Aberrant Response Detection: Incorporating Cumulative Sum Control Chart and Change-Point Analysis

Siyu Wan
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://doi.org/10.7275/35961231 https://scholarworks.umass.edu/dissertations_2/2915

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Aberrant Response Detection:

Incorporating Cumulative Sum Control Chart and Change-Point Analysis

A Dissertation Presented

by

SIYU WAN

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

DOCTOR OF PHILOSOPHY

September 2023

College of Education

Research, Educational Measurement, and Psychometrics
Aberrant Response Detection:
Incorporating Cumulative Sum Control Chart and Change-Point Analysis

A Dissertation Presented
by
SIYU WAN

Approved as to style and content by:

__________________________
Lisa A. Keller, Chair

__________________________
Scott Monroe, Member

__________________________
Bruce Skaggs, Member

__________________________
Shane Hammond
Associate Dean for Student Success
College of Education
ACKNOWLEDGMENTS

First and foremost, I express my deepest gratitude to my advisor, Prof. Lisa Keller, who not only chaired my dissertation committee but has also been an unwavering source of support throughout my academic journey. Her guidance and invaluable advice have shaped my academic and career development. I am truly fortunate to have her as my mentor. Prof. Keller’s expertise in psychometrics and her practical approach to problem-solving have inspired me to apply my knowledge in real-world contexts. In addition to providing insightful suggestions for my research, she has continuously encouraged and bolstered my confidence. I am confident that I am well-prepared for a successful professional career in psychometrics thanks to her guidance.

I would also like to express my sincere appreciation to my committee members, Prof. Scott Monroe and Prof. Bruce Skaggs. Prof. Monroe’s invaluable suggestions on research design have significantly contributed to my academic growth. His support has been instrumental in shaping the direction of my research. I am grateful to Prof. Skaggs for serving as the outside committee member for my dissertation. I appreciate his valuable time and insightful suggestions, which have encouraged me to think about making research accessible to individuals from diverse backgrounds.

I am incredibly fortunate to have received support and assistance from the REMP program. I extend my heartfelt thanks to Stephen Sireci, Craig Wells, April Zenisky, and Jennifer Randall. Their exceptional teaching and collaboration have enriched my academic experience. Each of them has had a profound impact on my personal and professional growth.
I would also like to express my gratitude to my colleagues at ABIM, particularly my manager, Dr. Jerome Clauser. Their support and understanding have been invaluable as I balanced the demands of completing my dissertation while starting a new position. Their guidance and encouragement have been instrumental.

Finally, I would like to acknowledge the unwavering support of my family and friends. My parents have always believed in me and instilled in me the confidence to give my best. My friends have been there for me during my moments of emotional lows, encouraging me and enabling me to successfully complete my doctoral project. Their love and encouragement have been a constant source of strength.
ABSTRACT

ABERRANT RESPONSE DETECTION: INCORPORATING CUMULATIVE SUM CONTROL CHART AND CHANGE-POINT ANALYSIS

SEPTEMBER 2023

SIYU WAN, B.S., CENTRAL CHINA NORMAL UNIVERSITY
M.A., UNIVERSITY OF GEORGIA
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professor Lisa Keller

The detection of aberrant responses in psychological and educational testing is crucial for maintaining the validity and reliability of assessment scores. This dissertation evaluates the performance of a new detection method, known as the bootstrap method, specifically targeting random guessing behaviors. A simulation design is used to compare the bootstrap method with traditional cumulative sum control chart (CUSUM) and change-point analysis (CPA) methods, considering various test lengths and severity rates.

The findings indicate that the bootstrap method did not exhibit superior performance in detecting aberrant responses compared to the traditional methods. It demonstrated lower detection power and higher type I error rates, particularly when examinees had similar probabilities of providing correct responses regardless of random guessing or thoughtful answering. While the traditional CUSUM and CPA methods remain effective, the bootstrap method requires further refinement to enhance its detection power.

To address these limitations, future research can explore the integration of
additional statistics into the bootstrap method to improve the accuracy of ability estimation, especially when a higher number of aberrant responses are present. This research contributes to the field of educational measurement by highlighting the need for ongoing refinement and optimization of the bootstrap method to effectively identify various forms of aberrant behaviors in psychological and educational testing settings.

The findings underscore the importance of robust detection methods for maintaining the integrity and validity of assessments, emphasizing the significance of ongoing research and development in this area.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
</tr>
<tr>
<td>ABSTRACT</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
</tbody>
</table>

## CHAPTER
### 1. INTRODUCTION

1.1 Background ........................................................................................................ 1  
1.2 Statement of Problem ......................................................................................... 3  
1.3 Purpose of Study .................................................................................................. 4  
1.4 Significance of Study .......................................................................................... 4  

### 2. LITERATURE REVIEW

2.1 Brief Review of Item Response Theory .............................................................. 5  
2.2 Detection of Aberrant Responses ........................................................................ 7  
2.3 Modeling of Aberrant Responses ......................................................................... 10  
2.3.1 Mixture Model ................................................................................................ 11  
2.3.2 Hybrid Model ................................................................................................ 12  
2.3.3 Graduate Change Model ............................................................................... 13  
2.4 Cumulative Sum Control Charts (CUSUM) ......................................................... 14  
2.4.1 One-Sided CUSUM ....................................................................................... 16  
2.4.2 Two-Sided CUSUM ...................................................................................... 18  
2.4.3 Identify Boundary Values of CUSUM .......................................................... 20  
2.5 Change-Point Analysis (CPA) ........................................................................... 22  
2.5.1 Four CPA Statistics ..................................................................................... 22  
2.5.1.1 $L_{max}$ based on the likelihood ratio test ............................................. 23  
2.5.1.2 $W_{max}$ based on the Wald test ............................................................. 24  
2.5.1.3 $S_{max}$ based on the score test ............................................................... 25
2.5.1.4 $R_{\text{max}}$ based on weighted residual ........................................ 25
2.5.2 Identifying Critical Values of PFS in CPA ........................................... 28
  2.5.2.1 Monte Carlo simulation ................................................................. 28
  2.5.2.2 FDR control method ................................................................. 29
2.6 Summary of Literature ........................................................................... 30

3. METHOD ....................................................................................................... 32
  3.1 General Method ......................................................................................... 32
    3.1.1 A New CUSUM Procedure Using the Bootstrap Analysis ............ 32
    3.1.2 Locating the Change Point of the New CUSUM Procedure ........ 36
  3.2 Simulation Design .................................................................................... 37
  3.3 Data Generation ....................................................................................... 38
  3.4 Data Analyses and Evaluation Criteria .................................................. 39

4. RESULTS ........................................................................................................ 42
  4.1 Baseline Condition without Random Guessing .................................... 42
  4.2 Overall Performance .............................................................................. 43
  4.3 Performance across Theta Levels ........................................................... 47

5. DISCUSSION .................................................................................................. 51
  5.1 Summary .................................................................................................. 51
  5.2 Limitations of Current Study ................................................................... 53
  5.3 Future Directions ..................................................................................... 55
  5.4 Conclusion ............................................................................................... 57

APPENDIX

PERFORMANCES OF THREE DETECTION METHODS .................................. 58
AVERAGE CORRECT CLASSIFICATION RATES ACROSS THETA LEVELS .............................................. 61

REFERENCES ................................................................................................. 63
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1. Simulation Conditions of Study 1</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Table 2. Indices for Aberrant Response Detection Methods</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Table 3. Boundary Values of CUSUM and Critical Value of CPA</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Table 4. Average Type I Errors under the Baseline Condition</td>
<td></td>
<td>42</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1. An Example of 3PL IRT ICC</td>
<td>........................................................................................................ 7</td>
<td></td>
</tr>
<tr>
<td>Figure 2. Summary of 36 PFS from Karabatsos (2013)</td>
<td>.............................................................. 10</td>
<td></td>
</tr>
<tr>
<td>Figure 3. An example of the one-sided CUSUM Chart</td>
<td>........................................................................ 18</td>
<td></td>
</tr>
<tr>
<td>Figure 4. The Example of Bootstrap CUSUM Chart Based on the New Approach</td>
<td>.......................................................... 33</td>
<td></td>
</tr>
<tr>
<td>Figure 5. The Bootstrap CUSUM Chart with Original Data and Five Samples</td>
<td>..................................................... 35</td>
<td></td>
</tr>
<tr>
<td>Figure 6. Bootstrap CUSUM Chart with Changes Shown in the Background</td>
<td>......................................................... 37</td>
<td></td>
</tr>
<tr>
<td>Figure 7. Average Type I Error of All Indices</td>
<td>.................................................................................. 44</td>
<td></td>
</tr>
<tr>
<td>Figure 8. Average Detection Power of All Indices</td>
<td>........................................................................... 45</td>
<td></td>
</tr>
<tr>
<td>Figure 9. Average CCR of All Indices</td>
<td>.................................................................................. 47</td>
<td></td>
</tr>
<tr>
<td>Figure 10. Average Power of Detection Methods across Ability Levels</td>
<td>.............................. 48</td>
<td></td>
</tr>
<tr>
<td>Figure 11. Power of Detection Methods across Ability Levels by Test Length</td>
<td>......................... 50</td>
<td></td>
</tr>
<tr>
<td>Figure 12. Power of Detection Methods across Ability Levels by Severity Rates</td>
<td>.......... 50</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Background

When using scores of any assessment, the underlying assumption is that examinees’ responses should not be affected by any unusual behaviors. However, various aberrant responses are commonly observed in practical testing situations. For example, Ward et al. (2017) reported that approximately 10% of examinees provided careless responses during a test. Apart from carelessness, test behaviors such as speededness, lack of motivation, cheating, pre-knowledge of items, the warm-up effect, and fatigue can all lead to aberrant responses (Sinharay, 2017b; Zhang et al., 2020). For example, test speededness typically occurs when examinees do not have sufficient time to thoroughly contemplate each test question (Bejar, 1985), thereby compromising the reliability and validity of the test (Lu & Sireci, 2007). Aberrant responses also undermine the accuracy of the examinees' ability estimation, leading to invalid conclusions and inferences based on assessment scores (Shao, 2016).

Therefore, detecting aberrant responses is essential to maintain test score validity.

The existing literature presents several approaches to detect aberrant responses of examinees. The first approach involves directly modeling aberrant behaviors based on varying assumptions. Apart from traditional unidimensional item response models (IRT), researchers have designed models such as the mixture model (Bolt, Cohen, & Wollack, 2002), the hybrid model (Yamamoto & Everson, 1997), and the graduate change model (Wollack & Cohen, 2004) to detect aberrant responses. By fitting these models to examinees’ responses and comparing model-fit statistics to the
unidimensional IRT model, researchers can identify aberrant responses.

This dissertation, however, concentrates on another approach, termed as the “outlier detection” approach, where responses are modeled using regular IRT models and responses or response patterns deviating from the typical response model are flagged. In the initial development of this approach, researchers employed different person-fit statistics (PFS) to identify test takers demonstrating abnormal item response patterns (which resulted in either artificially high- or low-test scores). Those test takers were then separated from those exhibiting normal item response patterns (Karabatsos, 2003). Nevertheless, traditional PFS has not shown strong performance in detection (de la Torre & Deng, 2008), and the majority of person-fit research in the literature focused on paper-and-pencil examinations (van Krimpen-Stoop & Meijer, 2000).

