© 2017 Edison M. Choe
ADVANCEMENTS IN TEST SECURITY: PREVENTIVE TEST ASSEMBLY METHODS AND CHANGE-POINT DETECTION OF COMPROMISED ITEMS IN ADAPTIVE TESTING

BY

EDISON M. CHOE

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology in the Graduate College of the University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:
Professor Hua-Hua Chang, Chair
Associate Professor Jinming Zhang
Assistant Professor Hans-Friedrich Köhn
Professor Carolyn J. Anderson
Associate Professor Steven A. Culpepper
Chapter 1: Test security is as longstanding as testing itself. The broad issues are introduced very briefly, then the specific matters pertaining to computerized adaptive testing (CAT) are discussed. The framework of CAT is established in detail, then the vital security concerns of item exposure and test overlap are given a rigorous theoretical treatment. Ultimately, the asymptotic distribution of mean test overlap rate under random item selection is proven.

Chapter 2: The assembly of linear test forms has traditionally been performed manually by test development specialists. However, manual test assembly (MTA) is a labor and time intensive process that generally produces suboptimal forms, which is why the task has increasingly been delegated to computers running automated algorithms. The standard paradigm of automatic test assembly (ATA) is mixed-integer linear programming (MILP), a mathematical optimization technique that allows the specification of desired test characteristics as a system of linear inequalities to be solved computationally. MILP with the conventional branch-and-bound algorithm guarantees an exact solution whenever feasible, but infeasibility and long computational times are two common difficulties, especially with large item pools and complex constraints such as selecting item sets or controlling for test overlap. In order to mitigate these complications, this chapter proposes a modified ATA procedure called common block assembly (CBA), which uses a stratified shadow-test approach to construct item blocks that are subsequently pieced together into full forms. Based on a previously operational item pool with an extended set of constraints, CBA can effortlessly obtain optimal solutions that outperform MTA in terms of overall test quality.

Chapter 3: Whereas multistage testing (MST) typically routes to a pre-assembled module, on-the-fly MST (OMST) adaptively assembles a module at each stage in real-time. Although OMST produces more individualized
forms with finer measurement precision, imposing exposure control and non-statistical constraints remain a challenge. The scripted testing method is introduced as a simply yet effective way to overcome these issues.

Chapter 4: Despite common operationalization, measurement efficiency of CAT should not only be assessed in terms of the number of items administered but also the time it takes to complete the test. To this end, a recent study introduced a novel item selection criterion that maximizes Fisher information per unit of expected response time, which was shown to effectively reduce the average completion time for a fixed-length test with minimal decrease in the accuracy of ability estimation. As this method also resulted in extremely unbalanced exposure of items, however, $a$-stratification with $b$-blocking was recommended as a means for counterbalancing. Although exceptionally effective in this regard, it comes at substantial costs of attenuating the reduction of average testing time, increasing the variance of testing times, and further decreasing estimation accuracy. Therefore, this chapter investigates several alternative methods for item exposure control, of which the most promising is a simple modification of maximizing Fisher information per unit of centered expected response time. The key advantage of the proposed method is the flexibility in choosing a centering value according to a desired distribution of testing times and level of exposure control. Moreover, the centered expected response time can be exponentially weighted to calibrate the degree of measurement precision. The results of extensive simulations, with item pools and examinees that are both simulated and real, demonstrate that optimally chosen centering and weighting values can markedly reduce the mean and variance of both testing times and test overlap, all without much compromise in estimation accuracy.

Chapter 5: Item compromise persists in undermining the integrity of testing, even secure administrations of CAT with sophisticated item exposure controls. In a novel approach to addressing this perennial test security issue, a recent article introduced a sequential procedure for detecting compromised items in which a significant increase in the proportion of correct responses for each item in the pool is statistically monitored after each exposure. In addition to actual responses, response times are valuable information with tremendous potential to reveal items that may have been leaked. Specifically, examinees that have preknowledge of an item would presumably respond more quickly to it than those who do not. Therefore, this chapter proposes
several augmented methods for the detection of compromised items, all involving simultaneous monitoring of changes in both the proportion correct and average response time for each operational item in the pool. Simulation results indicate that the consideration of response times can afford marked improvements over the analysis of responses alone.

Chapter 6: In direct continuation of Chapter 5, three additional methods of item compromise detection are examined: 1) extension of comparing two proportions, including binomial and Fisher’s exact tests; 2) generalized likelihood ratio test (GLRT); 3) nonparametric techniques comparing empirical distribution functions (EDFs), specifically the Kolmogorov-Smirnov (KS) and Kuiper’s tests. According to simulation results, GLRT in particular is demonstrated to be quite capable of detecting compromised items quickly and accurately, even with only a small chance of an examinee having preknowledge.

Chapter 7: Test security is ultimately a matter of test validity. Thus, the body of research in this thesis seeks to protect validity by improving security from a psychometric perspective. Needless to say, much work still remains in advancing the field to better inform practice.
To my lovely fiancé, family, and friends for their love and support.
ACKNOWLEDGEMENTS

First and foremost, I cannot thank Professor Hua-Hua Chang enough for his invaluable direction and support as my adviser throughout my graduate years. Much of this thesis is built upon the foundational work he has established over the course of his illustrious career in this field. I also extend my deepest gratitude to Professor Jinming Zhang for inspiring and guiding the latter part of this thesis. The proposed change-point detection methods for compromised items are directly based on his recent papers that pioneered the original idea. Additionally, I very much appreciate Professors Frieder Köhn, Carolyn Anderson, and Steve Culpepper for generously sharing the burden as committee members. Last but certainly not least, I am greatly indebted to Justin Kern for being my closest colleague and de facto peer mentor during our time here at UIUC - now on to bigger and better things!
# TABLE OF CONTENTS

## CHAPTER 1 INTRODUCTION
1.1 Brief Overview of Test Security ........................................ 1
1.2 CAT Framework .......................................................... 2
1.3 Item Exposure and Test Overlap in CAT ............................ 7
1.4 Table and Figures ......................................................... 23

## CHAPTER 2 COMMON BLOCK METHOD OF ATA USING A
STRATIFIED SHADOW-TEST APPROACH ............................... 25
2.1 Introduction ............................................................... 25
2.2 Method ................................................................. 39
2.3 Results ................................................................. 42
2.4 Discussion ............................................................. 43
2.5 Tables and Figures ..................................................... 44

## CHAPTER 3 THE SCRIPTED TESTING METHOD FOR CAT
AND ON-THE-FLY MST ......................................................... 56
3.1 Introduction ............................................................. 56
3.2 Method ................................................................. 57
3.3 Results ................................................................. 60
3.4 Discussion ............................................................. 60
3.5 Tables and Figure ....................................................... 61

## CHAPTER 4 OPTIMIZING THE USE OF RESPONSE TIMES
FOR ITEM SELECTION IN CAT ............................................... 63
4.1 Introduction ............................................................. 63
4.2 RT Framework .......................................................... 64
4.3 Motivation ............................................................. 66
4.4 General Method ........................................................ 67
4.5 Study 1: Simulated Item Pools and Examinees .................... 71
4.6 Study 2: Real Item Pool and Examinees ........................... 75
4.7 Discussion ............................................................. 76
4.8 Tables and Figures ....................................................... 79
1.1 Brief Overview of Test Security

Test security is as longstanding as testing itself. This is expected given that the primary purpose of testing throughout history has been used to evaluate a person on some ability, with the results often having significant consequences for the test-taker. Perhaps now more than ever, we live in a society where test performance dictates one’s academic trajectory, professional prospects, and even livelihood. In such a high-stakes testing culture, it is hardly surprising that many test-takers would stoop to any means to gain an edge. The common, tried-and-true tactics include copying answers from neighbors, stealing physical test forms, hiring proxies, and sharing test items by word of mouth or braindumps (i.e., online forums or discussion boards). Some less common, technologically advanced tactics include recording exams via spy cameras, consulting outside assistance via earpieces, and hacking into test servers to change grades or steal items. Therefore, it is imperative for test-users and developers to increase vigilance in security.

The subject of test security is very broad, with issues ranging from the design of secure delivery systems and physical security to investigations of misconduct and legal matters (Wollack & Fremer, 2013). This thesis in no way attempts to address all of these issues. In fact, the scope is much narrower, focusing just on the prevention and detection of item compromise from a psychometric perspective, primarily in computerized adaptive testing (CAT). Thus, the next section of this chapter lays the important groundwork for CAT, which will be necessary to fully comprehend Chapters 3 and onward. This first chapter concludes with an in-depth, theoretical exposition of item exposure and test overlap, which are two fundamental measures of how secure a test design is from potential item compromise (Way, 1998). Chapters 2-
focus on prevention by controlling for item exposure and test overlap in linear, CAT, and multistage testing (MST) designs, respectively. Chapters 5-6 focus on detection of item compromise in CAT by implementing real-time statistical monitoring procedures. Chapter 7 briefly discusses the broader validity implications of test security, summarizes the general contributions of this thesis, and concludes with final remarks.

1.2 CAT Framework

1.2.1 Item Response Theory

The primary purpose of CAT is to measure an examinee’s latent trait(s) of interest as efficiently as possible. As such, the core of any CAT system is an adaptive algorithm that strives to select the most appropriate sequence of items for the test-taker. Any such algorithm requires a way to relate the latent trait(s) to the psychometric properties of items, which is principally fulfilled by a class of models within the item response theory (IRT) framework. In particular, the three parameter logistic model (3PLM; Lord & Novick, 1968) is routinely used for applications measuring univariate ability with dichotomous items. It is typically parameterized as

\[ P(X_{ij} = 1|\theta) = P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}, \]

in which \( X_{ij} \) is a binary random variable mapping the \( i \)th examinee’s response to the \( j \)th item as either 1 for correct or 0 for incorrect, and \( \theta \) is the latent ability parameter. Hence, function \( P_j(\theta_i) \) outputs the conditional probability of correctly answering item \( j \) given the examinee’s ability \( \theta_i \), where \( a_j \), \( b_j \), and \( c_j \) represent the item discrimination, difficulty, and pseudo-guessing parameters, respectively. \( \theta_i \) and \( b_j \) are always scaled on the same continuous metric, which grants a direct and meaningful link between the test taker’s ability and the item’s difficulty.
1.2.2 Item Selection and Ability Estimation

Item selection algorithms are commonly based on the Fisher information, which can be derived for a 3PLM item as

\[ I_j(\theta_i) = -E\left(\frac{\partial^2}{\partial \theta_i^2} \log L(\theta_i|x_{ij})\right) = a_j^2 \left(\frac{1 - P_j(\theta_i)}{P_j(\theta_i)}\right) \left(\frac{P_j(\theta_i) - c_j}{1 - c_j}\right)^2. \]  

(1.2)

Note that \( L(\theta_i|x_{ij}) \) is the likelihood function of \( \theta_i \) given an observed response \( x_{ij} \):

\[ L(\theta_i|x_{ij}) = P_j(\theta_i)^{x_{ij}}[1 - P_j(\theta_i)]^{1-x_{ij}}. \]  

(1.3)

The classic maximum Fisher information (MI) method chooses the next item with the largest \( I_j(\hat{\theta}_i) \), where \( \hat{\theta}_i \) is the interim maximum likelihood estimate (MLE) of \( \theta_i \) based on the examinee’s answers to the previous items (Lord, 1980). Specifically, given observed responses to a set of \( k \) items, \( x = \{x_1, x_2, \ldots, x_k\} \), the MLE of \( \theta_i \) is computed as

\[ \hat{\theta}_i^{ML} = \arg \max_{\theta_i} L(\theta_i|x_i) = \arg \max_{\theta_i} \prod_{j=1}^{k} P_j(\theta_i)^{x_{ij}}[1 - P_j(\theta_i)]^{1-x_{ij}}, \]  

(1.4)

and its estimated standard error is the inverse square root of the cumulative Fisher information across the \( k \) items:

\[ SE(\hat{\theta}_i^{ML}) \approx \frac{1}{\sqrt{I(k)(\hat{\theta}_i^{ML})}} = \frac{1}{\sqrt{\sum_{j=1}^{k} I_j(\hat{\theta}_i^{ML})}}. \]  

(1.5)

Technically, \( \hat{\theta}_i^{ML} \) fundamentally assumes local independence of responses (i.e., conditional independence of \( X_{i1}, \ldots, X_{ik} \) given \( \theta_i \)), which is necessarily violated in CAT since the modus operandi of any adaptive algorithm is to select subsequent items based on the previous responses (Mislevy & Chang, 2000). Nevertheless, Chang and Ying (2009) used martingale theory to prove that, under mild regularity conditions with an infinite item pool, \( \hat{\theta}_i^{ML} \) for MI still converges in law as follows:

\[ \hat{\theta}_i^{ML} \overset{c}{\rightarrow} \mathcal{N}\left(\theta_i, \frac{1}{I(k)(\theta_i)}\right) \]  

as \( k \to \infty \)  

(1.6)
(see also Chang, 2015). As a result, the MI method theoretically produces an unbiased estimate and maximizes the measurement precision of $\theta_i$ in the long run when using MLE.

