier.

van der Linden, W., Scrams, D., & Schnipke, D. L. (1999). Using response-time constraints to control speedness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195-210.

van der Linden, W. J. (2006a). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

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

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

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant

response-time patterns in adaptive testing. *Psychometrika*, *73*, 365384.

van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, *31*(3), 283-304.

van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of childrens behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*, 141-177.

van Gerven, P. W. M., Paas, F. G. W. C., van Merrinboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: effects of worked examples on training efficiency. *Learning and Instruction*, *12*(1), 87-105.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). Cusum-based person-fit statistics for adpative testing. *Journal of Educational and Behavioral Statistics*, *26*, 199-217.

Verhelst, N. D., Verstralen, H. H., & Jansen, M. G. H. (2013). A logistic model for time-limit tests. In *Handbook of modern item response theory* (p. 169-185). Springer, New York, NY.

VinuelaNavarro, V., Erichsen, J. T., Williams, C., & Woodhouse, J. M. (2017). Saccades and fixations in children with delayed reading skills. *Ophthalmic and Physiological Optics*, *37*, 531-541.

Vitu, F. (1991). The existence of a center of gravity effect during reading. *Vision Research*, *31*, 1289-1313.

Volodin, N., & Adams, R. (1995). Identifying and estimating a D-dimensional item response model. In *International Objective Measurement Workshop.* University of

California, Berkeley, California.

Wang, T., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*, 119-134.

Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: the kohonen package. *Journal of Statistical Software*, *21*, 1-19.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207-244.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307-320.

Wollack, J. A., & Cizek, G. J. (2016). Section IIa detecting similarity, answer copying, and aberrance. In *Handbook of quantitative methods for detecting cheating on tests* (p. 39-114). Routledge.

Wollack, J. A., & Fremer, J. J. (2013). *Introduction: The test security threat.* New York: Chapman & Hall.

Wright, B. D., & Stone, M. H. (1979). Best test design.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (1998). Conquest: Generalized item response modeling software [computer software and manual]. *Camberwell, Victoria: Australian Council for Educational Research*.

Yan, J. (2016). som: Self-organizing map. R package version 0.3-5.1. [Computer software

manual]. Retrieved from https://CRAN.R-project.org/package=som

Yoss, R. E., Moyer, N. J., & Hollenhorst, R. W. (1970). Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. *Neurology*, *20*, 545-545.

Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement*, *40*(8), 592-607.

Zopluoglu, C., & Davenport, E. C. (2012). The empirical power and type i error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72*(6), 975-1000.

Zubin, J., & Steinhauer, S. R. (1983). The metamorphosis of schizophrenia: from chronicity to vulnerability. *Psychological Medicine*, *13*, 551-571.