Recently, two statistical process control (SPC) methods—cumulative sum control chart (CUSUM) and change-point analysis (CPA)—were introduced to detect aberrant responses. CUSUM identifies aberrant responses by cumulatively summing the residuals between the observed and expected scores on each item in the sequence of item administration. When the summed residual surpasses a certain threshold value, the method signals a misfit and aberrant response (Sinha, 2016). Conversely, CPA postulates that the model itself or the model parameters undergo a structural change before and after the change point (Sinha, 2017b). It thereby detects the existence or absence of one or more change points in a sequence of random variables and determines the location of the point.

Both SPC methods, CUSUM and CPA, have been widely used in recent years to detect aberrant response behaviors in a test (CUSUM: Armstrong & Shi, 2009a,
One benefit of the CUSUM method is its allowance for easy visualization, providing a clear, intuitive, and comprehensive depiction of the entire test-taking process. By observing the plot, researchers can visually identify multiple change points. Furthermore, CPA not only identifies examinees with aberrant responses but also automatically and accurately locates change points where an examinee started to alter their behaviors.

1.2 Statement of Problem

While CUSUM and CPA offer notable advantages, they also present certain limitations. For CUSUM, a manual check of the control chart is needed to locate the change point which can be time-consuming and laborious, particularly for large sample sizes. And although CPA identifies the change point automatically, its efficiency may be significantly affected if the change point occurs at the initial or final stage of the test sequence. Additionally, CPA requires more intensive computations than CUSUM.

To address these challenges, Wayne A. Taylor proposed a novel procedure in 2000 that amalgamates the CUSUM control chart and CPA (Taylor, 2000a). Following the plotting of the CUSUM chart, this procedure uses bootstrap analysis to identify change points based on the confidence level. Then, CPA is utilized to automatically locate the change point. Despite its promising nature, the algorithm has not been validated in the field of educational measurement, and no empirical studies have been conducted to test this method.
1.3  **Purpose of Study**

The purpose of this dissertation is to validate the procedure introduced by Taylor (2000a) through a simulation study that systematically evaluates its performance. Specifically, in the simulation study, I aim to assess the performance of various indices fitted to this new procedure. The efficiency of the new procedure will be compared to the traditional CUSUM and CPA methods. This study seeks to answer the following research questions:

1. Which index better detects random guessing behaviors using the new method?
2. Does the new method outperform the traditional CUSUM or CPA methods in detecting rapid guessing responses?

1.4  **Significance of Study**

Identifying aberrant responses is key to maintaining the validity of any tests. The two SPC methods, CUSUM and CPA, have been shown to do the job efficiently. However, they are also limited in dealing with large samples or locating change points that occur at certain places of the test. This new approach developed by Taylor (2000a) brings promise to the field of educational measurement in that it combines the advantages of both CUSUM and CPA, and it has been successfully applied in other fields such as medicine (Xu et al., 2015), astronomy (Chang et al., 2012), and climate (Arif et al., 2017; Bisai et al., 2014). It is unclear, however, whether it can detect aberrant responses correctly and efficiently in educational tests. This study will provide first and valuable insights into the application of this innovative approach.
CHAPTER 2
LITERATURE REVIEW

2.1 Brief Review of Item Response Theory

IRT is a dominant paradigm in the psychometric area for test design, analysis, and scoring. IRT provides a framework in which the relationship between a response of a test item and the latent ability that the test intends to measure is explained by mathematical models. More specifically, an item response is modeled as a probability of correctly answering the item given the item’s characteristics and examinee’s ability. (Hambleton, Swaminathan, & Rogers, 1991).

IRT models have the following assumptions: (1) unidimensionality: there is only one dominant proficiency/trait being assessed in an assessment. This latent trait determines the performance of items. And while there are multidimensional IRT models, this study will only focus on unidimensional IRT models as they are the most commonly used in current test practices; (2) local independence: item answers are mutually independent given a proficiency level. In other words, the response on an item does not affect the response on another after conditioning on the ability of the examinee; (3) monotonicity: when the proficiency of an examinee increases, the probability of a right answer also grows. A chosen IRT model captures the true relationship between the latent ability and the observable item response.

The IRT models are referred to as the 1, 2, or 3-parameter logistic (PL) model, depending on the number of parameters that take into consideration. Equation (1) is the widely used 3PL logistic model:

\[
P(y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j) \frac{\exp [Da_j(\theta_i - b_j)]}{1 + \exp [Da_j(\theta_i - b_j)]}
\] (2.1.1)
Where $y_{ij} = 1$ represents the correct response of item $j$ for a person $i$; $\theta_i$ is the ability level of person $i$; $a_j, b_j,$ and $c_j$ are the discrimination, difficulty, and guessing parameters for item $j$; $P_{ij}(\theta)$ is the probability of person $i$ answering item $j$ correctly. $D$ is a scaling constant ($D = 1.7$ to scale the logistic to the normal ogive metric; $D = 1$ to use the logistic metric). The 3PL model becomes a 2PL model when $c_j = 0$. Additionally, the 3PL model becomes the Rasch model when $c_j = 0$ and $a_j = 1$.

We usually assume that $\theta$ is a continuous random variable following a standardized normal distribution so that most examinees’ ability values will fall between -3 and 3. The higher the ability level, the higher the probability of the correct answer to items. The item parameter $a_j$ describes the discrimination of the item, that is how this item discriminates people with different proficiency levels. The parameter $b_j$ represents the difficulty of the item, which corresponds to the point on the ability scale where the probability of a correct answer equals $(1 + c_j)/2$. For multiple choices questions, the parameter $c_j$ account for the probability of a correct response by guessing.

The item characteristic curve (ICC) can be created to graphically show an item's function using the item parameters of the model. The following graph shows the ICC of the 3PL IRT model with $a = 1.1, b = 0.5$, and $c = 0.25$, in which the probabilities of correct responses are plotted against ability values.
Figure 1. An Example of 3PL IRT ICC

2.2 Detection of Aberrant Responses

No matter which of these aberrant behaviors is present, the examinee's performance on the item will usually be affected. For example, seeking help from a friend who already took the test will lead an examinee to get the right answers to questions on a test that they are unable to accurately answer fairly. When an examinee responds carelessly, he or she may answer certain questions poorly even though they are ones that they are capable of answering right. Therefore, the IRT model could not capture the true relationship between the latent ability of examinees and the observable item responses. Careless response violates the IRT assumptions and undermines the accuracy of the ability estimation of examinees, which might lead to invalid conclusions and inferences based on assessment scores (Shao, 2016).
The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) pointed out: when IRT models are used for test development, evidence of model fit needed to be provided. Before SPC methods were introduced to the educational measurement field, much of the previous literature discussing aberrant response detection relies on the use of PFS. The common theme among these methods is to find responses that do not fit an IRT model.

Meijer and Sijtsma (2001), Karabatsos (2013), and Rupp (2013) provided comprehensive reviews of various PFS. Figure 1.1 is a summary of the 36 PFS from Karabatsos's (2013) article, including both parametric and non-parametric methods. Currently, there is no single PFS that can guarantee to be the most efficient one to detect all kinds of aberrant responses. However, the $l_z$ statistic developed by Drasgow, Levin, and Williams (1985) is one of the most widely used IRT-based PFS. We will explain this index in detail as an example for this thesis. Here is the equation of $l_z$ statistic:

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}},$$  \hspace{1cm} (2.2.1)

where $l_0$ is the log-likelihood function for $i^{th}$ examinee’s observed score. $E(l_0)$ and $Var(l_0)$ are the expectation and variance of $l_0$. Levine and Rubin developed the $l_0$ firstly in 1979:

$$l_0 = \sum_{i=1}^{n}\{u_{ij}lnP_{ij}(\theta) + (1 - u_{ij})ln[1 - P_{ij}(\theta)]\},$$  \hspace{1cm} (2.2.2)

where $u_{ij}$ is the response of $i^{th}$ examinee on the $j^{th}$ item. $P_{ij}$ is the probability of the examinee with the ability $\theta$ answering the $j^{th}$ item correctly.
\( E(l_0) \) and \( Var(l_0) \) are given below:

\[
E(l_0) = \sum_{j=1}^{j} \left\{ P_{ij}(\theta) \ln P_{ij}(\theta) + \left( 1 - P_{ij}(\theta) \right) \ln [1 - P_{ij}(\theta)] \right\}, \quad (2.2.3)
\]

\[
Var(l_0) = \sum_{j=1}^{j} \left\{ P_{ij}(\theta) \left( 1 - P_{ij}(\theta) \right) \ln \left[ \frac{P_{ij}(\theta)}{1 - P_{ij}(\theta)} \right]^2 \right\}. \quad (2.2.4)
\]

Because of the central limit theorem, the null distribution of \( l_z \) is assumed to be standard normal for long tests. However, Molenaar and Hoijtink (1990) believed the null distribution of \( l_z \) is only standard normal when we use true \( \theta \). In actuality, we never know the true \( \theta \) of a person; instead, an estimate \( \hat{\theta} \) is utilized. Examinees’ estimated \( \hat{\theta} \) can be significantly higher or lower than their true \( \theta \) if they exhibit any aberrant responses. It will affect the distributional characteristic of \( l_z \).

There were a few studies found when using the maximum likelihood estimation method to estimate \( \theta \), the variance of \( l_z \) was smaller than the expectation of standard normal distribution, especially for tests with short to moderate length (< 50 items; van Krimpen-Stoop & Meijer, 1999; Nering, 1995). De La Torre and Deng (2008) developed a strategy that used a Bayesian procedure to adjust the ability estimates and null distribution in order to overcome these issues. Three distinct aberrant responses—lack of motivation, cheating, and speed—were used to examine their proposed methodology. The largest power under the simulation settings was 0.3, and neither the \( l_z \) nor their proposed approach had an acceptable power for speededness identification.
2.3 Modeling of Aberrant Responses

In the field of education and psychology assessments, aberrant responses are very common, such as carelessness, speededness, lack of motivation, cheating, pre-knowledge of items on the test, warm-up effect, and fatigue. This section focuses on a very common aberrant behavior, speededness, and its models. Speededness models can also be easily applied to modeling other aberrant behaviors, such as warm-up effects (Shao, 2016), and randomly guessing items at end of the test (Yu & Cheng, 2019).

Test speededness occurs when not all respondents have sufficient time to fully consider the answer for each question on a test within a fixed time limit (Bejar, 1985). Under the IRT framework, Hambleton and Swaminathan (1985) pointed out that
unidimensional IRT models implicitly assume that the test is unspeeded. The unidimensional assumption will be violated if a test is speeded. This section mainly discusses three widely used categories: mixture models, hybrid models, and graduate change models.