A notable drawback of MLE, however, is that the estimation can be volatile or even infeasible when there is little to no variation in responses, especially early on when only a few items have been answered. In particular, when responses are either all correct or all incorrect, $\theta_i$ is estimated to be $\infty$ or $-\infty$, respectively. Therefore, a Bayes estimator called the expected a posteriori (EAP) is often employed as an alternative, which calculates the expected value of the posterior distribution of $\theta$ given $x$ (over the parameter space $\Theta$) as follows:

$$\hat{\theta}_i^{EAP} = E_{\theta_i} f(\theta_i | x_i) = \int_{\Theta} \theta_i \frac{L(\theta_i | x_i) g(\theta_i)}{\int_{\Theta} L(\theta_i | x_i) g(\theta_i) d\theta_i} d\theta_i = \frac{\int_{\Theta} \theta_i L(\theta_i | x_i) g(\theta_i) d\theta_i}{\int_{\Theta} L(\theta_i | x_i) g(\theta_i) d\theta_i}. \quad (1.7)$$

(Bock & Mislevy, 1982). Note that $g(\theta_i)$ is a prior density function of $\theta_i$, which is usually set as uniform or standard normal in the absence of a more informative prior. Since the above expression is often analytically intractable, it can be numerically approximated as

$$\hat{\theta}_i^{EAP} \approx \frac{\sum \theta_q L(\theta_q | x_i) g(\theta_q)}{\sum Q L(\theta_q | x_i) g(\theta_q)}, \quad (1.8)$$

where $Q$ is a finite set of quadrature nodes ($\theta_q \in Q$) that is representative of $\Theta$. Also, given sufficiently large $k$ and $|Q|$, the standard error of $\hat{\theta}_i^{EAP}$ is well-approximated by the posterior standard deviation of $\theta_i$ (De Ayala, Schafer, & Sava-Bolesta, 1995), which in turn can be numerically approximated by

$$SE(\hat{\theta}_i^{EAP}) \approx PSD(\theta_i) \approx \sqrt{\frac{\sum Q (\theta_q - \hat{\theta}_i^{EAP})^2 L(\theta_q | x_i) g(\theta_q)}{\sum Q L(\theta_q | x_i) g(\theta_q)}}. \quad (1.9)$$

Assuming local independence, Chang and Stout (1993) proved the asymptotic posterior normality of $\theta_i$ under weak regularity conditions. Specifically, with large $k$,

$$f(\theta_i | x_i) \approx N\left(\hat{\theta}_i^{ML}, \frac{1}{I^{(k)}(\hat{\theta}_i^{ML})}\right), \quad (1.10)$$
which is essentially the Bayesian central limit theorem generalized to the case of independent but not identically distributed $X_{i1}, \ldots, X_{ik}|\theta_i$. Thus, it can be inferred that $\hat{\theta}^{EAP}_i \approx \hat{\theta}^{ML}_i$ and $SE(\hat{\theta}^{EAP}_i) \approx I^{(k)}(\hat{\theta}^{ML}_i)^{-1/2}$ for a large number of items, but technically only when local independence holds. In the context of CAT where an examinee’s response sequence is inherently dependent, the asymptotic distribution of EAP has not yet been rigorously established.

Large sample properties of MLE and EAP notwithstanding, tests have finite length in practice, sometimes as short as 10 to 20 scored items per section (e.g., ASVAB subtests). Numerous simulation studies that have investigated small sample behavior of these estimators (e.g., Weiss, 1982; T. Wang & Vispoel, 1998; van der Linden & Pashley, 2010) generally confirm the classic bias-variance tradeoff between maximum likelihood and Bayesian estimation: MLE tends to have lower bias but higher standard error, while EAP tends to have higher bias (towards the prior mean) but lower standard error. Nevertheless, differences are practically negligible for moderate test lengths, or at least 30 items according to T. Wang and Vispoel (1998). In addition, a fairly common practice is to use a combination of MLE and EAP (van der Linden & Pashley, 2010). For example, EAP could act as a provisional fail-safe if an infeasibility occurs with MLE.

1.2.3 Item Exposure Control

Regardless of the choice between estimators, the unrestricted form of MI is highly efficient in terms of ability estimation. Its optimal measurement efficiency, however, comes at the heavy cost of extremely unbalanced item pool usage, as items with large $a$ parameters are disproportionately favored due to their high information (Chang & Ying, 1999; Chang, Qian, & Ying, 2001; Hau & Chang, 2001). In fact, small $a$ items are seldom if ever used, which is clearly an inefficient management of resources. Furthermore, high exposure items are at greater risk of compromise to the detriment of test security. A poignant lesson was learned by ETS back in 1994 when 20 colluders from Kaplan were able to determine that, through collective memorization, the GRE CAT effectively administered only about 200 items despite an undoubtedly larger item pool (ETS and Test Cheating, 2012). To make matters worse, greater imbalance in item exposure directly results in larger test overlap be-
tween examinees (Chen, Ankenmann, & Spray, 2003), which increases the risk of collusion between test takers. Therefore, MI is now almost always restricted with some form of item exposure control in practice. Georgiadou et al. (2007) provides a fairly comprehensive review of various strategies.

Two of the most widely implemented strategies are the classic randomesque (Kingsbury & Zara, 1989) and Sympton-Hetter (SH; Hetter & Sympson, 1997; Sympson & Hetter, 1985) methods. The randomesque method introduces some randomness to the MI process as follows: 1) determine a set of \( k \) items with the largest Fisher information at the interim estimate of \( \theta \); 2) randomly select one of those items to administer. Although simple in concept, the arduous part is choosing an optimal \( k \) for the desired level of trade-off between measurement accuracy and item exposure control, which must be done via CAT simulations. On the other hand, the SH method probabilistically enforces a maximum exposure rate as follows: 1) representing the event of selecting item \( j \) as \( S_j \) and the event of administering item \( j \) as \( A_j \), the probability of administering an item given that it has been selected is \( P(A_j|S_j) = P(A_j \cap S_j)/P(S_j) \); 2) recognizing that an item can only be administered if it has been selected, or \( A_j \subseteq S_j \), \( P(A_j \cap S_j) = P(A_j) \), which is the actual exposure rate of the item; 3) setting the maximum exposure rate at \( r \), or \( P(A_j) = r \), the probability of administering the selected item \( j \) is set to be \( P(A_j|S_j) = r/P(S_j) \). Although effective in theory, a practical limitation of the SH method is that the probability of selecting item \( j \), \( P(S_j) \), can only be estimated through iterated CAT simulations until a stable value is obtained, which may take as many as 100-150 repetitions (van der Linden, 2003). Furthermore, since unselected items cannot be administered, SH is unable to increase exposure for underexposed items (Chang & Ying, 1999).

In light of such complications with restricted MI techniques, a notably distinct alternative is the \( a \)-stratified with \( b \)-blocking design (ASB; Chang et al., 2001), which achieves balance in item pool usage through an innovative item selection procedure. The ASB method first partitions the item bank into several blocks according to the magnitude of \( b \) values, sorts each block according to the magnitude of \( a \) values, then forms new strata by grouping items with the same rank order of \( a \) across the blocks. The rationale behind \( b \)-blocking is to ensure a balanced distribution of difficulties in each stratum for item pools that exhibit a correlation between \( a \) and \( b \), which should be examined by practitioners (Wingersky & Lord, 1984; Chang et al., 2001;
Chang & van der Linden, 2003). Figure 1.1 provides a diminutive illustration of the general process.

Ultimately, the CAT administration is divided into successive stages, proceeding from the stratum with the lowest to highest $a$ values for best results (Hau & Chang, 2001). In essence, high $a$ items have an unduly large influence in ability estimation when uncertainty is still high (Chang & Ying, 1999). Consequently, for short-length tests in particular, advancing from low to high discrimination items has been shown to curtail the underestimation of examinees who make inadvertent mistakes at the beginning (Chang & Ying, 2008). At any given stage, the next item chosen is the one that maximizes the $b$-matching criterion:

$$B_j(\theta_i) = |\theta_i - b_j|^{-1}. \quad (1.11)$$

In other words, the item whose difficulty is closest to the interim ability estimate is selected next from the current stratum. Note that $b$-matching is equivalent to MI for Rasch or 1PLM items (i.e., $a = 1$ and $c = 0$), which is suboptimal for 3PLM items in terms of maximizing information. Nevertheless, by coercing items to be drawn more evenly across the item pool in this way, ASB has been shown to dramatically improve the balance of item exposure with a marginal decrease in the accuracy of $\theta$ estimation.

1.3 Item Exposure and Test Overlap in CAT

In light of the serious security implications of a poorly utilized item pool, a more rigorous understanding of item exposure and test overlap is in order. The following sections methodically present these vital security metrics within a mathematical framework, ultimately building to a proof of the asymptotic distribution of the mean test overlap rate under random item selection.

1.3.1 Definition of Item Exposure Rate

Consider a CAT window consisting of $n$ examinees ($i = 1, \ldots, n$) and a pool of $m$ items ($j = 1, \ldots, m$). For each examinee $i$, define $U_i$ to be an $m$-vector
of binary indicators of whether the $j$th item was administered:

$$U_i = [U_{i1}, \ldots, U_{im}]'. \quad (1.12)$$

In other words, $U_{ij} = 1$ if examinee $i$ receives item $j$ and $U_{ij} = 0$ otherwise. Thus, the test length for every examinee can be expressed as an $n$-vector $L$ consisting of squared lengths of each $U_i$, or more intuitively, the sums of each $U_{ij}$ across all $m$ items (since $U_{ij}^2 = U_{ij}$):

$$L = [||U_1||^2, \ldots, ||U_n||^2]' = \left[L_1 = \sum_{j=1}^{m} U_{1j}, \ldots, L_n = \sum_{j=1}^{m} U_{nj}\right]'. \quad (1.13)$$

Likewise, the exposure count for every item can be expressed as an $m$-vector $V$ consisting of the sums of each $U_{ij}$ across all $n$ examinees:

$$V = \sum_{i=1}^{n} U_i = \left[V_1 = \sum_{i=1}^{n} U_{i1}, \ldots, V_m = \sum_{i=1}^{n} U_{im}\right]' , \quad (1.14)$$

and the total item exposure count $N$ can be computed as

$$N = \sum_{j=1}^{m} V_j = \sum_{i=1}^{n} L_i. \quad (1.15)$$

Consequently, the vector of item exposure rates, which can be interpreted as the mean vector of $U_{ij}$'s across examinees, is given as

$$\overline{U} = V/n = [\overline{U_1} = V_1/n, \ldots, \overline{U_m} = V_m/n]' , \quad (1.16)$$

the sum of which is equal to the average test length across examinees:

$$1' \overline{U} = \sum_{j=1}^{m} \overline{U_j} = \frac{1}{n} \sum_{j=1}^{m} V_j = \frac{1}{n} \sum_{i=1}^{n} L_i = L, \quad (1.17)$$

Note that for fixed-length CAT in which every examinee receives the same number of items $L$ (i.e., $L_i = L \forall i$), $L = L$ and $N = nL$. Table 1.1 summarizes all of this information.
1.3.2 Distributional Properties of Item Exposure Rate

Assuming a random sample of examinees, $U_i$ is independently distributed across $i$ as $m$-dimensional Bernoulli with the probability vector given as $p_i = [P(U_{i1} = 1) = p_{i1}, \ldots, P(U_{im} = 1) = p_{im}]'$:

$$U_i \sim \text{Bernoulli}_m(p_i).$$ (1.18)

With the additional assumption that items are administered completely at random, $p_i = p = [L/m, \ldots, L/m]' \forall i$, meaning $U_i$ is also identically distributed across $i$ as

$$U_i \overset{\text{iid}}{\sim} \text{Bernoulli}_m(p)$$ (1.19)

with the following expectation:

$$\mu = p = \left[\frac{L}{m}, \ldots, \frac{L}{m}\right]' .$$ (1.20)

Furthermore, the variance of any $U_{ij}$ can be derived as

$$\text{Var}(U_{ij}) = E(U_{ij}^2) - E(U_{ij})^2 = \frac{L}{m} - \left(\frac{L}{m}\right)^2 = \frac{L}{m} \left(1 - \frac{L}{m}\right) ,$$ (1.21)

and the covariance between any two items, $j$ and $j'$, can be derived as

$$\text{Cov}(U_{ij}, U_{ij'}) = E(U_{ij}U_{ij'}) - E(U_{ij})E(U_{ij'})$$

$$= \frac{L}{m} \left(\frac{L - 1}{m - 1}\right) - \left(\frac{L}{m}\right)^2$$

$$= \frac{L}{m} \left(\frac{L - m}{m(m - 1)}\right).$$ (1.22)

Therefore, the covariance of $U_i$ is the following $m \times m$ matrix with a compound symmetric (or exchangeable) structure:

$$\Sigma = \begin{bmatrix}
\frac{L}{m} \left(1 - \frac{L}{m}\right) & \frac{L}{m} \left(\frac{L - m}{m(m - 1)}\right) & \cdots \\
\frac{L}{m} \left(\frac{L - m}{m(m - 1)}\right) & \frac{L}{m} \left(1 - \frac{L}{m}\right) & \cdots \\
\vdots & \vdots & \ddots
\end{bmatrix}.$$ (1.23)
and the multivariate central limit theorem assures the following asymptotic normality for $L < m$:

$$Y = \sqrt{n}(\bar{U} - \mu) \xrightarrow{d} \mathcal{N}_m(0, \Sigma).$$  \hfill (1.24)

If $\Sigma$ happened to be invertible (which it is not), then the asymptotic distribution could be expressed as

$$Z = \Sigma^{-1/2}Y \xrightarrow{d} \mathcal{N}_m(0, I),$$  \hfill (1.25)

and $Z'Z$ would converge in distribution to $\chi^2$ with $m$ degrees of freedom:

$$Z'Z \xrightarrow{d} \chi^2(m).$$  \hfill (1.26)

However, the singularity of $\Sigma$ is easily verified by observing that the columns (or rows) sum to a zero vector:

$$\Sigma \mathbf{1} = \left[ \frac{L}{m} \left( 1 - \frac{L}{m} \right) + (m - 1) \frac{L}{m} \left( \frac{L - m}{m(m-1)} \right), \ldots \right] = 0,$$

indicating linear dependence of columns (or rows). Moreover, for any $m \times m$ exchangeable covariance matrix with the diagonal elements equal to $\sigma^2$ and the off-diagonal elements equal to $\delta$, there are two possible patterns of eigenvalues $\lambda_j$:

- if $\delta \geq 0$, then $\lambda_1 = \sigma^2 + (m - 1)\delta$ and $\lambda_2, \ldots, \lambda_m = \sigma^2 - \delta$;
- if $\delta \leq 0$, then $\lambda_1, \ldots, \lambda_{m-1} = \sigma^2 - \delta$ and $\lambda_m = \sigma^2 + (m - 1)\delta$.