2.3.1 Mixture Model

Based on the mixture Rasch model (Rost, 1990), Bolt, Cohen, and Wollack (2002) modeled speededness by classifying examinees into two classes: speeded and non-speeded. As shown in equation (2) (the same as \( a_j = 1 \) and \( c_j = 0 \) in equation 2.1.1), the probability of answering the item correctly for different class still follow the regular IRT model. But item parameters for each class are different. By imposing constraints on item parameters, the mixture model achieves detecting speededness:

\[
P_{ij|g} = \frac{\exp (\theta_{ig} - b_{jg})}{1 + \exp (\theta_{ig} - b_{jg})}.
\]

Where \( g \) is the index of latent class (speeded or non-speeded), \( b_{jg} \) is the difficulty parameter of item \( j \) in class \( g \); \( \theta_{ig} \) is the ability level of person \( i \) in class \( g \). \( P_{ij|g} \) is the probability of person \( i \) answering item \( j \) correctly. Taking the difficulty parameter as an example, for items at the early stage of a test (probably not affected by speededness), they are constrained to be equal across two classes. But for items at the end of the test (assumed to be affected by speededness), they are constrained to be more difficult in the speeded class than in the non-speeded class. This model also assumes all examinees start to speed at the same location.

This mixture Rasch model has been further extended to the mixture 2PL model (Bolt, Mroch, & Kim, 2003), two-dimensional mixture 2PL model (De Boeck, Cho, &
Wilson, 2011), and mixture hierarchical model (Wang & Xu, 2015).

### 2.3.2 Hybrid Model

Yamamoto and Everson (1997) proposed a 2PL hybrid model, which assumed that every item before the speeding point still follows the ordinary 2PL model, but items after the point are modeled by random guessing. The model assumes that examinees in the speeded group will randomly choose a response option when running out of time. The hybrid model takes the following form:

$$P_{ij|g}^* = \left( \frac{\exp \left( a_j(\theta_i - b_j) \right)}{1 + \exp \left( a_j(\theta_i - b_j) \right)} \right)^{1-I_{j|g}} \ast \left( r_{j|g} \right)^{1-I_{j|g}}. \quad (2.3.2)$$

Where \( \frac{\exp \left( a_j(\theta_i - b_j) \right)}{1+\exp \left( a_j(\theta_i - b_j) \right)} \) is the ordinal 2PL model which is the probability of an examinee \( i \) with latent trait \( \theta_i \) answering item \( j \) correctly. Unlike the mixture model, this model incorporates several speeded classes based on their speeding points, \( g \) represents latent classes, \( r_{j|g} \) is the probability of examinees in the \( g \)th class who randomly guess the answer on the \( j \)th item, \( I_{j|g} \) is the indicator function. When \( I_{j|g} = 1 \), it represents the examinee in class \( g \) speeded on the \( j \)th item, the probability of correctly answer is \( r_{j|g} \), which is the random guessing rate. When \( I_{j|g} = 0 \), it represents the examinee is not speeded with the probability of correct answer of \( \frac{\exp \left( a_j(\theta_i - b_j) \right)}{1+\exp \left( a_j(\theta_i - b_j) \right)} \). This model has a strong assumption that the random guessing rate of item \( I \), \( r_{j|g} \), is fixed for all speeded examinees, which equals the reciprocal of the number of response options of a multiple-choice item \( i \).

Yamamoto and Everson (1997) used a simulation study to show item and ability parameter estimation accuracies were dramatically increased compared to the
ordinary 2PL model when speededness occurred and the hybrid model was indeed the true model. Yu and Cheng (2022) adopted this model to generate data that examinees randomly answered items at the end of the test because of carelessness.

2.3.3 Graduate Change Model

Unlike the hybrid model, which assumes the random guessing rate is fixed after the speededness point, Wollack and Cohen (2004) firstly proposed the graduate change model that allows for a gradual decline in the probability of correct responses to generate data. This model mimics the effect of increasing time pressure on test takers (Yu & Cheng, 2022). In 2008, Goegebeur et al successfully showed how to fit this model and estimate parameters. Here is the 2PL graduate change model (Suh, Cho, & Wollack, 2012):

\[
P_{ij}^* = \frac{\exp \left( a_j (\theta_i - b_j) \right)}{1 + \exp \left( a_j (\theta_i - b_j) \right)} \times \min \left( 1, \left[ 1 - \left( \frac{j}{J} - \eta_i \right) \right] \right)^{\lambda_i},
\]

(2.3.3)

where \( \frac{\exp \left( a_j (\theta_i - b_j) \right)}{1 + \exp \left( a_j (\theta_i - b_j) \right)} \) is the ordinary 2PL model, and \( J \) is the total item number of a test. \( \eta \) is at which percentage of the test an examinee starts to speed \( (0 \leq \eta \leq 1) \). For example, if \( \eta_i = 0.7 \), which means this examinee starts to speed from 70% of the test. \( \lambda \) is the speededness rate parameter, which controls the decreased speed of probability. The larger the value \( \lambda \) takes, the faster \( P_{ij}^* \) will decrease. If the examinee does not speed, \( \eta_i = 1 \) and \( \lambda_i = 0 \), then \( \min \left( 1, \left[ 1 - \left( \frac{j}{J} - \eta_i \right) \right] \right)^{\lambda_i} = 1 \), thus this model will be the ordinal 2PL model. With two different \( \lambda \), two examinees who have the same ability level and speededness location might have very different probabilities of answering the same item correctly.
Equation 2.2.3 has been extended to the 3PL version of the graduate change model by Goergebeur et al. (2008). The 2PL gradual change model has been easily adapted to the model warm-up effect by Shao (2016).

In summary, all the above-mentioned models used different assumptions to model a very common aberrant behavior in testing: speededness. Each model has its own set of assumptions. Mixture models treat the testing process as a mixture of normal responding behavior and speededness with different item parameters of ordinary IRT models. Examinees will be classified into 2 groups: speeded vs. non-speeded. While hybrid models classified examinees into multiple groups based on their speeding points, hybrid models also make a stringent assumption that all examinees will randomly guess answers to items after the speeding point. Graduate change models agreed that the examinee will randomly choose an option when running out of time. But unlike hybrid models which assume the change is abrupt, graduate change models propose a graduate decline in the probability of correct answers. Compared to the previous two models, graduate the change model can better represent the real-life behaviors of examinees (Goergebeur et al., 2008; Suh et al., 2012).

2.4 Cumulative Sum Control Charts (CUSUM)

The CUSUM chart is a widely used method of SPC. Yu and Cheng (2022) defined SPC as “a collection of methods for monitoring, controlling, and improving a random process through statistical analysis” (p. 2). It was initially developed and used to monitor product quality in production or manufacturing areas. By using SPC, the product quality can be actively measured and charted at the same time while manufactured things are mass-produced. For example, if one company makes
chocolate beans, each bag should have a certain amount of beans. If each bag has too many beans that will increase the costs of materials for the company. But customers will not be satisfied if there are too few beans in a bag. Therefore, the company needs to control the number of chocolate beans during the production process, which can be achieved through SPC.

There are many different types of control charts – Montgomery (2013) provided a detailed review. The CUSUM (Page, 1954) is an established SPC method that has been widely used in finance (Bodnar & Schmid, 2007), supply chain management (Lee & Wang, 2003), and health care management (Spiegelhalter et al., 2003). Recently, SPC was introduced to educational measurement fields to detect aberrant responses. For example, Omar (2010) used Shewhart's mean and standard deviation charts to measure and monitor the consistency of rating performance items in operational assessments.

In 1998, Bradlow, Weiss, and Cho firstly adopted the CUSUM method to detect four types of aberrant behaviors, representing the effects of warm-up, fatigue, sub-expertise, and lack-of-fit, in a CAT test. van Krimpen-Stoop and Meijer (2002) applied the CUSUM to detect pre-knowledge for polytomous items. Egberink et al. (2010) used the CUSUM method to detect inconsistent item score responses to the Workplace Big Five personality questionnaire. Lee and Lewis (2021) used this method to identify reused items that were exposed in continuous testing.

Like many other SPC methods, CUSUM requires identifying a variable that represents the quality of the process needing statistical control (Omar, 2010). Researchers have proposed various CUSUM indices to represent the test-taking quality
in an educational test. van Krimpen-Stoop and Meijer (2000) proposed 8 CUSUM indices, they were some kinds of residuals (weighted or unweighted) between the expected and observed score of an item. Other CUSUM indices were based on the likelihood ratio test (Armstrong & Shi; 2009a Sinharay, 2016). These indices can be categorized into two groups: one-sided and two-sided. This section will review twelve CUSUM statistics and their corresponding procedures.

2.4.1 One-Sided CUSUM

One-sided indices consider the direction of change in the examinee's performance: an upward change means that the examinee has a good/correct response; a downward change means that the examinee has a worse/incorrect response. All eight indices in van Krimpen-Stoop and Meijer's (2000) article belong to this group. They are labeled as $T_{1j}$ to $T_{8j}$. Given a test with $N$ items, $j$ denotes the $j^{th}$ item in a test ($j > 1$). $Y_j$ is the observed score of the $j^{th}$ item. $\hat{\theta}_j$ is the estimated ability of the person based on his/her responses to first $j$ items. $\hat{\theta}_n$ is the ability estimate based on the whole test. $P_j(\hat{\theta})$ is the probability of answering item $j$ correctly based on the corresponding IRT model. $I(\hat{\theta}_j)$ and $I(\hat{\theta}_n)$ are the test information calculated at $\hat{\theta}_j$ and $\hat{\theta}_n$, respectively.

\begin{align*}
T_{1j} &= \frac{1}{N} \{Y_j - P_j(\hat{\theta}_j)\}, \quad (2.4.1) \\
T_{2j} &= T_{1j} \times \left\{ P_j(\hat{\theta}_j)[1 - P_j(\hat{\theta}_j)] \right\}^{-\frac{1}{2}}, \quad (2.4.2) \\
T_{3j} &= T_{1j} \times \left\{ I(\hat{\theta}_j) \right\}^{-\frac{1}{2}}, \quad (2.4.3) \\
T_{4j} &= T_{1j} \times \sqrt{j}, \quad (2.4.4) \\
T_{5j} &= \frac{1}{N} \{Y_j - P_j(\hat{\theta}_n)\}, \quad (2.4.5)
\end{align*}
\[ T_{6j} = T_{5j} \times \left\{ P_j(\hat{\theta}_n)[1 - P_j(\hat{\theta}_n)] \right\}^{-\frac{1}{2}}, \]

\[ T_{7j} = T_{5j} \times \left\{ I(\hat{\theta}_n) \right\}^{-\frac{1}{2}}, \]

\[ T_{8j} = T_{k5} \times \sqrt{j}. \]

\( T_{1j} \) to \( T_{4j} \) are functions of residuals, where probability is evaluated at the estimated ability of \( \hat{\theta}_j \). \( T_{5j} \) to \( T_{8j} \) use probability that is evaluated using \( \hat{\theta}_n \). All those CUSUM indices are summed across consecutive items. For each examinee, after each administered item \( j \), the CUSUM can be shown as:

\[ C_0^+ = C_0^- = 0, \]

\[ C_j^+ = \{0, T_j + C_{j-1}^+\}, \]

\[ C_j^- = \{0, T_j + C_{j-1}^-\}. \]

The start points \( C_0^+ \) and \( C_0^- \) are 0. \( C_j^+ \) and \( C_j^- \) are the cumulative sum of the consecutive positive and negative residuals \( T_j \). Based on equations 2.4.10 and 2.4.11, we know \( C_j^+ \) is always positive and \( C_j^- \) is always negative. A series of consecutive positive values of \( T_j \) will make \( C_j^+ \) larger, while consecutive negative values of \( T_j \) will make \( C_j^- \) smaller. Let UB and LB represent the pre-specified upper and lower bound, respectively. If \( C_j^+ > UB \), or \( C_j^- < LB \) at some points, we can identify this response pattern as aberrant. Otherwise, we will classify this item score pattern as fitting the unidimensional IRT model. The CUSUM control chart can be created based on the scatter plots of \( C_j^+ \) and \( C_j^- \) at every item point \( j \). Figure 3 is a CUSUM control chart based on \( T_{5j} \).
2.4.2 Two-Sided CUSUM

One-sided indices $T_1$ to $T_8$ will be plugged into functions 2.4.10 and 2.4.11 to obtain $C_j^+$ and $C_j^-$, respectively. There are two CUSUM values, positive or negative, associated with each item when we adopt one-sided indices. However, two-sided CUSUM integrates information from $C_j^+$ and $C_j^-$ together. It will only be one cumulative value when we utilize two-sided indices. There are four common two-sided indices, we denote them as $T_9$ to $T_{12}$.

$T_9$ was first developed in Bradlow et al. (1998) article:

$$T_9 = \frac{\left| \sum_{j=1}^{n} [Y_j - P_j(\hat{\theta}_j)] \right|}{\sqrt{\sum_{j=1}^{n} P_j(\hat{\theta}_j)[1 - P_j(\hat{\theta}_j)]}}.$$  \hspace{1cm} (2.4.12)
\( T_{9j} \) will be plugged into:

\[
C_{T9} = \max_{1 \leq j \leq n} T_{9j}.
\]  

(2.4.13)

For each item. Unlike previous indices that are some functions of residuals, Armstrong and Shi (2009a) suggested likelihood ratio statistics \( T_{10j} \) for the CUSUM. Their statistic \( T_j \) were defined differently in \( C_j^+ \) and \( C_j^- \). \( T_{10j}^U \) and \( T_{10j}^L \) are likelihood ratio statistics for testing whether there is an “aberrant upward shift” and “aberrant downward shift” of the probability of a correct answer, respectively:

\[
T_{10j}^U = \ln \frac{g^U_j[P_j(\theta)]}{P_j(\theta)},
\]

(2.4.14)

\[
T_{10j}^L = \ln \frac{g^L_j[P_j(\theta)]}{P_j(\theta)},
\]

(2.4.15)

where \( g^U_j[P_j(\theta)] \) and \( g^L_j[P_j(\theta)] \) are continuous curves that match an aberrant upward and downward shift, respectively. Armstrong and Shi (2009a) provided detailed steps for calculating them in their article. Then cumulative values based on \( T_{10j} \) are:

\[
C_{T_{10}} = \max_{1 \leq j \leq n} \left( C_{T_{10j}}^+ \right) - \min_{1 \leq j \leq n} \left( C_{T_{10j}}^- \right),
\]

(2.4.16)

where

\[
C_{T_{10j}}^+ = \max(0, C_{T_{10j-1}}^+ + T_{10j-1}^U)
\]

(2.4.17)

\[
C_{T_{10j}}^- = \min(0, C_{T_{10j-1}}^- + T_{10j-1}^L).
\]

(2.4.18)

van Krimpen-Stoop and Meijer (2001) also proposed a two-sided CUSUM index based on the \( l_z \):

\[
T_{11j}^U = l_z + 0.5,
\]

(2.4.19)

\[
T_{11j}^L = l_z - 0.5.
\]

(2.4.20)
The procedure of calculating $C_{T_{11}}$ is very similar to $C_{T_{10}}$:

$$C_{T_{11}} = \max_{1 \leq j \leq n} \left( C_{T_{11j}}^+ \right) - \min_{1 \leq j \leq n} \left( C_{T_{11j}}^- \right),$$  \hfill (2.4.21)

where

$$C_{T_{11j}}^+ = \max \left( 0, C_{T_{11,j-1}}^+ + T_{11,j-1}^U \right)$$  \hfill (2.4.22)

$$C_{T_{11j}}^- = \min \left( 0, C_{T_{11,j-1}}^- + T_{11,j-1}^L \right).$$  \hfill (2.4.23)

Using the one-sided index $T_{1j}$, Sinharay (2016) suggested another two-sided CUSUM procedure:

$$T_{12j} = T_{1j} = \frac{1}{N} \left\{ X_j - \bar{p}_j(\hat{\theta}_j) \right\},$$  \hfill (2.4.24)

$$C_{T_{12}} = \max_{1 \leq j \leq n} \left( C_{T_{12j}}^+ \right) - \min_{1 \leq j \leq n} \left( C_{T_{12j}}^- \right),$$  \hfill (2.4.25)

where $C_{T_{12j}}^+$ and $C_{T_{12j}}^-$ can be obtained by the general procedure of the CUSUM, that is,

$$C_{T_{12j}}^+ = \max \left\{ 0, T_{12j} + C_{T_{12,j-1}}^+ \right\}, C_{T_{12j}}^- = \min \left\{ 0, T_{12j} + C_{T_{12,j-1}}^- \right\}.$$

### 2.4.3 Identify Boundary Values of CUSUM

For traditional CUSUM procedures, that variable needing statistical control usually follows the normal distribution, such as the weight of tea bags. It is easy to define UB and LB using the mean and standard deviation of that variable. However, the null distributions of proposed CUSUM indices are usually far away from normal. For example, the residual $T_{5j}$ given in equation 2.4.5 follows a bimodal distribution.

There are several methods to determine UB and LB. In general, it is necessary to define a level of statistical significance (usually at a 5% level) first. Then the most extreme value is found and the UB and LB are determined by choosing that value for which 2.5% of the most extreme values lie above UB or below LB. This process can be
achieved through Monte Carlo simulation or based on the empirical dataset at hand.

Most previous studies used Monte Carlo simulation to find UB and LB. When using the one-sided CUSUM indices, the following steps are an example:

1. 1000 examinees' abilities ($\theta$) are generated from the standard normal distribution.
2. Item scores are generated based on examinees’ $\theta$s and item parameters. The probability of endorsing items, the CUSUM index, and its cumulative values (C+ and C-) were calculated.
3. The maximal C+ and minimal C- values of each examinee were collected. Then the LB and UB were identified as 2.5% and 97.5% percentile of 1000 extreme values, respectively.
4. The previous steps were repeated 100 times, and a final LB and UB was the average value across 100 replications.

If the testing organization has large sample sizes of examinees, an alternative method is to use the empirical dataset at hand and select groups of examinees with approximately the same $\hat{\theta}$ value. Then UB and LB values for these groups of examinees can be determined by step 3. There is a drawback of using empirical item responses instead of simulation: the existing misfitting item scores might affect boundary values. However, we expect most item score patterns will fit the underlying IRT model in reality. To investigate the influence of misfitting item responses, Meijer (2002) used both simulation and empirical datasets and found similar bounds. So, I would recommend using an empirical dataset if your sample size is big enough.
2.5 Change-Point Analysis (CPA)

CPA is another popular SPC method. It was first developed by Page in 1954. Since then, it has been extensively used in fields including medicine (Aminikhanghahi & Cook, 2017; Kass-Hout et al., 2012), environmental climate (Suhaila & Yusop, 2018; Yu & Ruggieri, 2019), economics (Thies & Molnár, 2018), and industry (Maleki et al., 2016; Mortaji et al., 2015).

CPA has also been used in educational testing in recent years to detect abnormal changes in the mean score of international testing programs over time (Lee & von Davier, 2013), or changes in the performance of test items over time due to potentially exposed item pool (Zhang, 2013), or aberrant response during the test-taking process due to item pre-knowledge, speededness, person misfit (Shao et al., 2016; Sinharay, 2016, 2017a, 2017b), or carelessness (Yu & Cheng, 2022). It takes a completely different approach than CUSUM, which intends to detect whether there has been any change in the parameter(s) underlying a sequence of random variables. But the same as CUSUM, it also requires using PFS to achieve detection.

2.5.1 Four CPA Statistics

If there is a change point in an examinee’s response sequence, the sequence can be divided into two subsequences: \( Y_1, Y_2, \ldots, Y_j \) (subsequence 1), and \( Y_{j+1}, Y_{j+2}, \ldots, Y_n \) (subsequence 2). There is a fundamental difference between the two subsequences statistically, and the PFS of CPA can quantify this difference. If the PFS is maximum and significant at the \( j^{th} \) item, it indicates the change point is at the \( j^{th} \) item.

There are four common PFS indices in psychology and education testing: (1) \( L_{max} \) based on the likelihood ratio test (LRT), (2) \( W_{max} \) based on the Wald test, (3)
$S_{\text{max}}$ based on the score test, and (4) $R_{\text{max}}$ based on weighted residual. The first three PFS were proposed by Shao et al. (2016) and Sinharay (2016, 2017a, 2017b, 2017c), while the last one was proposed by Yu and Cheng (2019). The development of these four indices shares the same principle. Let $\theta_{1j}$ be the true ability of an examinee's underlying score on items 1 to $j$, $\theta_{2j}$ be the underlying scores on items $(j+1)$ to $n$. $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ are corresponding maximum likelihood estimates (MLEs) ability estimates based on responses before and after the change point ($j^{th}$ item).

### 2.5.1.1 $L_{\text{max}}$ based on the likelihood ratio test

According to Shao et al. (2016) and Sinharay (2016), when $j$ is known, the LRT statistic for testing the null hypothesis $\theta_{1j} = \theta_{2j}$ is:

$$L_{jn} = -2 \{L(\hat{\theta}_n; Y_1, Y_2, ..., Y_n) - L(\hat{\theta}_{1j}; Y_1, Y_2, ..., Y_j) - L(\hat{\theta}_{2j}; Y_{j+1}, Y_{j+2}, ..., Y_n)\}, \quad (2.5.1)$$

where $\hat{\theta}_n$ is the ability estimate of an examinee based on all responses. We can obtain these estimates based on traditional IRT models. Taking $L(\hat{\theta}_{1j}; Y_1, Y_2, ..., Y_j)$ as an example, the specific LRT function is given below:

$$L(\hat{\theta}_{1j}; Y_1, Y_2, ..., Y_j) = \sum_{j=1}^{n} \left[ Y_j \log P_j(\hat{\theta}_{1j}) + (1 - Y_j) \log \left\{1 - P_j(\hat{\theta}_{1j})\right\}\right]. \quad (2.5.2)$$

$P_i(\theta_{ij})$ is the probability of $i^{th}$ person with $\theta_i$ answering $j^{th}$ item correctly based on different IRT models (equation 2.1.1). The LRT statistics can be used on both linear tests and CAT. When $j$ is known, the null hypothesis of $L_{jn}$ follows a $\chi^2$ distribution with degree of freedom equals to 1 asymptotically (Finkelman et al., 2010).

In actuality, the change point is unknown, and the number of possible change points is $(n-1)$. As a result, the corresponding statistic is calculated as the maximum of
all potential change points. The PFS is defined as:

\[ L_{\text{max}} = \max_{1 \leq j \leq n-1} \{ L_j \} \quad (2.5.3) \]

if \( L_{\text{max}} \) is significantly greater than 0, the response is labeled as aberrant, with the change point considered to be the point at which \( L_{\text{max}} \) is reached.