(Note that if $\delta = 0$, then the covariance matrix is diagonal and all eigenvalues are just equal to $\sigma^2$.) In the current application, $\delta < 0$ since $L - m < 0$, so the eigenvalues of $\Sigma$ can be determined as follows:

$$\lambda_1, \ldots, \lambda_{m-1} = \frac{L}{m} \left( 1 - \frac{L}{m} \right) - \frac{L}{m} \left( \frac{L - m}{m(m-1)} \right) = \frac{L(m-L)}{m(m-1)} > 0,$$

$$\lambda_m = \frac{L}{m} \left( 1 - \frac{L}{m} \right) + (m - 1) \frac{L}{m} \left( \frac{L - m}{m(m-1)} \right) = 0.$$

Thus, $\Sigma$ is positive semidefinite with a rank of $m - 1$, which is the number of non-zero eigenvalues.
Nevertheless, there exists a modified approach that does not require taking the inverse of $\Sigma$. First, define $\Sigma^*$ to be an $m \times m$, symmetric, and idempotent matrix: as such, $(\Sigma^*)^2 = \Sigma^* = (\Sigma^*)'$. The trace of a square matrix is equal to the sum of its eigenvalues, and every eigenvalue of $\Sigma^*$ is either 1 or 0, so it follows that $r = \text{rank}(\Sigma^*) = \text{trace}(\Sigma^*)$. $\Sigma^*$ is also known as an orthogonal projection matrix, because multiplying it to any $m$-vector projects the vector onto an orthogonal $r$-dimensional subspace. Next, define $Z^*$ as follows:

$$Z^* \xrightarrow{d} \mathcal{N}_m(0, I).$$

(1.30)

Then $\Sigma^*Z^*$, the orthogonal projection of $Z^*$ onto an $r$-dimensional subspace via $\Sigma^*$, has the following distribution:

$$\Sigma^*Z^* \xrightarrow{d} \mathcal{N}_m(0, (\Sigma^*)'\Sigma^*) = \mathcal{N}_m(0, \Sigma^*).$$

(1.31)

Finally, it can be shown that the squared length of $\Sigma^*Z^*$ has an asymptotic $\chi^2$ distribution with $r$ degrees of freedom:

$$||\Sigma^*Z^*||^2 = (\Sigma^*Z^*)'(\Sigma^*Z^*) = (Z^*)'\Sigma^*Z^* \xrightarrow{d} \chi^2(r).$$

(1.32)

To take advantage of this result, $Y$ needs to be transformed in such a way that $\Sigma$ is correspondingly transformed into an orthogonal projection matrix with the same rank:

$$AY \xrightarrow{d} \mathcal{N}_m(0, \Sigma^* = A\Sigma A').$$

(1.33)

To find the appropriate transformation matrix $A$, the key is to make the trace of $\Sigma^*$ equal to the rank of $\Sigma$. Since the diagonal entries of $\Sigma$ are all equal, this can be achieved by setting $\text{rank}(\Sigma) = m - 1$ equal to $\text{trace}(\Sigma)$ multiplied by some constant $a$:

$$m - 1 = a \sum_{j=1}^{m} \frac{L}{m} \left(1 - \frac{L}{m}\right) = a \left[ \frac{m \cdot L}{m} \left(1 - \frac{L}{m}\right) \right].$$

(1.34)

Then solving for $a$:

$$a = \frac{m(m - 1)}{L(m - L)},$$

(1.35)

which happens to be the inverse of the non-zero eigenvalue of $\Sigma$. Hence, $A$
is simply the scalar matrix $\sqrt{a}I$, which can be verified as follows:

$$\mathbf{AY} = (\sqrt{a}I) \mathbf{Y} = \sqrt{a} \mathbf{Y}, \quad (1.36)$$

$$\Sigma' = \mathbf{A} \Sigma' = (\sqrt{a}I) \Sigma (\sqrt{a}I) = \sqrt{a} \Sigma$$

$$= \begin{bmatrix}
1 - \frac{1}{m} & -\frac{1}{m} & \cdots \\
-\frac{1}{m} & 1 - \frac{1}{m} & \cdots \\
\vdots & \vdots & \ddots
\end{bmatrix}
= \mathbf{I} - \mathbf{qq}', \quad (1.37)$$

where $\mathbf{q} = [1/\sqrt{m}, \ldots, 1/\sqrt{m}]'$. $\Sigma'$ is clearly symmetric, its idempotence can be checked as follows:

$$(\Sigma')^2 = (\mathbf{I} - \mathbf{qq}')(\mathbf{I} - \mathbf{qq}')$$

$$= \mathbf{I}^2 - 2\mathbf{Iqq} + \mathbf{qq}'\mathbf{qq}'$$

$$= \mathbf{I} - 2\mathbf{qq}' + \mathbf{qq}' \quad \text{(since } \mathbf{q}'\mathbf{q} = 1) \quad (1.38)$$

$$= \mathbf{I} - \mathbf{qq}'$$

$$= \Sigma',$$

and $\text{rank}(\Sigma') = \text{trace}(\Sigma') = m - 1$. (If $\Sigma$ happened to be invertible, then $\mathbf{A} = \Sigma^{-1/2}$, thereby reducing $\mathbf{AY}$ and $\Sigma'$ back to $\mathbf{Z}$ and $\mathbf{I}$, respectively. Note that $\mathbf{I}$ is technically a projection matrix with full rank of $m$.) Therefore, assuming that items are selected completely at random,

$$(\mathbf{AY})' \mathbf{AY} = \mathbf{a} \mathbf{Y}' \mathbf{Y} = an(\bar{U} - \mu)'(\bar{U} - \mu) \xrightarrow{d} \chi^2(m - 1), \quad (1.39)$$

which can also be expressed as

$$Q = \left(\frac{m - 1}{m - L}\right) \frac{n \sum_{j=1}^{m} (\bar{U}_j - L/m)^2}{L/m} \xrightarrow{d} \chi^2(m - 1), \quad (1.40)$$

or equivalently in terms of counts,

$$Q = \left(\frac{m - 1}{m - L}\right) \frac{\sum_{j=1}^{m} (V_j - nL/m)^2}{nL/m} \xrightarrow{d} \chi^2(m - 1). \quad (1.41)$$
Note that when $L = 1$, $Q$ reduces to the familiar Pearson’s $\chi^2$ test statistic.

Alternatively, define $G$ as follows:

$$G = \frac{m - L}{n(m - 1)} Q = \frac{\sum_{j=1}^{m} (U_j - L/m)^2}{L/m},$$

which is the “$\chi^2$” statistic that was originally proposed by Chang and Ying (1999) as a measure of item pool utilization (i.e., greater the value of $G$, greater the distributional skewness or imbalance of item exposure rates). Moreover, it is known that multiplying a constant $c$ to a $\chi^2$ random variable with $\nu$ degrees of freedom results in a gamma random variable with shape and scale parameters of $k = \nu/2$ and $\theta = 2c$, respectively. In the current context, $c = (m - L)/(n(m - 1))$ and $\nu = m - 1$. Therefore, under the null assumption of completely random item selection, $G$ actually converges to a gamma distribution (as opposed to a $\chi^2$ distribution) as follows:

$$G \xrightarrow{d} \Gamma \left( k = \frac{m - 1}{2}, \theta = \frac{2(m - L)}{n(m - 1)} \right).$$

According to this result, the asymptotic expectation and variance of $G$ are, respectively,

$$\lim_{n \to \infty} E(G_n) = k\theta = \left( \frac{m - 1}{2} \right) \left( \frac{2(m - L)}{n(m - 1)} \right) = \frac{m - L}{n},$$

$$\lim_{n \to \infty} Var(G_n) = k\theta^2 = \left( \frac{m - 1}{2} \right)^2 \left( \frac{2(m - L)}{n(m - 1)} \right)^2 = \frac{2(m - L)^2}{n^2(m - 1)^2}.$$

In actuality, however, the variance of $G$ as specified above does not quite converge to its true value, meaning a slight correction needs to be made to $Q$. This is explained after the following discussion on test overlap.

### 1.3.3 Definition of Test Overlap Rate

Test overlap is defined here as the number of shared items between a pair of examinees $i$ and $i'$, which can be quantified as

$$W_{ii'} = U'_i U'_i = \sum_{j=1}^{m} U_{ij} U_{i'j}.$$
\( W_{ii'} \) is also called the pairwise test overlap to be particular, but pairwise will be assumed unless otherwise stated. Assuming a fixed-length CAT of \( L \) items with completely random item selection, \( W_{ii'} \) has a hypergeometric distribution (Chang & Zhang, 2002; Chen et al., 2003). As such, its probability is determined as

\[
P(W_{ii'} = w) = \frac{\binom{L}{w} \binom{m-L}{L-w}}{\binom{m}{L}},
\]

and, its expected value and variance are, respectively,

\[
E(W_{ii'}) = \frac{L^2}{m},
\]

\[
Var(W_{ii'}) = \frac{L^2(m-L)^2}{m^2(m-1)}.
\]

Next, defining test overlap rate as \( R_{ii'} = W_{ii'}/L \), its expected value and variance are easily derived as

\[
E(R_{ii'}) = \frac{1}{L} E(W_{ii'}) = \frac{L}{m},
\]

\[
Var(R_{ii'}) = \frac{1}{L^2} Var(W_{ii'}) = \frac{(m-L)^2}{m^2(m-1)}.
\]

The average test overlap rate is then literally the sum of \( R_{ii'} \) over all \( \binom{n}{2} \) possible pairs of examinees divided by \( \binom{n}{2} \):

\[
\bar{R} = \frac{\sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} R_{ii'}}{\binom{n}{2}} = \frac{\sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} W_{ii'}}{L \binom{n}{2}}.
\]
Note that the sum of $W_{ii'}$ over all pairs of examinees can also be expressed as

$$
\sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} W_{ii'} = \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} \sum_{j=1}^{m} U_{ij} U_{i'j} \\
= \sum_{j=1}^{m} \left( \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} U_{ij} U_{i'j} \right) \\
= \sum_{j=1}^{m} \binom{V_j}{2}.
$$

(1.53)

This makes sense since the term in the braces represents the number of examinee pairs both having received item $j$, which is equivalently but more efficiently computed as $\binom{V_j}{2}$. Therefore, a computation-friendly formulation of average test overlap rate is

$$
\bar{R} = \frac{\sum_{j=1}^{m} \binom{V_j}{2}}{L \binom{n}{2}} = \frac{n}{L(n-1)} \sum_{j=1}^{m} U_j^2 - \frac{1}{n-1},
$$

(1.54)

which was originally derived by Chen, Ankenmann, and Spray (2003). Moreover, they showed that for a large number of examinees, $\bar{R}$ is a linear function of the sample variance of item exposure rates ($s^2 \{ U \}$):

$$
\lim_{n \to \infty} \bar{R} = \frac{m}{L} s^2 \{ U \} + \frac{L}{m}.
$$

(1.55)

In other words, a higher average test overlap rate is a direct consequence of greater imbalance in item pool utilization.

Generalizing the test overlap construct, define the $a$th test overlap as the number of items that the $i$th examinee has in common with a group of $a$ other examinees (Chang & Zhang, 2002). Also, define $\alpha_a$ as a set of $a$ examinees apart from the $i$th examinee, $\alpha_a = \{i_1, \ldots, i_a\}$. Then a binary vector, in which 1 indicates an item that is common to at least one out of $a$ examinees and 0 indicates an item that is seen by none, can be expressed as

$$
U_{\alpha_a} = 1_{Z^+} \left( \sum_{k=1}^{a} U_{ik} \right) = \left[ 1_{Z^+} \left( \sum_{k=1}^{a} U_{ik1} \right), \ldots, 1_{Z^+} \left( \sum_{k=1}^{a} U_{ikm} \right) \right]'.
$$

(1.56)
where \(1_{\mathbb{Z}^+}\) is an indicator function that converts any positive integer into 1 (and 0 otherwise). Therefore, the \(a\)th test overlap can be quantified as

\[
W_{iaa} = U_i' U_{\alpha a} = \sum_{j=1}^{m} U_{ij}\left[1_{\mathbb{Z}^+}\left(\sum_{k=1}^{a} U_{ikj}\right)\right].
\]  

(1.57)

Note that in the special case of \(a = 1\), \(\alpha_1\) is just a single examinee \(i_1 = i'\), so \(W_{ia1}\) reduces to the pairwise test overlap \(W_{ii'}\). In general, \(W_{iaa}\) belongs to a hypergeometric family that requires a lengthy recursive procedure to determine its probability, so the theoretical distribution proven by Chang and Zhang (2002) will not be reproduced here. On the other hand, defining the \(a\)th test overlap rate as \(R_{iaa} = W_{iaa}/L\), the empirical average of \(R_{iaa}\) is literally the sum of all \(n\binom{n-1}{a}\) possible pairings of each examinee \(i\) with a unique group of \(a\) examinees divided by \(n\binom{n-1}{a}\):

\[
\bar{R}_{\alpha a} = \frac{\sum_{i=1}^{n} \sum_{\alpha_a \in A_i} R_{iaa}}{n \binom{n-1}{a}} = \frac{\sum_{i=1}^{n} \sum_{\alpha_a \in A_i} W_{iaa}}{L n \binom{n-1}{a}},
\]  

(1.58)

where \(A_i\) is the set of all \(\binom{n-1}{a}\) unique sets of \(\alpha_a\) given examinee \(i\). Alternatively, the sum of all \(W_{iaa}\) can be expressed as

\[
\sum_{i=1}^{n} \sum_{\alpha_a \in A_i} W_{iaa} = \sum_{i=1}^{n} \sum_{\alpha_a \in A_i} \sum_{j=1}^{m} U_{ij}\left[1_{\mathbb{Z}^+}\left(\sum_{k=1}^{a} U_{ikj}\right)\right] = \sum_{j=1}^{m} \left\{\sum_{i=1}^{n} \sum_{\alpha_a \in A_i} U_{ij}\left[1_{\mathbb{Z}^+}\left(\sum_{k=1}^{a} U_{ikj}\right)\right]\right\},
\]  

(1.59)

where the term in the braces represents the total number of \(i\) and \(\alpha_a\) pairings in which \(U_{ij} = 1\) and at least one of \(\{U_{i1j}, \ldots, U_{iaj}\}\) equal 1 for a given item \(j\). For example, if \(a = 3\), then there are three kinds of such pairings and their respective counts:

- \(U_{ij} = 1\) and all of \(\{U_{i1j}, U_{i2j}, U_{i3j}\}\) equal 1: \(4 \binom{V_j}{4} \binom{n-V_j}{0}\);
- \(U_{ij} = 1\) and exactly two of \(\{U_{i1j}, U_{i2j}, U_{i3j}\}\) equal 1: \(3 \binom{V_j}{3} \binom{n-V_j}{1}\);
\* $U_{ij} = 1$ and exactly one of \{ $U_{i_1j}, U_{i_2j}, U_{i_3j}$ \} equal 1: $2 \left( \binom{V_j}{2} \right) \left( \frac{n - V_j}{2} \right)$.