2.5.1.2 \( W_{\text{max}} \) based on the Wald test

The Wald test can also be used to test whether an examinee’s ability (denoted by \( \theta \)) changed from subsequence 1 to subsequence 2. The equation of the Wald test is:

\[ W_j = \frac{(\hat{\theta}_{1j} - \hat{\theta}_{2j})^2}{I_{1j}(\hat{\theta}_n) + I_{2j}(\hat{\theta}_n)} \quad (2.5.4) \]

where \( I \) denotes the sum of Fisher’s information for all items in the corresponding response sequences (sequences 1 and 2). The item information is a measure of the accuracy that an item can provide to a respondent with a particular ability value in the IRT. The greater the amount of information, the better the item is for measuring the examinee. It is worth noting that the information is calculated using the \( \hat{\theta}_n \), which is the ability estimate of an examinee based on all responses. The test information function is given below:

\[ I_j(\theta) = \sum_{i=1}^{n} P_i(\theta_{ij}) \left[ 1 - P_i(\theta_{ij}) \right] \quad (2.5.5) \]

\( P_i(\theta_{ij}) \) is the probability of \( i^{th} \) person with \( \theta_i \) answering \( j^{th} \) item correctly based on different IRT models (equation 2.1.1). Similar to \( L_{\text{max}} \), the PFS of the Wald test is defined as:

\[ W_{\text{max}} = \max_{1 \leq j \leq n-1} \{ W_j \} \quad (2.5.6) \]

for evaluating the null hypothesis of no change in ability (\( \theta_{1j} = \theta_{2j} \)) versus the
alternative of a change. However, previous research found the power of \( W_{\text{max}} \) is small when the change happens near the first or last observations of a sequence (Andrews, 1993). Thus, Andrews (1993) suggested limiting the location of the change point to approximately 70% of the middle of the entire response sequence. Therefore, \( W_{\text{max}} = \max_{n_1 \leq j \leq n-n_1} \{W_j\} \), \( n_1 \) is taken as an integer near 0.15\( n \) to enhance detection effectiveness.

2.5.1.3 \( S_{\text{max}} \) based on the score test

The score test statistic to test the null hypothesis that \( \theta_{1j} = \theta_{2j} \), when \( j \) is known, is given by

\[
S_{jn} = \left( \frac{\nabla(\hat{\theta}_n; Y_1, Y_2, ..., Y_j)}{I_1(\hat{\theta}_n)} \right)^2 + \left( \frac{\nabla(\hat{\theta}_n; Y_{j+1}, Y_{j+2}, ..., Y_n)}{I_2(\hat{\theta}_n)} \right)^2,
\]

(2.5.7)

where \( \nabla(\hat{\theta}_n; Y_1, Y_2, ..., Y_j) \) and \( \nabla(\hat{\theta}_n; Y_{j+1}, Y_{j+2}, ..., Y_n) \) are the first-order derivatives with respect to \( \hat{\theta}_n \) of the log-likelihood of subsequences 1 and 2, respectively. The test statistic is the maximum of the \( S_{jn} \):

\[
S_{\text{max}} = \max_{1 \leq j \leq n-1} \{S_{jn}\}.
\]

(2.5.8)

A response pattern will be flagged if \( S_{\text{max}} \) is significantly larger than 0, and the change point is estimated to be the point where \( S_{\text{max}} \) is achieved. Similar to \( W_{\text{max}} \), the location of the change point can be limited from \( n_1 \) to \( n - n_1 \) to enhance detection effectiveness.

2.5.1.4 \( R_{\text{max}} \) based on weighted residual

To detect random guessing behaviors of examinees, Yu and Cheng (2019) proposed another PFS indicator based on weighted residuals. The difference between the observed item response (observed score) and the expected score given a
respondent's latent trait (estimate) and the model is described as a residual. For normal test-takers, the observed score pattern should be very close to the expected score pattern, and a small residual is anticipated. However, we expect to see relatively large residuals of aberrant response behavior.

The principle of this index is to find a point at which the whole response sequence can be divided into two subsequences, and which minimizes the difference between the average absolute weighted residuals (ABWR) of the two subsequences. Yu and Cheng (2019) constructed the PFS based on the graded response model (GRM; Samejima, 1969). Here we provide the equation of weighted residual for dichotomous items for consistency:

\[
 r_j(\hat{\theta}) = \frac{Y_j - P(\hat{\theta})}{P(\hat{\theta})},
\]  

(2.5.9)

where the numerator is the usual residual term, that is, the difference between observed and expected scores. The denominator is the probability of a person with \( \hat{\theta} \) answering \( j^{th} \) item correctly based on different IRT models. To detect inattentiveness at the end of the test, the weighted item residual \( r_j(\hat{\theta}) \) is used to formulate a test statistic \( R_{jn} \):

\[
 R_{jn} = \frac{1}{n-j} \sum_{j+1}^{n} |r_j(\hat{\theta}_{1j})| - \frac{1}{j} \sum_{1}^{j} |r_j(\hat{\theta}_{1j})|.
\]

(2.5.10)

Because Yu and Cheng (2019) wanted to detect aberrant responses that happened at a late stage of the test, equation 2.5.10 used the ability estimate based on normal responses before the change point. If we plan to detect aberrant responses at the early stage of the test, we can use the following equation:
\[ R_{jn} = \frac{1}{j} \sum_{1}^{j} |r_j(\hat{\theta}_{2j})| - \frac{1}{n - j} \sum_{j+1}^{n} |r_j(\hat{\theta}_{2j})| . \]  

Finally, the test statistic is the maximum of the \( R_{jn} \):

\[ R_{\text{max}} = \max_{1 \leq j \leq n-1} \{ R_{jn} \} . \]

However, unlike previous indices, \( L_{\text{max}} \), \( W_{\text{max}} \) and \( S_{\text{max}} \), \( R_{\text{max}} \) does not test the null hypothesis \( \theta_{1j} = \theta_{2j} \). It does not have an analytical asymptotic distribution, either. For \( R_{jn} \), the ABWR reflects the degree of deviation between the observed and expected score patterns of that subseries. When the difference between the ABWR of the subseries before and after the change point exceeds a certain range, it indicates that the examinee has an aberrant response in the test. Yu and Cheng (2019) adopted the Monte Carlo simulation approach to construct the null distribution of \( R_{\text{max}} \).

The four PFS of CPA are all based on the principle of determining whether there is a point at which the sequence of responses can be divided into two parts with fundamentally different statistical properties and locating the change point. The previous three PFS, \( L_{\text{max}} \), \( W_{\text{max}} \) and \( S_{\text{max}} \), are used to test the null hypothesis \( \theta_{1j} = \theta_{2j} \). They were developed as two-sided test indicators. That is, when the detection goal is only to check whether there is an abnormal response in the sequence, and the type of abnormality is not specified, it is better to use these indicators. Of course, these three indicators are also available in one-sided form, when the goal is to detect a specific type of aberrant response (e.g., speededness), it is appropriate to use one-sided indicators.

Moreover, in practice, \( L_{\text{max}} \), \( W_{\text{max}} \) and \( S_{\text{max}} \), are more suitable for high-stakes, large-scale educational tests. The \( R_{\text{max}} \), is not used to test the null hypothesis, but to
test whether there is any abnormality in the test. When the test administrator has a clear understanding of the specific aberrant behavior to be tested, $R_{max}$ is more appropriate. In practice, $R_{max}$ is more suitable for low-stakes psychometric tests.

2.5.2 Identifying Critical Values of PFS in CPA

When using CPA for aberrant response detection, it is necessary to use PFS, so the determination of the PFS critical values is very important. If the PFS thresholds are not selected appropriately, the accuracy of the detection will be significantly reduced, and the value of CPA will be greatly diminished. This thesis introduces two widely used methods: The Monte Carlo simulation method provided by Worsley (1979) and the false discovery rate (FDR) control method proposed by Storey and Tibshirani (2003).

2.5.2.1 Monte Carlo simulation

The procedure of using Monte Carlo simulation to identify the critical points in CPA is very similar to the CUSUM control chart. The procedures are given below:

1. Generating normal responses of 1,000 examinees: The latent abilities of these 1,000 examinees follow the standard normal distribution, $N (0, 1)$. Based on the examinees’ true abilities and item parameters, we can generate a response matrix without any aberrant behavior.

2. Based on each response matrix, we can calculate $L_{max}$, $W_{max}$, $S_{max}$, or $R_{max}$ for each examinee. It is worth noting that we will use theta estimates ($\hat{\theta}$) to obtain these PFS values. Because we will never know the true latent ability of an examinee in reality.

3. The 1,000 $L_{max}$, $W_{max}$, $S_{max}$, or $R_{max}$ values of examinees can construct
the distribution of the null hypothesis. If we pick a level of significance (e.g., 0.05), the critical value is the 50th largest number among these 1,000 values for this replication.

(4). After step 3, we will have a critical value. Repeat step 1 to step 3 B (e.g., B=100) times, we can have B critical values. The final critical value is the average of them.

2.5.2.2 FDR control method

For a test with N subjects, we need to test N hypotheses simultaneously (i.e., each examinee is tested for aberrant responses). So, we will have N times of comparisons with the critical value, which belongs to the multiple-test. Shao et al. (2016) believed that the critical value should not be set at 0.05 or 0.01 as the significance level according to common practice but should be corrected. There are two common correction methods: (1) the Bonferroni correction: it corrects the significance level to 0.05/N or 0.01/N. However, the sample size N is usually very large in practice, this method is too strict and conservative. It is difficult to reject the null hypothesis; (2) the other method is to control for the FDR (Benjamini & Hochberg, 1995).

FDR is widely used to correct for multiple comparisons in genomics studies (e.g., Benjamini& Hochberg, 1995). FDR represents the expected proportion of false discoveries among all discoveries. This method aims to control the FDR in an acceptable range. For example, if we found there are 100 examinees with aberrant responses, 90 of them performed aberrantly, and the remaining 10 people performed normally. The FDR is 0.1 for this case. There are different ways to estimate FDR. Shao (2016) adopted the procedure proposed by Storey and Tibshirani (2003). Here we use
$L_{max}$ as an example to briefly explain the procedure:

(1). Based on the empirical dataset, we randomly permute the item responses B (e.g., B=100) times and compute the $L_{max}$ value for each person. The $L_{max}$ values of all examinees for each replication can be considered to be those under the null hypothesis of no change (or no person misfit).

(2) Let T be the unknown critical value of $L_{max}$, FDS is defined as:

$$FDR = \frac{B^{-1} \sum_{b=1}^{B} \sum_{i=1}^{N} I(L_{max} > T)}{\sum_{i=1}^{N} I(L_{max} > T)}$$

where b represents the order number of permutations, and I is the indicator function. When $L_{max} > T, I = 1$, it will be flagged as an abnormality; otherwise, $I = 0$. The value of FDR can be adjusted according to the needs of research and application. The statistical community generally recommends setting it to 0.2. Thus, a minimum T value satisfying $FDR \leq 0.2$ can be obtained, as well as the critical value of $L_{max}$.

2.6 Summary of Literature

To overcome the limitation of traditional PFS, such as $l_z$, SPC technical has been introduced to the educational measurement field. CUSUM and CPA are both aberrant response detection methods that are used to analyze whether there are change points in an examinee's response sequence, and both can be classified as "change point analysis" in a broader sense. The CUSUM procedure sums a series of positive or negative residuals (the difference between observed and expected scores) in the order of items to obtain one-sided or two-sided PFS. It can also sum log-likelihood ratios. As a result, after each item, the method can update the CUSUM value. The basic idea
behind CPA is to see if the entire response sequence can be divided into two subsequences at a certain point and if the difference in some statistical property between these two is significant.

The main advantage of CUSUM is that it provides the visualization to quickly identify the location of aberrant responses. In addition, proctors can use CUSUM to monitor test-taking behaviors and intervene in a timely manner. Furthermore, by looking at the charts, CUSUM allows for intuitive and simple multiple change points analysis. Despite all advantages previously discussed, there are still some limitations of the CUSUM method. Firstly, in order to fully understand an examinee’s test-taking process, we need to manually check the CUSUM control chart to locate the change point and to identify potential aberrant behaviors. It can be time-consuming and laborious for a large sample size.

In contrast, CPA has the advantage that it not only determines whether an examinee’s response has change points directly by PFS but also locates the most likely location of a change point if it exists. This is particularly important in large-scale tests in terms of saving human resources. However, CPA has its limitation. When the change-point is located in the first or last stage of the sequence, the effectiveness of CPA will be greatly affected.

Sinharay (2016) implemented a study in a CAT setting and found that CPA-based PFS \((L_{\text{max}}, W_{\text{max}}, S_{\text{max}})\) outperformed CUSUM-based PFS in terms of detection effectiveness. When deciding which method to use in practice, Hawkins et al. (2003) suggested that CUSUM-based PFS is more effective when the examinees' response model before and after the change point is known. However, if one or more
model parameters are unknown, CPA-based PFS is preferred.

CHAPTER 3

METHOD

3.1 General Method

Taylor (2000a) proposed a new procedure that incorporates the CUSUM control chart and CPA. After plotting the CUSUM chart, this new procedure uses bootstrap analysis to identify the existence of a change-point based on the confidence level. Then two estimators are used to locate the change point automatically. This procedure has been used in fields of medicine (Xu et al., 2015), astronomy (Chang et al., 2012), and climate (Arif et al., 2017; Bisai et al., 2014). Despite its widespread use, no study has applied this approach to detect an aberrant response in educational or psychological tests. This dissertation will use a simulation study to evaluate the performance of this procedure. Before explaining the details of the simulation, the General Method section describes essential steps of this new approach that will be applied throughout the simulation involved in this study.

3.1.1 A New CUSUM Procedure Using the Bootstrap Analysis

Let $X_j$ represents the $j^{th}$ observation of the variable needing SPC. It can be the CUSUM index (e.g., $T_j$ in equations 2.4.1) for $j^{th}$ item in a test. When there are $n$ items in the test, the mean value is:

$$
\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.
$$

(3.1.1)

The CUSUM value always starts at 0:

$$
C_0 = 0.
$$

(3.1.2)
The following CUSUM points ($C_j$) are calculated by adding the difference between current $X_j$ and $\bar{X}$:

$$C_j = C_{j-1} + (X_j - \bar{X}). \quad (3.1.3)$$

We will have $n$ CUSUM values: $C_1, C_2, \ldots, C_n$. For this approach, the cumulative sums are not the cumulative sums of the residual values. Instead, they are the cumulative sums of differences between the residual and the average. By plotting $C_1$ to $C_n$, we will obtain a CUSUM control chart. Figure 4 is the example provided by Taylor (2000a), it is a plot of US Trade Deficit Data. The new CUSUM chart always starts at 0 and ends at 0. A CUSUM chart segment with an upward slope indicates a period in which the values tend to be higher than the overall average. Similarly, a segment with a downward slope indicates a period of time when the values are lower than the overall average. A sudden shift or change in the average is indicated by a sudden change in the direction of the CUSUM. When the CUSUM chart follows a relatively straight path, it indicates that the average did not change.

Figure 4. The Example of Bootstrap CUSUM Chart Based on the New Approach

Before running the bootstrap analysis, an estimate of the magnitude of the
changes is required to create a chart boundary. It is calculated as:

\[ C_{diff} = C_{max} - C_{min}, \]  

(3.1.4)

\[ C_{max} = \max_{j=0,1,...,n} C_j, \]  

(3.1.5)

\[ C_{min} = \min_{j=0,1,...,n} C_j. \]  

(3.1.6)

Once the boundary estimator \( C_{diff} \) is obtained, the bootstrap analysis can be performed. A single bootstrap is executed as follows:

1. Randomly reordering original \( X_j \) values, denoted as \( X'_1, ... , X'_n \), as our bootstrap sample, which is also known as sampling without replacement.

2. Based on the bootstrap sample, we can repeat previous steps to obtain corresponding CUSUM values as \( C'_1, C'_2, ... , C'_n \).

3. The magnitude of change for the bootstrap CUSUM is calculated as

\[ C'_{diff} = C'_{max} - C'_{min}. \]

4. If the original magnitude of change is larger than the magnitude of change of bootstrap CUSUM, \( C_{diff} > C'_{diff} \), this bootstrap analysis will be counted.

After performing bootstrap analysis \( M \) times, we can find \( k \) times of \( C_{diff} > C'_{diff} \). Then the confidence level (CL) that a change has occurred as a percentage is defined as:

\[ CL = \frac{k}{M} \times 100 \]  

(3.1.7)

Usually, 90% or 95% confidence is required before we can state that a significant change has been detected. The rationale behind bootstrapping is that the
bootstrap samples reflect random data reordering that mimics the behavior of the CUSUM if no change has occurred. We can estimate how much $C'_{\text{diff}}$ would differ if no change occurred by running a large number of bootstrap samples. This result can then be compared to the $C_{\text{diff}}$ obtained from the data in its original order to see if it is consistent with what we would expect if no change occurred.

The CUSUM chart in Figure 4 of the data in its original order is combined with the CUSUM charts from 5 separate bootstrap samples in Figure 5 (Taylor, 2000a). Compared to the original CUSUM line, bootstrap lines tend to stay closer to zero. This raises the possibility that a change has occurred. Ideally, rather than bootstrapping, we would like to determine the distribution of $C_{\text{diff}}$ based on all possible reorderings of the data. However, this is generally not feasible. If a test has 24 items, the total number of possible reordering is $24! \approx 6.2 \times 10^{23}$. Therefore, the bootstrap analysis is usually performed 1,000 times to estimate the null distribution of $C_{\text{diff}}$. By increasing the number of bootstrap samples, we can have a better estimate. However, for most applications, 1000 bootstraps are sufficient.

![Figure 5. The Bootstrap CUSUM Chart with Original Data and Five Samples](image-url)
The bootstrapping strategy is a distribution-free method with only one assumption, which is an independent error structure. It can be expressed as:

\[ X_j = \mu_j + \varepsilon_j, \tag{3.1.8} \]

where \( \mu_j \) is the average value of item \( j \). Normally, \( \mu_j = \mu_{j-1} \) except for a small number of values of \( j \) that are called change points. Meanwhile, \( \varepsilon_j \) is a random error correlated with the \( j^{th} \) value. It is assumed that \( \varepsilon_j \) are independent with a mean of 0. Taylor (2000b) also provided a procedure to evaluate this assumption. Autoregressive time series data such as stock prices are not appropriate for this method.

### 3.1.2 Locating the Change Point of the New CUSUM Procedure

Once we used the bootstrap analysis to determine if a change has occurred, we can calculate two estimators to find the change-point location. The first one is the CUSUM estimator:

\[ |S_m| = \max_{j=0,1,\ldots,n} |C_j|, \tag{3.1.9} \]

where it is the point furthest from 0 in the CUSUM chart. Point \( m \) estimates the last point before the change. The point \( m+1 \) represents the first point after the change.

Another estimator of when the change occurred is the mean square error (MSE):

\[ MSE_m = \sum_{j=1}^{m} (X_j - \bar{X}_1)^2 + \sum_{j=m+1}^{n} (X_j - \bar{X}_2)^2, \tag{3.1.10} \]

where \( \bar{X}_1 = \frac{\sum_{j=1}^{m} X_j}{m} \), and \( \bar{X}_2 = \frac{\sum_{j=m+1}^{n} X_j}{n-m} \). The rationale behind this is the MSE estimator is similar to CPA. It divides the response sequence into two parts: 1 to \( m \), and \( m+1 \) to \( n \). After obtaining the average of each segment, we can use \( MSE_m \) to see
how well the data fits the two estimated averages. The value of m that minimizes \( MSE_m \) is the best estimator of the last point before the change. And the point \( m+1 \) estimates the first point after the change.

As soon as a change is identified, the data can be separated into two subsequences. Then the same bootstrapping procedure can be applied to those two subsequences separately to determine if there is another change point that can further divide the response sequence. In this manner, we can identify multiple change points with confidence levels. Figure 6 presents the CUSUM chart from Figure 4 with the significant changes shown in the background by Taylor (2000a). It appears that the slope of the CUSUM chart changed twice. The blue background represents the period of change.

![CUSUM Chart](image)

Figure 6. Bootstrap CUSUM Chart with Changes Shown in the Background

### 3.2 Simulation Design

This study aims to investigate whether traditional CUSUM indices (Equations 2.4.5 to 2.4.8) can be used for a new procedure, which we refer to as "bootstrap CUSUM" in this thesis. The performance of the bootstrap CUSUM will also be
compared with that of well-developed CUSUM and CPA for detecting random responses.

To simulate the scenario of random guessing, we use the hybrid model developed by Yamamoto and Everson (1997). Ten percent of the examinees are chosen at random to have random responses. Different severity rates of random responses are considered, which refers to the proportion of items affected by those examinees with aberrant behaviors. The last 30%, 40%, and 50% of the items in the test, referred to as low, medium, and high levels of severity, respectively, are chosen to be answered aberrantly for those respondents.

Specifically, each respondent has a probability of 0.2 to respond correctly to those items regardless of their true ability or the item's true difficulty. For respondents without aberrant behavior, their response patterns will be generated based on the ordinary 2PL IRT model.