Adding all the counts together across all items,

$$
\sum_{i=1}^{n} \sum_{\alpha_3 \in A_i} W_{i_3 \alpha} = \sum_{j=1}^{m} \left[ 4 \binom{V_j}{4} \binom{n - V_j}{0} + 3 \binom{V_j}{3} \binom{n - V_j}{1} + 2 \binom{V_j}{2} \binom{n - V_j}{2} \right],
$$

(1.60)

which can be generalized to any $a$:

$$
\sum_{i=1}^{n} \sum_{\alpha_a \in A_i} W_{i_a \alpha} = \sum_{j=1}^{m} \sum_{k=1}^{a} (a - k + 2) \binom{V_j}{a-k+2} \binom{n - V_j}{k-1}.
$$

(1.61)

Therefore, the average $a$th test overlap rate can be equivalently reformulated as

$$
\bar{R}_{\alpha_a} = \frac{\sum_{j=1}^{m} \sum_{k=1}^{a} (a - k + 2) \binom{V_j}{a-k+2} \binom{n - V_j}{k-1}}{Ln \binom{n - 1}{a}}.
$$

(1.62)

This is a far more efficient calculation than directly computing $W_{i_a \alpha}$ for all $n \binom{n-1}{a}$ possible pairings, which becomes exponentially demanding for increasing $n$ and $a$.

1.3.4 The Asymptotic Distribution of Mean Test Overlap Rate

To establish the distribution of average pairwise test overlap rate, note that $\bar{R}$ and $G$ are both functions of the same quantities: $n$, $m$, $L$, and $U_j$. Thus, it is possible to derive an algebraic relationship between them. First, using the formula for $\bar{R}$, solve for the sum of squared item exposure rates:

$$
\sum_{j=1}^{m} \bar{U}_j^2 = \frac{L(n-1)}{n} \left( \bar{R} + \frac{1}{n-1} \right) = \frac{L(n-1)}{n} \bar{R} + \frac{L}{n}.
$$

(1.63)

Next, expand the squared term of $G$, work out the summations, and substitute the above result:

$$
G = \frac{\sum_{j=1}^{m} (\bar{U}_j - L/m)^2}{L/m}
$$
Therefore, a straightforward linear transformation of $\bar{R}$ with multiplicative and additive constants $m(n - 1)/n$ and $m/n - L$, respectively, is exactly $G$ and should be distributed as such.

Inversely, $\bar{R}$ can be expressed as a linear function of $G$:

$$\bar{R} = \frac{n}{m(n - 1)} \left( G + \frac{m}{n} - L \right).$$

Thus, its asymptotic distribution is also within the gamma family, but the addition of a constant precludes a convenient specification. Nevertheless, the true expectation and variance of $\bar{R}$ can be derived using the known distributional properties of $R_{ii'}$:

$$E(\bar{R}) = \left( \frac{n}{2} \right)^{-1} \sum_{i=1}^{n-1} \sum_{i' = i+1}^{n} E(R_{ii'}) = \left( \frac{n}{2} \right)^{-1} \left( \frac{n}{2} \right) \frac{L}{m} = \frac{L}{m},$$

$$Var(\bar{R}) = Var\left( \left( \frac{n}{2} \right)^{-1} \sum_{i=1}^{n-1} \sum_{i' = i+1}^{n} R_{ii'} \right)$$

$$= \left( \frac{n}{2} \right)^{-2} Var\left( \sum_{i=1}^{n-1} \sum_{i' = i+1}^{n} R_{ii'} \right)$$
\begin{align*}
  &= \left( \binom{n}{2} \right)^{-2} \left[ \sum_{i=1}^{n-1} \sum_{i'='i+1}^{n} \text{Var}(R_{ii'}) + 2 \sum_{\mathcal{R}} \text{Cov}(R_{ii'}, R_{i'i''}) \right] \\
  &= \left( \binom{n}{2} \right)^{-2} \sum_{i=1}^{n-1} \sum_{i'='i+1}^{n} \text{Var}(R_{ii'}) \\
  &= \left( \binom{n}{2} \right)^{-2} \binom{n}{2} \frac{(m - L)^2}{m^2(m - 1)} \\
  &= \frac{2(m - L)^2}{m^2(m - 1)}(n - 1).
\end{align*}

(1.67)

Note that the covariance term, where \( \mathcal{R} \) is the set of all unique pairs of \( R_{ii'} \) (i.e., \( \mathcal{R} = \{\{R_{12}, R_{13}\}, \{R_{12}, R_{14}\}, \ldots, \{R_{n-2,n-1}, R_{n-1,n}\}\} \)) drops out because it is equal to 0. The derivation can be simplified by first recognizing that there are two kinds of covariance pairs: 1) among 3 examinees (i.e., \( \text{Cov}(R_{ii'}, R_{i'i''}) \), \( \text{Cov}(R_{ii'}, R_{i'i''}) \), or \( \text{Cov}(R_{ii'}, R_{i'i''}) \)); 2) among 4 examinees (i.e., \( \text{Cov}(R_{ii'}, R_{i'i''}) \)). In the latter case, the covariance is 0 since \( R_{ii'} \) and \( R_{i'i''} \) are clearly independent. The former case is not as obvious due to an examinee overlap between pairs, so it suffices to derive just this part:

\[
\begin{align*}
  \text{Cov}(R_{ii'}, R_{i'i''}) &= \frac{1}{L^2} \text{Cov}(W_{ii'}, W_{i'i''}) \\
  &= \frac{1}{L^2} \text{Cov} \left( \sum_{j=1}^{m} U_{ij} U_{i'j}, \sum_{j=1}^{m} U_{i'j} U_{i''j} \right) \\
  &= \frac{1}{L^2} \left[ \sum_{j=1}^{m} \text{Cov}(U_{ij} U_{i'j}, U_{i'j} U_{i''j}) + \sum_{j \neq j'} \text{Cov}(U_{ij} U_{i'j}, U_{i'j'} U_{i''j'}) \right] \\
  &= \frac{1}{L^2} \left[ m \text{Cov}(U_{ij} U_{i'j}, U_{i'j} U_{i''j}) + m(m - 1) \text{Cov}(U_{ij} U_{i'j}, U_{i'j'} U_{i''j'}) \right] \\
  &= \frac{1}{L^2} \left[ m \{ E(U_{ij} U_{i'j}^2 U_{i''j}) - E(U_{ij} U_{i'j}) E(U_{i'j} U_{i''j}) \} \\
  &\quad + m(m - 1) \{ E(U_{ij} U_{i'j} U_{i''j}) - E(U_{ij} U_{i'j}) E(U_{i'j} U_{i''j}) \} \right] \\
  &= \frac{1}{L^2} \left[ m \left\{ \left( \frac{L}{m} \right)^3 - \left( \frac{L}{m} \right)^4 \right\} \\
  &\quad + m(m - 1) \left\{ \left( \frac{L}{m} \right)^3 \left( \frac{L - 1}{m - 1} \right) - \left( \frac{L}{m} \right)^4 \right\} \right] \\
  &= 0.
\end{align*}
\]
1.3.5 Corrections for $G$ and $Q$ Statistics

Based on the results above, the true expectation and variance of $G$ are now readily found:

\[
E(G) = \frac{m(n-1)}{n} E(\overline{R}) + \frac{m}{n} - L \\
= \left(\frac{m(n-1)}{n}\right) \left(\frac{L}{m}\right) + \frac{m}{n} - L \quad (1.68)
\]

\[
= \frac{m - L}{n},
\]

\[
Var(G) = \left(\frac{m(n-1)}{n}\right)^2 Var(\overline{R}) \\
= \left(\frac{m(n-1)}{n}\right)^2 \left(\frac{2(m-L)^2}{m^2(m-1)n(n-1)}\right) \quad (1.69)
\]

\[
= \frac{2(m-L)^2(n-1)}{(m-1)n^3}.
\]

Note that the true expected value is the same as previous asymptotic result, but the variance is off by a factor of $(n - 1)/n$:

\[
\lim_{n \to \infty} E(G_n) = E(G), \quad (1.70)
\]

\[
\left(\frac{n - 1}{n}\right) \lim_{n \to \infty} Var(G_n) = Var(G). \quad (1.71)
\]

This correction can be made in the asymptotic distribution of $G$ by taking advantage of the properties of the gamma distribution, namely

\[
E(G) = k\theta = \frac{m - L}{n}, \quad (1.72)
\]

\[
Var(G) = k\theta^2 = \frac{2(m-L)^2(n-1)}{(m-1)n^3}. \quad (1.73)
\]

Solving for $k$ and $\theta$, the corrected asymptotic distribution of $G$ is

\[
G \overset{d}{\rightarrow} \Gamma \left(k = \left(\frac{m-1}{2}\right) \left(\frac{n}{n-1}\right), \theta = \left(\frac{2(m-L)}{n(n-1)}\right) \left(\frac{n-1}{n}\right)\right). \quad (1.74)
\]

The distribution can also be reparameterized as follows:

\[
G \overset{d}{\rightarrow} \Gamma \left(\frac{\nu}{2}, 2c\right), \quad (1.75)
\]
where \( \nu \) and \( c \) are

\[
\nu = (m - 1) \left( \frac{n}{n-1} \right), \quad (1.76)
\]

\[
c = \left( \frac{m - L}{n(m-1)} \right) \left( \frac{n-1}{n} \right). \quad (1.77)
\]

Then it becomes clear that \( Q^* = G/c \) has a chi-square distribution with \( \nu \) degrees of freedom:

\[
Q^* = \frac{G}{c} = \left( \frac{n}{n-1} \right) \left( \frac{m-1}{m-L} \right) \frac{n \sum_{j=1}^{m} (\bar{U}_j - L/m)^2}{L/m} \xrightarrow{d} \chi^2(\nu) \quad (1.78)
\]

or equivalently in terms of exposure counts,

\[
Q^* = \left( \frac{n}{n-1} \right) \left( \frac{m-1}{m-L} \right) \frac{\sum_{j=1}^{m} (V_j - nL/m)^2}{nL/m} \xrightarrow{d} \chi^2(\nu). \quad (1.79)
\]

In other words, the uncorrected \( Q \) has a negative bias which is corrected by multiplying \( n/(n-1) \) to both its value and degrees of freedom. Of course, this correction becomes increasingly negligible as \( n \) increases, since \( \lim_{n \to \infty} \frac{n}{n-1} = 1 \).

### 1.3.6 Asymptotic Normality of \( G \) and \( Q^* \)

Finally, it is worth noting that all of the asymptotic results discussed above have been in terms of the sample size of examinees (i.e., \( n \to \infty \)). As an added bonus, perhaps unsurprisingly, \( Q^* \) and \( G \) also converge to normality as the item pool increases (i.e., \( m \to \infty \)) thanks to CLT. In the case of \( Q^* \), its degrees of freedom \( \nu \) can be interpreted as the number of independent standard normal RV’s that are squared then summed, so

\[
Q^* \xrightarrow{n \to \infty} \chi^2(\nu) \xrightarrow{m \to \infty} \mathcal{N}(\nu, 2\nu). \quad (1.80)
\]

Likewise, in the case of \( G \), its shape parameter \( k \) can be interpreted as the number of independent exponential RV’s with \( \lambda = 1/\theta \) that are summed (as
gamma is a generalization of the Erlang distribution allowing \( k \in \mathbb{R}^+ \), so

\[
G \xrightarrow{n \to \infty} \Gamma(k, \theta) \xrightarrow{m \to \infty} \mathcal{N}(k\theta, k\theta^2).
\] (1.81)

For empirical confirmation, a simulation study was conducted to compare the sampling distributions of \( Q^* \) and \( G \) with their corresponding corrected, uncorrected, and normal densities under completely random item selection. The kernel densities were estimated with 100,000 replications for many combinations of \( n, m, \) and \( L \), but for brevity, the results of only \( G \) for the following four conditions are summarized in Figure 1.2: \( \{n = 5, m = 5, L = 2\}, \{n = 5, m = 50, L = 20\}, \{n = 50, m = 5, L = 2\}, \{n = 50, m = 50, L = 20\} \). As expected, the corrected gamma more accurately represented \( G \), which was particularly noticeable with smaller sample sizes. Also, \( G \) showed convergence to the gamma distribution for increasing \( n \), and it showed convergence to the normal distribution for increasing \( m \). In fact, the asymptotic normality of \( G \) appeared to be quite robust, even with small \( n \).

### 1.3.7 Practical Implications and Conclusion

These results provide a theoretical basis for assessing the potential security risk of a CAT design. In particular, the average pairwise test overlap rate under random item selection represents the best case scenario of a completely balanced utilization of the item pool. Therefore, its asymptotic distribution is the null to which the observed value can be statistically compared. Furthermore, given that most operational CAT programs have hundreds of even thousands of examinees and items, invoking the asymptotic normality of \( G \) or \( Q^* \) is theoretically justified. In other words, the average test overlap rate can be computed, converted to the \( G \) or \( Q^* \) statistic, then conveniently tested against the null using a simple \( z \)-test. Sometimes it is desired to compare the performances between two or more potential item exposure controls. However, the presented results do not extend to the differences or ratios between the \( G \) or \( Q^* \) statistics. Lastly, the asymptotic distribution of the average \( a \)th test overlap rate is yet to be established.
1.4 Table and Figures

Table 1.1: Item exposures, \( U_{ij} = 1 \) or 0, for CAT window with \( i = 1, \ldots, n \) examinees and \( j = 1, \ldots, m \) items. For fixed-length CAT (i.e., \( L_i = L \forall i \)), \( N = nL \) and \( \bar{L} = L \).