3.3 Data Generation

The simulated sample comprises 1,000 examinees with abilities randomly generated from a standard normal distribution. The test follows the 2PL IRT model, and item difficulty parameters are also drawn from a standard normal distribution. Test length is manipulated at three levels: 40, 60, and 80. The response matrix is generated based on the simulated parameters and the hybrid model. In total, there are nine conditions: three test lengths (40, 60, and 80) × three severity rates (30%, 40%, and 50%), with each condition replicated 100 times. Table 1 summarizes the simulation conditions.
Table 1. Simulation Conditions of Study 1

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examinees</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
</tr>
<tr>
<td>Ability distribution ($\theta$)</td>
<td>N (0, 1)</td>
</tr>
<tr>
<td>Test</td>
<td>Length</td>
</tr>
<tr>
<td></td>
<td>40, 60, 80</td>
</tr>
<tr>
<td>Item parameters</td>
<td>Difficulty (b)</td>
</tr>
<tr>
<td></td>
<td>Discrimination (a)</td>
</tr>
<tr>
<td>Random guess</td>
<td>Severity rate (items are affected)</td>
</tr>
<tr>
<td></td>
<td>Probability of correct response on affected items</td>
</tr>
</tbody>
</table>

3.4 Data Analyses and Evaluation Criteria

For this study, we have chosen to use a total of six statistical indices to assess the effectiveness of different detection methods. Specifically, we have selected four indices for bootstrap CUSUM (S1 to S4), one index for traditional CUSUM, and one index for CPA. The equations for each of these indices are provided in Table 2. We have chosen these particular indices to represent the traditional CUSUM and CPA methods due to their demonstrated high detection power in a recent study conducted by Yu and Cheng (2022).

Table 2. Indices for Aberrant Response Detection Methods

<table>
<thead>
<tr>
<th>Detection method</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrap CUSUM</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{N} { Y_j - P_j(\hat{\theta}_n) }$</td>
</tr>
<tr>
<td></td>
<td>S2</td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{N} { Y_j - P_j(\hat{\theta}_n) } \times { P_j(\hat{\theta}_n)[1 - P_j(\hat{\theta}_n)] }^{\frac{1}{2}}$</td>
</tr>
<tr>
<td></td>
<td>S3</td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{N} { Y_j - P_j(\hat{\theta}_n) } \times [ I(\hat{\theta}_n) ]^{\frac{1}{2}}$</td>
</tr>
</tbody>
</table>
To determine the boundary values for the traditional CUSUM procedure and the critical point of the CPA statistic, we utilized Monte Carlo simulation. The results of this simulation are presented in Table 3, which provides the relevant values for each of these statistical methods.

Table 3. Boundary Values of CUSUM and Critical Value of CPA

<table>
<thead>
<tr>
<th>TL</th>
<th>Upper boundary</th>
<th>Lower boundary</th>
<th>CPA ($L_{max}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.261</td>
<td>-0.263</td>
<td>7.474</td>
</tr>
<tr>
<td>60</td>
<td>0.218</td>
<td>-0.218</td>
<td>7.939</td>
</tr>
<tr>
<td>80</td>
<td>0.191</td>
<td>-0.191</td>
<td>8.356</td>
</tr>
</tbody>
</table>

To evaluate the effectiveness of the three methods in detecting aberrant responses, we applied them to the generated datasets. Once the examinees were classified into normal or abnormal groups using these methods, we compared the correct classification rates (CCR), type I error, and power rates across replications. CCR, also known as the hit rate, refers to the probability of correctly classifying each examinee based on their behavior into either the normal or aberrant group. In other words, it represents the proportion of individuals who were correctly identified as
either exhibiting normal or aberrant behavior.

In addition to CCR, we also evaluated the type I error and power rates of the three methods. Type I error rate refers to the probability of incorrectly classifying a normal examinee as aberrant, while power rate refers to the probability of correctly identifying an aberrant examinee. These measures were evaluated across 100 replications to ensure the results were robust. To conduct these analyses, we utilized the R statistical software (R Core Team, 2021).
CHAPTER 4

RESULTS

To begin with, we evaluated the type I error rates for all six indices under the condition where all examinees performed normally. This served as our baseline type I error rate for reference purposes. Next, we investigated the performance of the bootstrap CUSUM method, and selected one index from this method that exhibited the highest power. We then compared this index with the traditional CUSUM and CPA indices.

To further investigate the efficiency of the detection power of these methods, we divided the examinees into five groups based on their true abilities. This allowed us to assess the effectiveness of the methods in detecting random guessing behaviors across a range of ability levels. By employing this approach, we were able to obtain a more comprehensive understanding of the strengths and limitations of each of the statistical methods under investigation.

4.1 Baseline Condition without Random Guessing

Initially, we assessed the basic type I error rates of all three detection methods under the baseline condition, where all simulated examinees performed normally. The average type I errors across 100 replications are presented in Table 4.

<table>
<thead>
<tr>
<th>Type 1 error</th>
<th>Test Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>0.050</td>
</tr>
</tbody>
</table>
It can be observed that the average type I errors of the two bootstrap CUSUM indices (S1 and S2), the traditional CUSUM index, and the CPA index are close to the nominal level of 0.05 and are not affected by test lengths. However, the average type I errors of the remaining two bootstrap CUSUM indices (S3 and S4) are larger than the nominal level of 0.05. Furthermore, the S3 index of the bootstrap CUSUM method exhibited an increase in type I errors as the test length increased.

### 4.2 Overall Performance

To thoroughly assess the performance of all three methods, we investigated the type I error rates across varying test lengths and severity rates. Figure 7 provides a visual representation of our findings. It is worth noting that, with the exception of the bootstrap index S3, the type I errors for all other bootstrap indices (the traditional CUSUM index and the CPA index) were found to be unaffected by both test lengths and severity rates. These values were consistently close to the nominal level of 0.05, demonstrating good control over false positives.

We observed that the type I error of the bootstrap index S3 was slightly influenced by severity rates when the test contained 60 items. However, in general, all six indices exhibited type I errors that were nearly identical to the baseline condition. This indicates that the selected indices maintain their desired level of control over false positives, even under varying conditions.
For more detailed information regarding the specific type I error values for each index under different conditions, please refer to Table A1 in Appendix A.

![Figure 7. Average Type I Error of All Indices](image)

Then we investigated power of all six indices across varying test lengths and severity rates. It was observed in Figure 8 that, across all indices, the power increased as the test lengths increased. Additionally, the detection power of index S3 from the bootstrap method, as well as the traditional CUSUM and CPA indices, also exhibited an increase as the proportion of affected items grew. However, the power of the remaining three bootstrap indices (S1, S2, and S4) remained relatively unaffected by severity rates.

Among the four bootstrap indices, index S4 demonstrated the highest detection power. However, it is important to note that all four bootstrap indices had relatively lower detection power when compared to the traditional CUSUM and CPA methods.
For example, in a scenario where a test consisted of 60 items and ten percent of individuals exhibited random guessing behaviors on the last 30% of items (18 items), the power values for indices S1 to S4 were 0.440, 0.439, 0.367, and 0.488, respectively. In comparison, the power for the traditional CUSUM method was 0.643, and the power for the traditional CPA method was 0.612.

For detailed information regarding the specific power values for each index under different conditions, please refer to Table A2 in Appendix A. This table provides a comprehensive breakdown of power values corresponding to various test lengths and severity rates, as depicted in Figure 8.

![Figure 8. Average Detection Power of All Indices](image)

Lastly, we delve into the correct classification rates (CCR) of all methods. As depicted in Figure 9, the CCR of all four indices followed a similar pattern to that of power, increasing as test lengths grew. Similar to the power analysis, index S3 from
the bootstrap method, along with the traditional CUSUM and CPA indices, was found to be influenced by severity rates, resulting in an increased CCR as the proportion of affected items rose. However, the CCR of the remaining three bootstrap indices (S1, S2, and S4) remained relatively unaffected by severity rates.

It is important to note that all three methods exhibited relatively high CCR, even in the case of the bootstrap method, which had lower detection power. This can be attributed to the fixed prevalence rate of 10%, implying that only 10% of the individuals (100 in total) displayed random guessing behaviors. Consequently, the majority of individuals performed normally. Even with such a low prevalence rate, index S3 achieved a CCR greater than 0.85 under all conditions. For detailed information regarding the CCR values of each index under different conditions, please refer to Table A2 in Appendix A.

We found that index S1 demonstrated a higher CCR than index S4 within the bootstrap method. This outcome aligns with our expectations, as the type I error of index S1 was lower than that of index S4. However, it is important to highlight that the traditional CUSUM and CPA methods still outperformed the bootstrap method in terms of CCR.


4.3 **Performance across Theta Levels**

All indices in this study are derived from variations of residuals between examinees' true scores and expected scores. To gain deeper insight into the performance of the bootstrap method, we classified the examinees into five groups based on their true abilities. These groups were defined as follows: group 1 ($\theta < -2$), group 2 ($-2 < \theta < -1$), group 3 ($-1 < \theta < 1$), group 4 ($1 < \theta < 2$), and group 5 ($\theta > 2$). It is worth noting that the distribution of examinees' true abilities was generated based on the standard normal distribution. Within this classification, approximately 68% of individuals fell into group 3, while 27% of individuals were categorized in either group 2 or group 4. The remaining 5% of individuals were distributed between group 1 and group 5.

Since the type I errors of all methods were not influenced by the conditions of random guessing, and the CCR follows a similar pattern to detection power, our
primary focus in this section will be on discussing the power of these methods. Figures that show CCR across theta groups are in Appendix B. Among the indices, it was found that index S4 demonstrated the highest power. Therefore, we will use it as a representative index to present the performance of the bootstrap method.

Figure 10 illustrates the effectiveness of all three methods across five different ability groups. It's important to note that the power was averaged across various test lengths and severity rates. Except for individuals in group 1 with abilities smaller than -2, the detection power of the indices increases as people's abilities improve.

Interestingly, the detection power is lowest when examinees' true abilities fall within the range of -2 to -1 (group 2). For the bootstrap method, the power is only around 0.1 in this group. It appears that individuals in this range exhibit similar probabilities of providing the correct answer whether they answer questions
thoughtfully or simply guess randomly. However, as ability levels increase, the
likelihood of answering questions correctly when responding thoughtfully also
increases. Consequently, all methods become more effective at detecting randomly
guessing behavior.

In groups 4 and 5, consisting of individuals with true abilities greater than 1 or
2, the detection power of all three methods exceeds 0.90. This high ability level makes
the residuals between true and expected scores more noticeable when randomly
guessing, allowing these methods to more easily identify such behavior. Conversely,
individuals in group 1 with abilities smaller than -2 may be more likely to provide
correct responses when randomly guessing compared to answering questions
thoughtfully. As a result, these methods can efficiently detect such behaviors as well.

Figures 11 and 12 illustrate the power of three detection methods across theta
levels in relation to test lengths and severity rates, respectively. In Figure 11, we
observe a similar pattern to Figure 10, where the power of all three methods increases
as the test length increases. On the other hand, Figure 12 also exhibits a similar pattern
to Figure 10, but with a difference in the effect of severity rates on the power of the
methods. Specifically, the power of the CPA and CUSUM methods increases as
severity rates increase, whereas the power of the bootstrap method remains unaffected
by severity rates.
Figure 11. Power of Detection Methods across Ability Levels by Test Length

Figure 12. Power of Detection Methods across Ability Levels by Severity Rates
CHAPTER 5

DISCUSSION

5.1 Summary

Aberrant responses are a common occurrence in tests and can significantly impact the reliability and generalizability of test scores. Given the increasing emphasis on test validity and security in psychological and educational assessments, the detection of aberrant responses has emerged as a crucial topic with both theoretical and practical implications. In recent years, two SPC methods, namely CUSUM and CPA, have been widely utilized to identify aberrant response behaviors in tests. However, both methods have their own limitations. The CUSUM method necessitates manual inspection to locate change points, while CPA requires intensive computational resources.

In response to these limitations, Wayne A. Taylor proposed a new procedure that combines CUSUM and CPA (Taylor, 2000a). This procedure has demonstrated success in diverse fields such as medicine, astronomy, and climate research. In order to classify examinees using this method, there is no need for a separate step to compute boundary or critical values. As opposed to the CPA method, the computation is less complex. Nonetheless, its efficacy in accurately and efficiently detecting aberrant responses in educational tests remains unclear. Consequently, this study employed a simulation design to evaluate the performance of the new bootstrap method.

The study focused on the detection of one specific aberrant behavior, random guessing behaviors, using four traditional CUSUM indices incorporated within the new bootstrap method. A comparison was made between the new method and the
Based on the analysis of the simulation study results, it can be concluded that the new bootstrap method did not outperform the traditional CUSUM and CPA methods in detecting aberrant responses. Specifically, some of the indices in the bootstrap method exhibited type I error rates higher than the nominal level of 0.05, even under the baseline condition. Furthermore, the index with the highest detection power in the bootstrap method still performed much worse than the traditional CUSUM or CPA method.

Similar to previous research (Shao et al., 2016; Sinharay, 2016; Tu et al., 2022; Yu & Cheng, 2019, 2022), we also found when test becomes longer, the detection power will also increase. It is true for all methods. However, unlike traditional CUSUM and CPA method, the severity rates did not affect the detection power for the bootstrap method, at least for the severity rates (30%, 40% and 50%) we investigated in this study.

To gain a deeper understanding of the underlying mechanisms of the detection methods, we further divided the examinees into groups based on their true abilities. As expected, the performance of the detection methods varied based on the disparity between examinees' expected scores and their true scores. When high-ability individuals engaged in random guessing, their probability of answering items correctly significantly decreased, making them more easily detectable by the methods. Similarly, for individuals with very low abilities, random guessing resulted in better performance than expected, which also facilitated detection.

However, for examinees with moderate to lower abilities, the detection of
random guessing behaviors posed more challenges. This is because their probabilities of providing the correct answers were similar, whether they answered questions thoughtfully or simply guessed randomly. Consequently, distinguishing between their genuine responses and random guessing became more difficult for the detection methods.

The findings of this study have important implications for educational measurement and test validity. The new method, although it did not demonstrate higher detection power compared to traditional CUSUM and CPA methods, contributes to the ongoing exploration and refinement of aberrant response detection techniques. By combining the strengths of CUSUM and CPA, the new method offers a unique approach to identifying aberrant responses in educational tests.

Moreover, the results of this study shed light on the limitations of the new method in terms of its lower detection power. It emphasizes the need for researchers and practitioners to carefully consider the specific context and requirements of their testing situations when selecting appropriate methods for detecting aberrant responses. Understanding the strengths and limitations of different methods enables informed decision-making and promotes the use of reliable and valid assessment practices.

5.2 Limitations of Current Study

While our research design revealed that the bootstrap method did not perform as well as the traditional CUSUM and CPA methods, it is important to note that this does not render the method entirely useless. There are several areas where improvements can be made to enhance its performance. The following limitations should be considered:
Limited set of CUSUM indices: In our study, we only incorporated four CUSUM indices into the bootstrap method, all of which are variations based on the relationship between expected scores and true scores. However, when examinees exhibit abnormal behavior across multiple items, it can lead to inaccurate ability estimates and subsequently affect expected scores. To address this limitation, future research can explore the inclusion of additional statistics or indices within the bootstrap method to potentially enhance its performance.

Simplified representation of aberrant response behavior: Aberrant response behavior is a complex phenomenon that cannot be fully captured by any single model. In our study, we focused on one specific mechanism of randomly guessing behaviors represented by the hybrid model. This model assumes that examinees make random choices after a single change point in the test. While this assumption serves as a strong representation, alternative models such as the graduate change models could be considered. Unlike the hybrid models that assume an abrupt change, graduate change models propose a gradual decline in the probability of correct answers, which may better reflect real-life situations. Examining the performance of the bootstrap method under different models, as noted by Yu and Cheng (2022) who observed different performances between traditional CUSUM and CPA methods, could provide further insights into the potential variability of the bootstrap method's performance.

Limited scope of aberrant response behaviors: This study solely focused on detecting randomly guessing behaviors using the bootstrap method. The performance of this method for identifying other types of aberrant behaviors, such as compromise items or warm-up effects, remains unknown. Future investigations should explore the
applicability and effectiveness of the bootstrap method in detecting a broader range of aberrant response behaviors.

Assumption of known item parameters: In our study, we assumed that item parameters were known and utilized true item parameters to estimate examinees’ abilities. Consequently, we did not manipulate severity rates, which represent the percentage of examinees exhibiting random guessing behaviors. While most testing programs calibrate item parameters during the pre-test stage and treat them as known in operational scoring, it is important to acknowledge that the true item parameters are typically unknown in real-world scenarios.

Addressing these limitations would contribute to a more comprehensive understanding of the bootstrap method and its application in detecting aberrant responses. By refining the method’s implementation, considering alternative models, exploring additional aberrant behaviors, and accounting for the uncertainty surrounding item parameters, future research can further enhance the effectiveness and practical utility of the bootstrap method in educational assessment contexts.

5.3 Future Directions

One potential future direction is to address the limitations identified in previous session. By focusing on improving the performance of the bootstrap method, such as by exploring the inclusion of additional indices or statistics and considering alternative models, researchers can enhance its effectiveness in detecting aberrant responses. Additionally, expanding the scope of aberrant behaviors beyond random guessing and accounting for item parameter uncertainty in real-world scenarios will contribute to a more comprehensive and practical application of the bootstrap method in educational
assessment contexts.

This study uses the unidimensional 2PL model to generate dichotomous item responses. Future research can consider using (1) polytomous IRT models, such as graded response model (GRM; Samejima, 1969) to include ordinal responses, or nominal responses model (NRM; Bock, 1972) to include nominal responses; and/or (2) multidimensional IRT models which can account for multiple underlying factors or hierarchy in the factor structure. It can further investigate the performance of the bootstrap method under different situations.

The article only discussed the scenario of a single change point in the response sequence and did not introduce the analysis of multiple change points. In fact, there is still limited research on multiple change point analysis in the field of psychological and educational measurement. However, in reality multiple change point phenomena occur often, where two or more effects may manifest in the same individual's response process during a test. For example, a participant may exhibit a practice effect in the early stages of the test, fatigue effect in the middle stages, and accelerated response in the later stages. Consequently, the individual's ability level may undergo multiple changes, resulting in the presence of multiple change points in the response sequence. In current measurement research and applications, multiple change points may hold greater importance than single change points, making research on multiple change point analysis more significant. Future research should focus on constructing IRT models for multiple change point aberrant responses and developing a series of indicators and methods in this area.

Detecting aberrant response patterns can be a complex task. All those detection
methods, being a statistical approach, provides valuable insights into classifying examinees as either exhibiting or not exhibiting aberrant responses. However, it is important to note that this classification is based on statistical inference and should be considered as a supplementary tool rather than the sole determinant for identifying abnormal responding examinees. To ensure a comprehensive and accurate assessment, it is necessary to consider additional sources of evidence such as response time, examinee self-reported surveys, and teacher evaluations. Shao (2016) adopted CPA based on response time data to detect speededness. The findings indicated that the detection results demonstrated well-controlled Type I error rates and exhibited high efficacy of the method. By incorporating multiple indicators, a more robust verification process can be established to identify and address abnormal respondent patterns.

5.4 Conclusion

In conclusion, while the new method may have lower detection power compared to traditional CUSUM and CPA methods, it still offers valuable insights and potential benefits in educational measurement and practical testing situations. Understanding its implications for test validity and considering its potential applications can inform researchers and practitioners in their efforts to detect aberrant responses and ensure the integrity of educational assessments.
APPENDIX A

PERFORMANCES OF THREE DETECTION METHODS

Average Type I Errors of All Statistics

<table>
<thead>
<tr>
<th>Severity (%)</th>
<th>Statistics</th>
<th>Test Length</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Bootstrap CUSUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.049</td>
<td>0.050</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.081</td>
<td>0.091</td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.059</td>
<td>0.062</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.044</td>
<td>0.048</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.051</td>
<td>0.053</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Bootstrap CUSUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.050</td>
<td>0.051</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.050</td>
<td>0.051</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.081</td>
<td>0.087</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.062</td>
<td>0.062</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.046</td>
<td>0.047</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.052</td>
<td>0.052</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Bootstrap CUSUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.049</td>
<td>0.050</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.048</td>
<td>0.051</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.080</td>
<td>0.089</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.061</td>
<td>0.063</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.045</td>
<td>0.045</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.052</td>
<td>0.053</td>
<td>0.050</td>
<td></td>
</tr>
</tbody>
</table>
Average Power of All Statistics

<table>
<thead>
<tr>
<th>Severity (%)</th>
<th>Statistics</th>
<th>Test Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>30</td>
<td>Bootstrap CUSUM</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.552</td>
</tr>
<tr>
<td>40</td>
<td>Bootstrap CUSUM</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.372</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.361</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.417</td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.568</td>
</tr>
<tr>
<td>50</td>
<td>Bootstrap CUSUM</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.373</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.418</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.401</td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.603</td>
</tr>
</tbody>
</table>
Average CCR of All Statistics

<table>
<thead>
<tr>
<th>Severity (%)</th>
<th>Statistics</th>
<th>Test Length</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>60</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Bootstrap CUSUM</td>
<td>0.890</td>
<td>0.899</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.891</td>
<td>0.899</td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.856</td>
<td>0.854</td>
<td>0.859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.886</td>
<td>0.893</td>
<td>0.901</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.917</td>
<td>0.921</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.909</td>
<td>0.914</td>
<td>0.920</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.910</td>
<td>0.916</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Bootstrap CUSUM</td>
<td>0.895</td>
<td>0.902</td>
<td>0.911</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.892</td>
<td>0.898</td>
<td>0.906</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.863</td>
<td>0.866</td>
<td>0.867</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.886</td>
<td>0.894</td>
<td>0.902</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.922</td>
<td>0.930</td>
<td>0.937</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.912</td>
<td>0.916</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.900</td>
<td>0.903</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Bootstrap CUSUM</td>
<td>0.898</td>
<td>0.907</td>
<td>0.912</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>0.894</td>
<td>0.900</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.870</td>
<td>0.871</td>
<td>0.872</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.885</td>
<td>0.895</td>
<td>0.900</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>0.929</td>
<td>0.938</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUSUM</td>
<td>0.914</td>
<td>0.920</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>0.916</td>
<td>0.920</td>
<td>0.927</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B

AVERAGE CORRECT CLASSIFICATION RATES ACROSS THETA LEVELS

Average CCR of Detection Methods across Ability Levels

![Graph showing the CCR of detection methods across ability levels by test length.]

CCR of Detection Methods across Ability Levels by Test Length

![Graph showing the CCR of detection methods across ability levels for different test lengths.]

61
CCR of Detection Methods across Ability Levels by Severity Rates
REFERENCES


Sinharay, S. (2017c). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403–421.


Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Application of latent trait and latent class models in the social sciences* (pp. 89–98).