<table>
<thead>
<tr>
<th>U_1</th>
<th>U_{11}</th>
<th>\ldots</th>
<th>U_{1j}</th>
<th>\ldots</th>
<th>U_{1m}</th>
<th>L_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>U_i</td>
<td>\vdots</td>
<td>U_{ij}</td>
<td>\ldots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>U_n</td>
<td>U_{n1}</td>
<td>\ldots</td>
<td>U_{nj}</td>
<td>\ldots</td>
<td>U_{nm}</td>
<td>L_m</td>
</tr>
</tbody>
</table>

\[ \sum_{i=1}^{n} U_i = V \quad V_1 \quad \ldots \quad V_j \quad \ldots \quad V_m \quad \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij} = N \]

\[ \bar{U} = V/n \quad \bar{U}_1 \quad \ldots \quad \bar{U}_j \quad \ldots \quad \bar{U}_m \quad N/n = \bar{L} \]

Figure 1.1: An illustration of the \( a \)-stratification with \( b \)-blocking (ASB) process.
Figure 1.2: Comparison of $G$’s empirical distribution (kernel density estimation using 100,000 replications) with corrected gamma, uncorrected gamma, and normal densities.
CHAPTER 2

COMMON BLOCK METHOD OF ATA USING A STRATIFIED SHADOW-TEST APPROACH

2.1 Introduction

2.1.1 Basics of Manual Test Assembly

The general development process for linear testing follows these broad steps:

1. Determine the test blueprint and specifications (e.g., test length, content balance, item types, etc.).

2. Establish an item pool that is potentially appropriate for operation.

3. Assemble the desired number of parallel forms in accordance with requirements.

4. Evaluate each form for psychometric quality and sufficiency of meeting requirements.

5. Modify any deficient forms by swapping items with another form or replacing items with new ones from the item pool.

6. Repeat steps 4 and 5 until all forms are deemed adequate for administration.

Step 1 is typically done by substantive experts in the field with approval from policy makers or some regulatory agency. Step 2 is a huge operation in itself involving item writing and review by content experts as well as pretesting and evaluation by psychometricians. Steps 3-6, which are the focus of this chapter, have traditionally been performed manually by test development specialists. Without any intention of trivializing, the manual test assembly (MTA) process is akin to trying to solve a gigantic puzzle, in which the
individual items are pieces that need to be put together according to a long and complicated set of rules.

One of the most important rules is test quality, which is primarily judged by the test information of form $f$ defined as

$$I_f(\theta) = \sum_{j=1}^{n} I_j(\theta), \quad (2.1)$$

where $\theta$ is the ability level of an examinee and $I_j(\theta)$ is the information function for item $j$ given $\theta$. In other words, test information at $\theta$ for form $f$ is the sum of the information at $\theta$ of all $n$ items on that form. Since the item pool used in study was calibrated using the Rasch (or 1PL) model:

$$P_j(\theta) = \frac{1}{1 + e^{-(\theta-b_j)}}, \quad (2.2)$$

item information was in turn computed as

$$I_j(\theta) = P_j(\theta)[1 - P_j(\theta)]. \quad (2.3)$$

The basic interpretation of test information is that higher the value at a given $\theta$, the more accurate the ability estimates of examinees at that level. This follows directly from the fact that the standard error of measurement is simply the inverse of the square root of the test information as follows:

$$SE(\theta) = \frac{1}{\sqrt{I_f(\theta)}}, \quad (2.4)$$

Consequently, the primary goal is to assemble all forms according to some desired measurement precision at various ability levels, which depends on the purpose of the test. Note that for the Rasch model, item information is maximized when $b_j = \theta$. Thus, for an educational program that seeks to measure examinees equally well across the broad ability spectrum, each form should be composed of items from a wide range of difficulties in equivalent quantities. On the other hand, for a professional licensing or certification organization that seeks to classify examinees as either proficient or not, each form should ideally consist of items with difficulties in the narrow range around the cut-score.
From the perspective of classical test theory (CTT), test quality can also be gauged by the $p$ values of items that make up the form. A $p$ value is the proportion of correct responses to an item, so it is an intuitive measure of item difficulty. However, it is important to note that $p$ values are sample variant, meaning their values are highly dependent on the abilities of the examinee group. Nevertheless, if the quality of examinees is not so drastically different from one administration to the next, then $p$ values are still convenient as a preliminary screening metric for test quality and form consistency. Generally speaking, items with values of $p$ close to 0 or 1 do not provide much useful information in terms of measurement, but they may still provide face validity or serve as fillers in case content specifications cannot be met with better items. Of course, the use of such items on a form should be kept at a minimum. Moreover, it is worthwhile to check that the distributions of $p$ values are similar across all forms to ensure a consistent assortment of item difficulties.

Test security is also another critical component of test assembly. Two issues in particular are form exposure and item exposure, which are related but not equivalent security risks when overlap exists between forms. On one hand, form exposure refers to the number of times a particular form is administered within a given testing window. It can be reduced by administering many different forms evenly, which is preferable when there is a serious concern of cheating by copying from neighboring test takers or if there is potential for the answer key to become compromised. These threats can be mitigated by administering the exam on isolated computer terminals at highly secure testing sites with professional proctors. Also, item order can randomized for every administration, thereby ensuring that no two administered forms are exactly alike. On the other hand, item exposure refers to the number of times a particular item is administered within a testing window. When all administered forms are unique, the item exposure is equal to the form exposure; however, when there is some overlap between forms, the exposure for items that appear on multiple forms is greater than the exposure of any single form. Regardless of how tightly security is controlled, nothing can prevent an examinee from memorizing an item and sharing it with future examinees, so controlling item exposure is a top priority.

Even from the rudimentary treatment of manual test assembly provided above, it should be clear that MTA is a labor and time intensive process
that generally produces suboptimal forms. Particularly for large-scale examination programs, the monumental undertaking of assembling multiple forms that are of high quality and security within the constraints of the test blueprint is typically beyond human capabilities. Thus, the task has increasingly been delegated to computers running automated algorithms. Specifically, recent advancements in test assembly theory and computing technology have enabled the emergence of a burgeoning field in measurement known as automated test assembly (ATA). Using newly developed tools and strategies within the ATA framework, it is entirely feasible to devise a fully automated test assembly design that efficiently accomplishes all objectives.

2.1.2 Basics of ATA Using Linear Programming

The standard paradigm of automatic test assembly (ATA) is mixed-integer linear programming (MILP), a mathematical optimization technique that allows the specification of desired test characteristics as a system of linear inequalities to be solved computationally. For this technique, there are three fundamental components that need to be established: decision variables, constraints, and objective function.

First, the decision variables are defined as

\[ x_{jf} = \begin{cases} 
1 & \text{if item } j \text{ is assigned to form } f \\
0 & \text{otherwise} 
\end{cases} \quad j = 1, \ldots, N; \quad f = 1, \ldots, M \tag{2.5} \]

where \( N \) is the total number of items in the pool and \( M \) is the total number of forms to be assembled. In other words, \( x_{jf} \) is a binary variable that indicates the decision of whether item \( j \) is selected for form \( f \), with 1 for yes and 0 for no.

Second, the constraints refer to the test specifications that need to be met, which are expressed as equations in a manner similar to the objective function. For example, to restrict the test length to be exactly \( n \) items for form \( f \), the constraint can be expressed as

\[ \sum_{j=1}^{N} x_{jf} = n. \tag{2.6} \]
To ensure that item \( j \) is selected no more than once across all \( M \) forms, the constraint can be expressed as

\[
\sum_{f=1}^{M} x_{jf} \leq 1. \tag{2.7}
\]

For a slightly more complex example, suppose the number of items on a certain topic must be limited to between \( l \) and \( u \) items on any given form. Letting \( S \) represent the set of all items relevant to this topic, the constraint is expressed as

\[
l \leq \sum_{j \in S} x_{jf} \leq u. \tag{2.8}\]

Lastly, the objective function to be optimized is stated as

\[
\min y \mid T_\theta - y \leq \sum_{j=1}^{N} I_j(\theta) x_{jf} \leq T_\theta + y, \quad y \in \mathbb{R} \tag{2.9}\]

where \( T_\theta \) is the target test information at a desired \( \theta \) and \( y \) is a real-valued objective variable to be minimized. In other words, the objective function is optimized by minimizing \( y \), which ensures the optimal selection of items from the pool such that the resulting test information for form \( f \) is as close as possible to the target information at \( \theta \). By specifying \( \theta \) to be the cut score, test information at the cut score can be tightly controlled across all assembled forms. As a general guideline, \( T_\theta \) for an \( n \)-item form should be set to \( n \) times the mean item information at the cut score \( \theta_c \) for the entire \( N \)-item pool:

\[
T_\theta = n \bar{I}(\theta_c) = \frac{n}{N} \sum_{j=1}^{N} I_j(\theta_c). \tag{2.10}\]

In this way, all test specifications and the objective function can be specified as a system of linear equations involving binary decision variables and the real-valued objective variable, which can be solved through the use of a computer program (hence the apt name for this technique, where “mixed-integer” refers to the combination of binary and real variables, and “linear” refers to the linearity of the equations). As a simple example, Table 2.1 illustrates a solution for three assembled forms (\( M = 3 \)) using an item pool of
ten items ($N = 10$). The following constraints were imposed:

\[
\sum_{j=1}^{10} x_{jf} = 3, \quad f = 1, 2, 3; \quad (2.11)
\]

\[
\sum_{f=1}^{3} x_{jf} \leq 1, \quad j = 1, \ldots, 10; \quad (2.12)
\]

\[
0 \leq \sum_{j \in S} x_{jf} \leq 1, \quad S = A, B, C, D. \quad (2.13)
\]

Given three-item forms ($n = 3$) and an average item information of 0.20 at a cut score of $\theta_c = 0.50$ (i.e., $I(0.50)$), the target test information was set to $T_{0.50} = 3(0.20) = 0.60$. Therefore, the objective function was given as

\[
\min y \quad | \quad 0.60 - y \leq \sum_{j=1}^{10} I_j(0.50)x_{jf} \leq 0.60 + y. \quad (2.14)
\]

In line with the specified requirements, items 2, 6, and 8 were selected for Form 1; items 1, 5, and 9 were selected for Form 2; and items 4, 7, and 10 were selected for Form 3. Note that every form has exactly three items with at most one item from each topic (A, B, C, D), and each available item is used no more than once across all forms. Furthermore, each form meets the test target information with an error of no more than $y = 0.01$.

The solution in the preceding illustration was engineered for demonstration purposes, although the problem is elementary enough to be feasibly solved by hand. However, test assembly problems in real-life are generally far more complicated with additional constraints, forms, and hundreds or thousands of items in the pool, hence the need for a computer program to find a solution. There are a few linear programming software that are commercially available (e.g., LINGO), but they tend to be prohibitively expensive and may not provide the level of flexibility necessary for more complex constraints. Fortunately, there is a freely available package called lpSolveAPI recently developed and periodically updated for R (Konis, 2014). Without delving too far into the technical details, lpSolveAPI is an application program interface (API) for a powerful linear programming solver called lp_Solve version 5.5. The solver is based on the branch-and-bound algorithm that guarantees an exact solution whenever feasible. At the very least, lpSolveAPI can
perform just as well as any commercial program for free, but the package truly shines for its customizability, which allows for the modeling of virtually any set of test specifications (Diao & van der Linden, 2011). Therefore, lpSolveAPI is the tool of choice for test assembly in proceeding applications.

2.1.3 Building an ATA Design

When building an ATA design, a critical consideration is the issue of feasibility, which refers to whether or not a solution exists for a specified target test information with a particular set of constraints. In other words, although it is possible to model any set of test specifications using lpSolveAPI, there is no guarantee that the program can actually find a solution. There are many factors that may render a model infeasible, but barring any inadvertent misspecifications or errors in programming, two of the most common causes are conflicting constraints and deficiencies in the item pool. Two constraints are said to conflict when it is mathematically impossible to satisfy both conditions, and an item pool is considered to be deficient if it cannot sustain the specified requirements due to a lack of certain items.

Referring back to the example in Table 2.1, suppose the content constraint in Equation 2.13 is changed to

$$\sum_{j \in S} x_{jf} = 1, \quad S = A, B, C, D,$$

or in other words, exactly one item must be chosen from each topic for any given form. Since there are a total of four topics, each form must be composed of exactly four items. However, this new constraint directly conflicts with Equation 2.11, which specifies that every form must have exactly three items. Even if the constraint in Equation 2.11 was removed, Equation 2.15 also conflicts with Equation 2.12, which specifies that an item cannot be used more than once across all three forms. Note that it is impossible to satisfy both constraints since topics C and D each consist of only two items, which reflects a deficiency in the item pool. In order for each of the three forms to consist of a unique item from each of the four topics, the item pool must have at least one additional item for each of the C and D topics. By doing so, however, $T_{0.50} = 0.60$ specified by the objective function in Equation 2.14
is no longer viable since that target value is based on a three-item form. Assuming that the mean item information at $\theta_c = 0.50$ remains at 0.20, the target value for the four-item tests should be increased to $T_{0.50} = 4(0.20) = 0.80$.

Unfortunately, identifying the sources of infeasibility becomes increasingly challenging with each additional constraint imposed on the test design. In reality, with complex designs involving dozens of constraints, the only practical course of action is to methodically search for the causes through trial and error. For instance, problematic constraints can be identified by running the program after each new constraint is added to the model, preferably in order of importance. The most recently added constraint that produces an infeasible solution needs to be modified in some way, either by relaxing the bounds or removing the constraint altogether. Additionally, the imposed target test information may be either too high or too low and thereby impossible to achieve under the specified constraints. Using the value obtained from Equation 2.10 as a starting point, $T_\theta$ can be changed incrementally until a feasible solution is found.

Another critical issue for building an ATA design is computing time, which refers to how long the program takes to find a solution to the model. Needless to say, computer hardware is a major factor, with faster processors and larger memories leading to quicker solutions. Nevertheless, the current discussion will not focus on hardware specifications, as upgrading hardware may not be practicable. Besides, performance for the task at hand should be roughly comparable across standard machines that are reasonably up-to-date. The more pertinent and controllable factors to computing speed involve various strategies of test assembly. That is to say, there are multiple ways of assembling forms or meeting constraints that will ultimately produce the same desired outcome, but certain methods are less computationally intensive and thus faster than others, often in the order of hours or even days. Perhaps as expected, however, there are particular drawbacks to such faster methods that need to be taken into account as well.

First of all, there are two basic strategies for assembling multiple forms: sequential and simultaneous. In sequential assembly, forms are assembled one at a time until the desired number of forms is reached. A simple example is illustrated in Figure 2.1a, where three unique forms of $n$ items each are assembled sequentially starting with an item pool of $N$ items. To assemble
the first form, there are $N + 1$ variables $(x_1, \ldots, x_N, y)$ that need to be solved. For each subsequent form, the number of available items in the pool decreases by $n$, so there are $N + 1 - n$ and $N + 1 - 2n$ variables that need to be solved to assemble the second and third forms, respectively. In this way, a purely sequential method produces the quickest solution in comparison to other methods by minimizing the number of variables on a single run of the program. However, since forms are assembled one by one, there is generally no straightforward way to specify constraints that apply across forms, in particular controlling for item exposure and test overlap. Additionally, since the solution is always optimized for each form, the “best” items are always used up first. Note that the term “best” here is used more broadly to describe items that the solver deems to be optimal given the specified constraints. Consequently, each successive form is less optimal than the previous one as the pool of available items becomes progressively deficient, which may eventually lead to an infeasible solution for a later form.

In simultaneous assembly, on the other hand, forms are assembled all at once in a single run. A simple example is illustrated in Figure 2.1b, where three unique forms of $n$ items each are assembled simultaneously with an item pool of $N$ items. In this case, the program needs to concurrently solve a total of $3N + 1$ variables $(x_{11}, \ldots, x_{N1}, x_{12}, \ldots, x_{N2}, x_{13}, \ldots, x_{N3}, y)$, which turns out to be exponentially slower than ultimately solving for $(N + 1) + (N + 1 - n) + (N + 1 - 2n) = 3(N - n + 1)$ variables using the sequential method. On the upside, simultaneous assembly allows for straightforward specifications of across-form constraints, and the issue of “best” items being used up on earlier forms is no longer pertinent. Nevertheless, the threat of infeasibility still looms large as the item pool may not be able to support the assembly of more than a certain number of forms at once.

In consideration of the pros and cons of both sequential and simultaneous assembly strategies, van der Linden (2005) introduced a hybrid method called shadow-test assembly. This technique is best explained with an example as in Figure 2.2, which depicts the assembly process of four unique forms in three steps:

1. Starting with an item pool of $N$ items, the first form of $n$ items is assembled along with a so-called shadow form of $3n$ items, which is basically a consolidation of the three subsequent forms that are yet
to be assembled. The fundamental idea is that the “best” items in the pool are evenly allocated between Form 1 and the shadow form, thus preventing Form 1 from monopolizing them right at the start (as would be the case in a purely sequential process). Hence, the shadow form acts as a temporary repository for these optimal items, which are promptly returned to the item pool after Form 1 is completed. This ensures that the item pool can support the optimal assembly of the next three forms, thereby reducing the chances of an infeasible solution later.

2. Next, the same process is repeated with an item pool of now $N - n$ items, except the shadow form is now composed of $2n$ items since there are two subsequent forms left to assemble.

3. Finally, with the remaining pool of $N - 2n$ items, the last two forms are assembled simultaneously. Form 4 can be thought of as the shadow form in this step, but there is no logical reason to return these items to the item pool and repeat the process once more.

Therefore, the shadow-test method can be viewed as a clever modification of the sequential method that counters the progressive depletion of optimal items through the simultaneous assembly of a shadow form for each real form. The major advantage of this strategy is that it is drastically faster than the purely simultaneous method with little, if any, sacrifice to the quality of forms assembled. However, the disadvantage of not being able to set across-form constraints still remains.

Besides assembly strategies for multiple forms, the matter of sequential v. simultaneous assembly is a crucial consideration when both stimulus and discrete items are involved. Stimulus items refer to a set of items that must appear together on a form (e.g., items with shared cases or visuals), and discrete items refer to single items that can appear alone on a form. Perhaps the more direct approach to assembling a form is selecting stimulus and discrete items simultaneously, in which case if a stimulus item is selected, all associated items in the set must be selected as well. This can be specified as
a constraint with an additional set of decision variables as follows:

$$z_{sf} = \begin{cases} 1 & \text{if stimulus } s \text{ is assigned to form } f \\ 0 & \text{otherwise} \end{cases}$$

\[ s = 1, \ldots, L; \quad f = 1, \ldots, M, \]  

(2.16)

where \( L \) is the total number of stimulus sets in the item pool, and

$$\sum_{j \in R_s} x_{jf} - |R_s|z_{sf} = 0, \quad (2.17)$$

where \( R_s \) is the set of all items in stimulus \( s \) and \(|R_s|\) is the number of items in that set. For example, suppose items 2, 4, and 9 belong to the first stimulus set \((s = 1, R_1 = \{2, 4, 9\}, |R_1| = 3)\). For the first form \((f = 1)\), the constraint would be expressed as

$$x_{21} + x_{41} + x_{91} - 3z_{11} = 0. \quad (2.18)$$

Consequently, if the first stimulus set is chosen \((z_{11} = 1)\), all the items in that set must be chosen as well \((x_{21} = 1, x_{41} = 1, x_{91} = 1)\) in order to satisfy the above equation; on the contrary, if the first stimulus set is ultimately not chosen \((z_{11} = 0)\), all the items in that set must not be chosen either \((x_{21} = 0, x_{41} = 0, x_{91} = 0)\) in order to satisfy the equation. The converse of these statements is true as well. Although straightforward to implement, this simultaneous method of assembly requires an additional \(LM\) decision variables, making it considerably slower than an otherwise identical design with just discrete items.

Alternatively, a sequential strategy can be employed, in which stimulus and discrete items are selected in two stages. An illustration of such a procedure is shown in Figure 2.3, where three forms are assembled simultaneously, each with \(l\) stimulus sets and a total of \(n\) items. In the first stage, \(l\) stimulus sets are selected for each form out of a total of \(L\) sets in the item pool. This is accomplished by treating stimulus sets as single entities with their own objective function and set of constraints. For example, analogous to Equations 2.9 - 2.13, the specifications can be given as

$$\min y^* \quad | T_\theta^* - y^* \leq \sum_{s=1}^{L} I_s^\theta(\theta)z_{sf} \leq T_\theta^* + y^*, \quad y^* \in \mathbb{R}; \quad (2.19)$$
\[ T_\theta^* = I^*(\theta_c) = \sum_{s=1}^{L} I_s^*(\theta_c); \]  

\[ \sum_{s=1}^{L} z_{sf} = l, \quad f = 1, 2, 3; \]  

\[ \sum_{f=1}^{3} z_{sf} \leq 1, \quad s = 1, \ldots, L; \]  

\[ 0 \leq \sum_{s \in S} z_{sf} \leq 1, \quad S = A, B, C, D, \]  

where \( y^* \) is the objective variable to be minimized, \( T_\theta^* \) is the target test information at \( \theta \) with only stimulus items, \( I_s^*(\theta) \) is the combined information of all items in stimulus set \( s \), and \( I^*(\theta_c) \) is the average set information across the cut-score all stimulus sets. In the second stage, \( k_f \) discrete items are selected for each form to make complete \( n \)-item forms. Note that the number of necessary discrete items may vary for each form, because there may be differing number of items across stimulus sets. However, \( k_f \) does not have to be specified directly; instead, the forms can be re-assembled from scratch with the following two constraints:

\[ \sum_{j \in Q} x_{jf} = |Q|; \]  

\[ \sum_{j \in \{R_1, \ldots, R_L\} \setminus Q} x_{jf} = 0, \]  

where \( Q \) is the set of all stimulus items selected in Stage 1, \( |Q| \) is the total number of items in \( Q \), and \( \{R_1, \ldots, R_L\} \setminus Q \) is the set of all stimulus items not selected in Stage 1. Additionally, the standard test specifications can be given in an equivalent manner as Equations 2.9 - 2.13. Ultimately, by splitting the assembly process into Stage 1 with \( LM + 1 \) variables and Stage 2 with \( NM + 1 \) variables, the sequential method is substantially quicker than the simultaneous method that requires a solution for \( M(L+N) + 1 \) variables all at once.

Lastly, the issue of allowing for overlap between forms deserves special attention. As previously mentioned, test overlap can be specified as an across-form constraint, in which case a purely simultaneous assembly method is unavoidable. A basic circular design is illustrated in Figure 2.4, in which
there is an overlap of \( v \) items between all adjacent pairs of forms (first and last forms are considered to be adjacent as well) and no overlap between non-adjacent pairs of forms. In this scenario, similarly to dealing with stimulus and discrete items simultaneously, a new set of decision variables need to be defined as follows:

\[
w_{jff'} = \begin{cases} 
1 & \text{if item } j \text{ overlaps between forms } f \text{ and } f' \\
0 & \text{otherwise},
\end{cases}
\]  \hspace{1cm} (2.26)

where \( j = 1, \ldots, N \), \( f = 1, \ldots, (M - 1) \), \( f' = 2, \ldots, M \), and \( f < f' \). In other words, an extra set of \( N \) variables are required for every pair of forms to be assembled, which happens to be an additional \( N \binom{M}{2} \) variables. Note that \( \binom{M}{2} \) is the number of all possible pairs of forms. Using these new decision variables, the number of overlapping items between all pairs of forms is specified by

\[
\sum_{j=1}^{N} w_{jff'} = \begin{cases} 
v & \text{if } \{f, f'\} \in \{\{f, f + 1\}, \{1, M\}\} \\
0 & \text{otherwise},
\end{cases}
\]  \hspace{1cm} (2.27)

and controlled by

\[
x_{jf} + x_{jf'} - 2w_{jff'} \geq 0; \hspace{1cm} (2.28)
\]
\[
x_{jf} + x_{jf'} - w_{jff'} \leq 1. \hspace{1cm} (2.29)
\]

Equation 2.28 ensures that if \( w_{jff'} = 1 \), then \( x_{jf} = 1 \) and \( x_{jf'} = 1 \). However, the converse of that statement does not necessarily hold: if \( x_{jf} = 1 \) and \( x_{jf'} = 1 \), then \( w_{jff'} \) can be either 0 or 1. This logical complication is remedied by Equation 2.29, which ensures that if \( x_{jf} = 1 \) and \( x_{jf'} = 1 \), then \( w_{jff'} = 1 \). Furthermore, both equations ensure that when \( w_{jff'} = 0 \), item \( j \) can appear at most once in either form. In summary, Equations 2.28 and 2.29 conjunctively enforce the following set of logical statements:

\[
x_{jf} = 1 \land x_{jf'} = 1 \iff w_{jff'} = 1;
\]
\[
x_{jf} = 0 \land x_{jf'} = 0 \implies w_{jff'} = 0;
\]
\[
x_{jf} = 1 \land x_{jf'} = 0 \implies w_{jff'} = 0;
\]
\[
x_{jf} = 0 \land x_{jf'} = 1 \implies w_{jff'} = 0.
\]  \hspace{1cm} (2.30)
Therefore, it is relatively straightforward to control for overlap using this strategy, with the ultimate benefit of reducing the total number of items required to assemble a given number of forms. Compared to four unique \(n\)-item forms that add up to \(4n\) items, the four overlapping \(n\)-item forms in Figure 2.4 add up to just \(4(n - v)\) items, which is a savings of \(4v\) items. However, the major limitation to this method is that \(N\left(\frac{M}{2}\right)\) extra decision variables \((w_{jff})\) are necessary in addition to the original \(NM + 1\) variables \((x_{jf}, y)\) for a grand total of \(N \left[\binom{M}{2} + M\right] + 1\) variables that need to be solved. Going back to the example of assembling \(M = 4\) forms in Figure 2.4, note that there are a total of \(10N + 1\) variables. Needless to say, the number of variables increases rapidly with each additional form, thus quickly rendering the model utterly unmanageable even for a high-performance machine.

Fortunately, there is a way to control for overlap without the need to introduce a new set of decision variables and constraints as given in Equations 2.26 - 2.29. This can be accomplished by implementing a block design such as the one shown in Figure 2.5. In this particular example, four blocks of \(n - 2v\) items each and four blocks of \(v\) items each are first assembled from the item pool. There are no overlapping items between any of these eight blocks. These blocks are then combined together post hoc in the manner illustrated to create four forms of \(n\) items each. Specifically, note that each form is composed of a unique block of \(n - 2v\) items, a block of \(v\) items that is shared with one form, and another block of \(v\) items that is shared with a different form. Moreover, note that item exposure is also strictly controlled, with a total of \(4(n - 2v)\) variables appearing only once and a total of \(4v\) items appearing exactly twice across all four forms. Therefore, the block strategy ultimately produces forms that are structurally equivalent to the ones produced by the previous technique in Figure 2.4. However, the degree of savings in computing time depends on the method used to assemble the blocks. Using the most computationally intensive method of assembling all eight blocks simultaneously, a total of \(8N + 1\) variables need to be solved, which is still \(2N\) less than the previous method. Utilizing the less intensive sequential or shadow-test methods to assemble the blocks would provide even faster results.
2.2 Method

In full consideration of the issues regarding the construction of a functional ATA design, a new procedure called common block assembly (CBA) is proposed as a practical method for successfully accomplishing the entire list of mission objectives. It can be understood as a parsimonious version of block assembly in which no block is unique to a single form, which is to say all blocks are common to at least two forms. In other words, no item can appear on only one form. A rudimentary example is given in Figure 2.6, where four \( n \)-item forms are assembled with four common blocks of \( n/2 \) items each (for a total of \( 2n \) items), which is only half as many blocks to assemble just as many forms as the block design in Figure 2.5. Specifically, each form is composed of two common blocks and each common block is shared by two forms, resulting in an overlap rate of 50\% \((n/2 \text{ items out of } n)\) between four pairs of forms (1 & 2, 2 & 3, 3 & 4, 4 & 1) and no overlap between the other two pairs of forms (1 & 3, 2 & 4). Also, the exposure rate for any of the \( 2n \) items is strictly controlled at 50\% (2 out of 4 forms).

Implementing the technique on a full scale, two designs are proposed as shown in Figures 2.7 and 2.8. On one hand, CBA Design 1 in Figure 2.7 illustrates an assembly of 12 forms using 14 common blocks, divided into 6 A blocks of \( v_1 \) items each, 4 B blocks of \( v_2 \) items each, and 4 C blocks of \( v_3 \) items each, for a total of \( 6v_1 + 4v_2 + 4v_3 \) items drawn from the item pool. As listed in Table 2.2, each form is composed of 2 A blocks, 1 B block, and 1 C block for a total of \( 2v_1 + v_2 + v_3 = n \) items. Any two forms share no more than one block, thereby controlling the overlap at 0, \( v_1 \), \( v_2 \), or \( v_3 \) items as shown in Table 2.3. Note that for any given form, there exists one other form with which it shares no items, meaning there is always one unique form. Also, each A block is assigned to 4 forms and each B or C block is assigned to 3 forms, thus limiting item exposure rate to either 33\% (4/12) or 25\% (3/12) assuming that the forms are administered uniformly throughout the testing cycle.

On the other hand, CBA Design 2 in Figure 2.8 illustrates an assembly of 16 forms using 16 common blocks, divided into 4 A blocks of \( v_1 \) items each, 4 B blocks of \( v_2 \) items each, 4 C blocks of \( v_3 \) items each, and 4 D blocks of \( v_4 \) items each, for a total of \( 4(v_1 + v_2 + v_3 + v_4) = 4n \) items drawn from the item pool. As listed in Table 2.4, each form is composed of 1 A block, 1 B
block, 1 C block, and 1 D block for a total of $v_1 + v_2 + v_3 + v_4 = n$ items. Any two forms share no more than one block, thereby controlling the overlap at 0, $v_1$, $v_2$, $v_3$, or $v_4$ items as shown in Table 2.5. Note that for any given form, there exists 3 other forms with which it shares no items. Also, any given block is assigned to 4 forms, thus limiting item exposure rate to 25% (4/16) assuming that the forms are administered uniformly throughout the testing cycle.

The functionality of these two designs was evaluated using an operational item pool consisting of over 2000 items from a large-scale licensure examination. Ultimately, the following test specifications were strictly met for all assembled forms of a design:

- $n = 350$ items per form on one run and $n = 300$ items per form on another run;
- Test information of at least $n$ times the mean item information at the cut score of $\theta_c = 0.55$;
- Content and item type coverage per form as shown in Table 2.6;
- Exactly 10% for $n = 350$ or 12% for $n = 300$ of “bad” items per form with $p$ values less than 0.25 or greater than 0.85;
- No enemy items appearing on same form;
- Difference in average test duration within 1 minute among all forms.

As detailed shortly, all of the above requirements were specified using a series of objective functions and constraints with the exception of the restriction on “bad” items. For this particular specification, the item pool was segregated into two groups according to $p$ values: “good” items with $0.25 \leq p \leq 0.85$ and “bad” items with $p < 0.25$ or $p > 0.85$. For Design 1, the A and B blocks were equally drawn from the “good” items while the C blocks were drawn from the “bad” items. For Design 2, the A, B, and C blocks were equally drawn from the “good” items while the D blocks were drawn from the “bad” items. The block characteristics for both designs are summarized in Table 2.7. The idea was that instead of removing the “bad” items altogether, their usage was limited to 10% on 350-item forms (35 items) and 12% on 300-item forms (36 items) for feasibility reasons. The remaining
90% of a 350-item form (315 items) and 88% of a 300-item form (264 items) were composed of “good” items. The specific form compositions for the four scenarios were as follows:

- Design 1 with \( n = 350 \): each of the 12 forms composed of 2 A blocks and 1 B block with 105 “good” items each (for a total of 315 “good” items) and 1 C block with 35 “bad” items;

- Design 1 with \( n = 300 \): each of the 12 forms composed of 2 A blocks and 1 B block with 88 “good” items each (for a total of 264 “good” items) and 1 C block with 36 “bad” items;

- Design 2 with \( n = 350 \): each of the 16 forms composed of 1 A block, 1 B block, and 1 C block with 105 “good” items each (for a total of 315 “good” items) and 1 C block with 35 “bad” items;

- Design 2 with \( n = 300 \): each of the 16 forms composed of 1 A block, 1 B block, and 1 C block with 88 “good” items each (for a total of 264 “good” items) and 1 C block with 36 “bad” items.

Given a sizeable item pool, the presence of both stimulus and discrete items in the item pool, and the need for a fairly large number of common blocks, the shadow-test method was implemented in conjunction with the sequential stimulus-discrete technique for assembling the blocks. In other words, in what could be called a stratified shadow-test approach, each real block was assembled simultaneously with a shadow block, both of which were filled with stimulus items in the first stage followed by discrete items in the second stage.

For the sake of brevity, the assembly strategy is described just for Design 2 with \( n = 300 \). The first goal was to assemble 12 blocks of 88 “good” items each (\( A_1, \ldots, C_4 \)), which was achieved through 11 iterations of the shadow-test method using the “good” item pool. On each of the iterations, the main block of 88 items was assembled simultaneously with a shadow block representing a composite of all subsequent blocks yet to be assembled with “good” items. To be specific, on the first iteration for assembling block \( A_1 \), the shadow block consisted of \( 88 \times 11 = 968 \) items since there were 11 blocks left to assemble. On the second iteration for assembling block \( A_2 \), the shadow block consisted of \( 88 \times 10 = 880 \) items since there were 10
blocks left to assemble. This pattern continued until the 11th iteration for assembling block C3, in which the shadow block consisted of 88 items as well and was thus designated as block C4 without the need for an extra iteration. Likewise, the second goal was to assemble the last 4 blocks of 36 “bad” items each (D1, …, D4), which was achieved through three subsequent iterations using the “bad” item pool. On each of the iterations, the main block of 36 items was assembled simultaneously with a shadow block representing a composite of all subsequent blocks yet to be assembled with “bad” items. To be specific, on the 12th iteration for assembling block D1, the shadow block consisted of $36 \times 3 = 108$ items since there were 3 blocks left to assemble. Continuing in this fashion until the 14th and final iteration for assembling block D3, the shadow block consisting of 38 items was designated as the final block D4 without the need for a 15th iteration.

Once the assembly of all 16 blocks was complete, they were carefully combined into full 300-item forms according to the Common Block Design 2 architecture laid out in Figure 2.8 and Table 2.4. The assembly procedure for the other scenarios, including Design 2 with 350-item forms, Design 1 with 300-item forms, and Design 1 with 350-item forms, is nearly identical to the process detailed above save for a few changes in values.

2.3 Results

Table 2.8 summarizes the most relevant features and performance results of the different designs considered in this study. As a baseline comparison of test quality, a summary of test information obtained from an operational set of manually assembled test forms is also provided. The crucial thing to note is that, in terms of test quality, the 350-item forms produced by either CB design vastly outperformed the 350-item forms produced by MTA. In particular, the mean test information at the cut score was over 10 points higher for the CB forms, and the test information was much tighter across forms. In fact, even the 300-item CB forms performed better than the 350-item MTA forms with slightly higher mean information at the cut score and much less variability overall. These comparisons are shown to great effect by the TICs in Figure 2.9, particularly around the cut-point.

Comparing the 300-item CB forms to their 350-item counterparts, there
was certainly a significant drop in test information by having 50 less items. The practical consequence of misclassification is shown in Table 2.9, where the rates were calculated based on a simulation of 200 examinees per form with 10 repetitions. The results revealed only marginal differences between the 350-item and 300-item CB forms in terms of classification accuracy. The most pronounced difference was between the MTA and CB forms on the rates of unqualified examinees passing, with the MTA forms having noticeably higher percentages.

Finally, there are two more striking observations to be made from Table 2.8. For one, the test duration was controlled extremely well by all of the CB designs, with the longest and shortest forms being less than one minute apart on average. This is quite an achievement considering the length of the exam. Lastly, the solution time shows how long the program took to find a solution from start to finish. Computing times are clearly dependent upon the quality of the machine, but it is still remarkable nonetheless that solutions to such difficult assembly problems can be found in a matter of seconds, especially compared to weeks or even months with MTA.

2.4 Discussion

In full consideration of everything that has been discussed, there is little doubt that ATA is far more efficient and precise than MTA. In particular, CBA has been demonstrated to be a promising ATA procedure when exact solutions are desired but unattainable through established methods due to the sheer size of the problem. This is especially the case when test overlap needs to be strictly controlled for test security. In a nutshell, CBA deconstructs the problem into manageable parts (i.e., blocks), solves them through shadow-testing, then puts the solved pieces together to form the whole solution without any loss in accuracy. As a possible next step, the performance of the stratified shadow-test approach with MILP in this context can be compared to other established ATA paradigms, such as heuristic (e.g., Swanson & Stocking, 1993; Stocking, Swanson, & Pearlman, 1993) or sampling (e.g., Belov, 2016) schemes. In particular, it would be an interesting proposition to implement and assess the utility of the general common block architecture with any of these alternative ATA methods.
2.5 Tables and Figures

Table 2.1: Example solution of exactly three items per form, no more than one item from each topic (A, B, C, D) per form, and no item appearing more than once across all forms with a target test information of 0.60 at $\theta = 0.50$.

<table>
<thead>
<tr>
<th>$j$</th>
<th>Topic</th>
<th>$I_j(0.50)$</th>
<th>$x_{j1}$</th>
<th>$x_{j2}$</th>
<th>$x_{j3}$</th>
<th>$\sum_{f=1}^{3} x_{jf}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0.20</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>0.21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>0.22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>0.20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>0.17</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>0.19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>D</td>
<td>0.19</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$\sum_{i=1}^{10} x_{jf} = 3$  $\sum_{j=1}^{3} \sum_{f=1}^{10} x_{jf} = 9$

$\sum_{j=1}^{N} I_j(0.50)x_{jf} = 0.60$  $T_{0.50} = 0.60$

$y = 0.01$
Table 2.2: Block composition of each of the 12 forms in CBA Design 1.

<table>
<thead>
<tr>
<th>Form</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>A4</td>
<td>B1</td>
<td>C1</td>
</tr>
<tr>
<td>2</td>
<td>A1</td>
<td>A5</td>
<td>B2</td>
<td>C2</td>
</tr>
<tr>
<td>3</td>
<td>A1</td>
<td>A6</td>
<td>B3</td>
<td>C3</td>
</tr>
<tr>
<td>4</td>
<td>A1</td>
<td>A2</td>
<td>B4</td>
<td>C4</td>
</tr>
<tr>
<td>5</td>
<td>A2</td>
<td>A5</td>
<td>B1</td>
<td>C3</td>
</tr>
<tr>
<td>6</td>
<td>A2</td>
<td>A6</td>
<td>B2</td>
<td>C1</td>
</tr>
<tr>
<td>7</td>
<td>A2</td>
<td>A3</td>
<td>B3</td>
<td>C2</td>
</tr>
<tr>
<td>8</td>
<td>A3</td>
<td>A5</td>
<td>B4</td>
<td>C1</td>
</tr>
<tr>
<td>9</td>
<td>A3</td>
<td>A6</td>
<td>B1</td>
<td>C4</td>
</tr>
<tr>
<td>10</td>
<td>A3</td>
<td>A4</td>
<td>B2</td>
<td>C3</td>
</tr>
<tr>
<td>11</td>
<td>A4</td>
<td>A5</td>
<td>B3</td>
<td>C4</td>
</tr>
<tr>
<td>12</td>
<td>A4</td>
<td>A6</td>
<td>B4</td>
<td>C2</td>
</tr>
</tbody>
</table>

Table 2.3: Number of overlapping items between forms in CBA Design 1.

<table>
<thead>
<tr>
<th>Form (f)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td>v₁</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4: Block composition of each of the 16 forms in CBA Design 2.

<table>
<thead>
<tr>
<th>Form</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>B1</td>
<td>C1</td>
<td>D1</td>
</tr>
<tr>
<td>2</td>
<td>A2</td>
<td>B2</td>
<td>C2</td>
<td>D1</td>
</tr>
<tr>
<td>3</td>
<td>A3</td>
<td>B3</td>
<td>C3</td>
<td>D1</td>
</tr>
<tr>
<td>4</td>
<td>A4</td>
<td>B4</td>
<td>C4</td>
<td>D1</td>
</tr>
<tr>
<td>5</td>
<td>A1</td>
<td>B2</td>
<td>C3</td>
<td>D2</td>
</tr>
<tr>
<td>6</td>
<td>A2</td>
<td>B1</td>
<td>C4</td>
<td>D2</td>
</tr>
<tr>
<td>7</td>
<td>A3</td>
<td>B4</td>
<td>C1</td>
<td>D2</td>
</tr>
<tr>
<td>8</td>
<td>A4</td>
<td>B3</td>
<td>C2</td>
<td>D2</td>
</tr>
<tr>
<td>9</td>
<td>A1</td>
<td>B3</td>
<td>C4</td>
<td>D3</td>
</tr>
<tr>
<td>10</td>
<td>A2</td>
<td>B4</td>
<td>C3</td>
<td>D3</td>
</tr>
<tr>
<td>11</td>
<td>A3</td>
<td>B1</td>
<td>C2</td>
<td>D3</td>
</tr>
<tr>
<td>12</td>
<td>A4</td>
<td>B2</td>
<td>C1</td>
<td>D3</td>
</tr>
<tr>
<td>13</td>
<td>A1</td>
<td>B4</td>
<td>C2</td>
<td>D4</td>
</tr>
<tr>
<td>14</td>
<td>A2</td>
<td>B3</td>
<td>C1</td>
<td>D4</td>
</tr>
<tr>
<td>15</td>
<td>A3</td>
<td>B2</td>
<td>C4</td>
<td>D4</td>
</tr>
<tr>
<td>16</td>
<td>A4</td>
<td>B1</td>
<td>C3</td>
<td>D4</td>
</tr>
</tbody>
</table>

Table 2.5: Number of overlapping items between forms in CBA Design 2.
Table 2.6: Test Blueprint.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Topic</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>8 - 16%</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4 - 10%</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>28 - 38%</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>6 - 12%</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>3 - 8%</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>8 - 16%</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>2 - 6%</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>8 - 16%</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>3 - 8%</td>
</tr>
<tr>
<td>2</td>
<td>J</td>
<td>15 - 20%</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10 - 20%</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>15 - 25%</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>25 - 40%</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5 - 10%</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>5 - 10%</td>
</tr>
</tbody>
</table>

Table 2.7: Block characteristics for CBA designs with 350 or 300 items.

<table>
<thead>
<tr>
<th>Design</th>
<th>Block Group</th>
<th># Blocks</th>
<th>Item Quality</th>
<th># Items/Block (p)</th>
<th># Items/Block (n = 350)</th>
<th># Items/Block (n = 300)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>6</td>
<td>Good</td>
<td>$v_1 = 105$</td>
<td>$v_1 = 88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4</td>
<td>Good</td>
<td>$v_2 = 105$</td>
<td>$v_2 = 88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>4</td>
<td>Bad</td>
<td>$v_3 = 35$</td>
<td>$v_3 = 36$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>4</td>
<td>Good</td>
<td>$v_1 = 105$</td>
<td>$v_1 = 88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4</td>
<td>Good</td>
<td>$v_2 = 105$</td>
<td>$v_2 = 88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>4</td>
<td>Good</td>
<td>$v_3 = 105$</td>
<td>$v_3 = 88$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>4</td>
<td>Bad</td>
<td>$v_4 = 35$</td>
<td>$v_4 = 36$</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.8: Comparison of MTA and CBA designs.

<table>
<thead>
<tr>
<th>Specs</th>
<th>MTA</th>
<th>CB Design 1</th>
<th>CB Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># Items/Form</td>
<td>350</td>
<td>350</td>
<td>300</td>
</tr>
<tr>
<td># Forms</td>
<td>-</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td># Blocks</td>
<td>-</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td># Items Total</td>
<td>-</td>
<td>1190</td>
<td>1400</td>
</tr>
<tr>
<td># Unique Forms</td>
<td>-</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>% Overlap</td>
<td>-</td>
<td>10-30</td>
<td>10-30</td>
</tr>
<tr>
<td>% Exposure</td>
<td>-</td>
<td>25-33</td>
<td>25</td>
</tr>
<tr>
<td>Mean Test Info</td>
<td>64.64</td>
<td>77.26</td>
<td>76.24</td>
</tr>
<tr>
<td>Min Test Info</td>
<td>61.78</td>
<td>77.81</td>
<td>75.57</td>
</tr>
<tr>
<td>Max Test Info</td>
<td>67.15</td>
<td>76.95</td>
<td>76.75</td>
</tr>
<tr>
<td>Mean Time (m)</td>
<td>-</td>
<td>348.36</td>
<td>348.43</td>
</tr>
<tr>
<td>Min Time (m)</td>
<td>-</td>
<td>347.83</td>
<td>347.96</td>
</tr>
<tr>
<td>Max Time (m)</td>
<td>-</td>
<td>348.63</td>
<td>348.82</td>
</tr>
<tr>
<td>Solution Time (s)</td>
<td>Weeks</td>
<td>14.79</td>
<td>18.09</td>
</tr>
</tbody>
</table>

Table 2.9: Misclassification rates for each design based on simulation of 200 examinees per form with 10 repetitions.

<table>
<thead>
<tr>
<th>Misclassification</th>
<th>MTA</th>
<th>CB Design 1</th>
<th>CB Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Fail</td>
<td>Qualified</td>
<td>2.73</td>
<td>2.80</td>
</tr>
<tr>
<td>% Pass</td>
<td>Unqualified</td>
<td>22.13</td>
<td>16.72</td>
</tr>
<tr>
<td>% Total</td>
<td>3.53</td>
<td>3.38</td>
<td>3.83</td>
</tr>
</tbody>
</table>

48
Figure 2.1: Sequential and simultaneous assembly of three unique $n$-item forms with $N$-item pool.
Figure 2.2: Shadow-test assembly of four unique $n$-item forms with $N$-item pool.
Figure 2.3: Simultaneous assembly of three unique forms with sequential selection stimulus and discrete items for each form.

Figure 2.4: Simultaneous assembly of four $n$-item forms with overlap of $v$ items between all pairs.
Figure 2.5: Block assembly of four $n$-item forms with overlap of $v$ items between all pairs.

Figure 2.6: Common block assembly of four $n$-item forms with overlap of $n/2$ items between all pairs.
Figure 2.7: Common block assembly of 12 $n$-item forms, each composed of two A blocks, one B block, and one C block.
Figure 2.8: Common block assembly of 16 n-item forms, each composed of one of every block.
Figure 2.9: Comparison of test information curves (TIC) for MTA forms with 350 items (MF-350), CB Design 1 forms with 350 and 300 items (D1-350 and D1-300), and CB Design 2 forms with 350 and 300 items (D2-350 and D2-300). The dashed vertical line is the designated cut-score of $\theta_c = 0.55$. 

Cut Score = 0.55
CHAPTER 3

THE SCRIPTED TESTING METHOD FOR CAT
AND ON-THE-FLY MST

3.1 Introduction

Multistage testing (MST) is a rising alternative to CAT, in which the test is administered as a series of modules (i.e., groups of items) in stages as opposed to a sequence of individual items. The prototypical paradigm of MST (PMST) with $M$ stages consists of the following five steps: 1) start examinees on a preassembled module of medium difficulty in the first stage ($s = 1$); 2) make an interim estimate of ability based on performance thus far ($\hat{\theta}_s$); 3) route them accordingly to a preassembled module that is easier or harder in the next stage; 4) repeat steps 2 and 3 until the completion of all $M$ stages; 5) evaluate the final estimate of ability ($\hat{\theta}_M$). From the test-taker’s perspective, the primary draw of MST over CAT is the allowance of item review within a module. From the test-user’s perspective, the preassembly of modules in PMST affords the benefits of test form review before administration and, perhaps more importantly, far greater control over item exposure and test constraints than what is possible with CAT. However, these advantages also come with general drawbacks of lower accuracy and security due to inherent limitations on estimation frequency and number of possible test forms, respectively.

In response to these shortcomings of PMST, on-the-fly MST (OMST) has been recently introduced in literature as a promising alternative (Zheng, Wang, Culbertson, & Chang, 2014). The fundamental difference is that, instead of routing to a preassembled module, OMST adaptively assembles a module at each stage in real-time according to $\hat{\theta}_s$. Two broad approaches to on-the-fly assembly have been proposed: 1) selecting items that shape the module information curve to fit a target TIF centered at $\hat{\theta}_s$ via an iterative optimization process (Han & Guo, 2014), and 2) selecting items that
maximize information at $\hat{\theta}_s$ (Zheng & Chang, 2015). The former method emphasizes test fairness by striving for equivalent measurement precision across all ability levels, while the latter method emphasizes accuracy by striving for maximum measurement precision at any given ability level. The latter method is the focus of this study and what is referenced by OMST hereafter. OMST has the major advantage of producing more individualized forms with finer measurement precision, thereby increasing test security and reliability relative to PMST. Moreover, a couple of OMST variations have been suggested to further improve estimation accuracy, including a hybrid method that gradually transitions from OMST to CAT (S. Wang, Lin, Chang, & Douglas, 2016) and more robust item selection methods that account for the uncertainty of interim ability estimates especially at the earlier stages (Tay, 2015).

However, on-the-fly assembly inevitably sacrifices the aforementioned strengths of preassembly, in particular the straightforward implementation of item exposure control and test constraints. Based on extensive simulations, Zheng and Chang (2015) generally recommended using the maximum priority index (Cheng & Chang, 2009) with a remedial item replacement step for any constraint violations, and the Sympon-Hetter method (Hetter & Sympon, 1997) with a stratified item bank for better use of underexposed items. Although shown to perform fairly well, this combination of procedures is quite involved and computationally expensive, thereby impeding practical implementation in many operational contexts. Therefore, this study proposes the Scripted Testing method (McKinley, Petersen, & Spray, 2014) as a simpler solution to the challenge of imposing exposure control and non-statistical constraints in OMST.

3.2 Method

The Scripted Testing method was devised as an item selection algorithm for CAT (Lee, Li, Petersen, & Gawlick, 2014) and generally proceeds as follows: For a test with $m$ items, there are $m$ slots to be filled, and an item is selected according to pre-defined rules for each slot. For the first slot, randomly select an item from a designated content area (collection). For each subsequent slot,

1. discard any enemies of items already administered in previous slots;
2. draw a designated number of items (selection length) from designated
collection according to some selection criterion (e.g., MFI or b-matching);

3. randomly select one from the set of candidates.

There are two distinct features of this method. First, a predetermined se-
quence of collections guarantees meeting content specifications for every ex-
aminee. The specific ordering may be determined either randomly or de-
liberately by content experts. Second, note that a randomesque method of
exposure control (Kingsbury & Zara, 1989) is depicted in steps 2 and 3, where
the selection length balances item usage at expense of ability estimation ac-
curacy. In other words, longer length generally results in better balance but
lower accuracy.

An example of a delivery script for CAT with 30 items is shown in Table
3.1. The collection sequence is predetermined. Also, total indicates the
number total items in the collection, and selection length is the number of
items drawn from the collection whose difficulties are closest in match to
current ability estimate, \( \hat{\theta} \). In slot 1, 25 candidate items drawn from a total
of 25 in collection 9, meaning the item is chosen completely at random within
the collection. In slot 2, 30 best items according to \( \hat{\theta}_1 \) are drawn from a total
of 32 in collection 12, from which an item is randomly selected. The process
is analogous for all subsequent slots until the end when the final ability
estimate, \( \hat{\theta}_{30} \), is obtained.

The adaptation of the Scripted Testing method to OMST is straightfor-
ward. For the first module, randomly select each item from a designated col-
collection. For each subsequent module, the process is the same as in scripted
CAT except the same ability estimate is used for the selection of all items
within the module. An example of a delivery script for a 3-stage OMST, with
10 items in each stage, is shown in Table 3.2. In on-the-fly (OTF) module 1,
note that the selection lengths for all slots are equal to the total, meaning all
items are chosen completely at random within the specified collections. In
OTF module 2, all of the items are selected based on \( \hat{\theta}_1 \), the interim ability
estimate after stage 1. Likewise, in OTF module 3, all of the items are se-
lected based on \( \hat{\theta}_2 \), the interim ability estimate after stage 2. \( \hat{\theta}_3 \) is the final
ability estimate after stage 3.

In this study, scripted CAT (SCAT) and scripted OMST (SOMST, with
3 or 4 evenly divided stages) were simulated with a fixed length of about
30 items. An operational pool of items (approximately 700 items satisfying 3PLM) was obtained from a large-scale math exam. The items represented about 25 collections and included enemy sets. \( N_k = 2,000 \) examinees were generated from each of about \( Q = 20 \) designated values of \( \theta_k \) \((k = 1, \ldots, Q)\) in the range of -4 to 4. The performance indicator of choice was the marginal reliability of \( \theta \), computed as

\[
MR(\hat{\theta}) = \frac{\sum_{k=1}^{Q} w_k \left( \theta_k - \sum_{k=1}^{Q} w_k \theta_k \right)^2}{\sum_{k=1}^{Q} w_k \left( \theta_k - \sum_{k=1}^{Q} w_k \theta_k \right)^2 + \sum_{k=1}^{Q} \frac{w_k}{N_k} \left( \sum_{j=1}^{N_k} \left( \hat{\theta}_{jk} - \theta_k \right)^2 \right)} \tag{3.1}
\]

where \( \hat{\theta}_{jk} \) is the MLE of the \( j \)th examinee with \( \theta_k \), and \( w_k \) is a posterior density weight for \( \theta_k \) that is empirically estimated from previous data by a nonparametric iterative algorithm (Mislevy, 1984). The numerator specifies true variance of \( \theta \) \((\sigma^2_T)\), and the denominator specifies the sum of true and error variances \((\sigma^2_T + \sigma^2_E)\). If equal weights and same sample size of 1 are used for all \( \theta_k \) (i.e., \( w_k = 1/Q \) and \( N_k = 1 \forall k \)), then the total sample size is \( Q \) and the expression simplifies to the more familiar form of reliability:

\[
\rho^2_{\theta,\theta} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E} = \frac{\sum_{k=1}^{Q} (\theta_k - \overline{\theta})^2}{\sum_{k=1}^{Q} (\theta_k - \overline{\theta})^2 + \sum_{k=1}^{Q} (\hat{\theta}_k - \theta_k)^2}. \tag{3.2}
\]

Note that the error variance is also called the mean squared error (MSE), which is commonly used on its own (or its square root) as a metric of overall estimation accuracy:

\[
\sigma^2_E = MSE = \frac{1}{Q} \sum_{k=1}^{Q} \left( \hat{\theta}_k - \theta_k \right)^2. \tag{3.3}
\]

The objective was to maximize \( MR(\theta) \) given the following exposure constraints:

- Number of unused items: 0;
- Maximum conditional exposure rate (MCER): 0.2, 0.3, or 0.4;
- Maximum marginal exposure rate (MMER): 0.2,
where the conditional exposure rate (CER) is the proportion of times an item has been administered to a group of examinees with same $\theta_k$, and the marginal exposure rate is the weighted average of the CERs with weights $w_k$. The objective was achieved by programming an optimization algorithm that searches for the smallest possible selection length for each slot within the constraints.

3.3 Results

The performance results are detailed in Table 3.3 and Figure 3.1. Completely random item selection is included in the table as a baseline reference. The empirical MCER, empirical MMER, and number of unused items are shown as confirmation that the imposed constraints were met. In brief, the following trends were observed:

- At each level of MCER, $MR(\theta)$ increases from SOMST-3 to SOMST-4, then it increases by a greater amount from SOMST-4 to SCAT.

- For each testing design, $MR(\theta)$ increases from MCER = 0.2 to 0.3, then it increases by a lesser amount from 0.3 to 0.4.

3.4 Discussion

The “best” choice of testing design and MCER is ultimately a judgement call based on the requirements of the test-user, but these preliminary results indicate that SOMST is certainly a capable design with comparable performance to SCAT in terms of $MR(\theta)$. In particular, it has been demonstrated that the Scripted Testing method is an incredibly simple yet powerful tool for controlling exposure rates and balancing content to exact specifications in OMST. The promising results and ease of implementation, especially for operations that are already running CAT, should encourage test-users to consider SOMST.
### 3.5 Tables and Figure

Table 3.1: Example of a CAT script for selecting 30 items.

<table>
<thead>
<tr>
<th>Slot</th>
<th>Collection</th>
<th>Total</th>
<th>Selection Length</th>
<th>$\bar{\theta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>25</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>32</td>
<td>30</td>
<td>$\bar{\theta}_1$</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>45</td>
<td>35</td>
<td>$\bar{\theta}_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\bar{\theta}_29$</td>
</tr>
<tr>
<td>30</td>
<td>18</td>
<td>27</td>
<td>9</td>
<td>$\bar{\theta}_{30}$</td>
</tr>
</tbody>
</table>

Table 3.2: Example of an OMST script with 3 stages of 10 items each.

<table>
<thead>
<tr>
<th>OTF MODULE 1</th>
<th>OTF MODULE 2</th>
<th>OTF MODULE 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot</td>
<td>Collection</td>
<td>Total</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>23</td>
</tr>
</tbody>
</table>
Table 3.3: Table of estimated marginal reliability of ability, $MR(\theta)$, for each testing design and maximum conditional exposure rate (MCER) pairing.

<table>
<thead>
<tr>
<th>Design (MCER)</th>
<th>$MR(\theta)$</th>
<th>Empirical MCER</th>
<th>Empirical MMER</th>
<th>Unused Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.7862</td>
<td>0.0670</td>
<td>0.0560</td>
<td>0</td>
</tr>
<tr>
<td>SCAT (0.2)</td>
<td>0.9024</td>
<td>0.1955</td>
<td>0.1072</td>
<td>0</td>
</tr>
<tr>
<td>SCAT (0.3)</td>
<td>0.9135</td>
<td>0.2960</td>
<td>0.1361</td>
<td>0</td>
</tr>
<tr>
<td>SCAT (0.4)</td>
<td>0.9189</td>
<td>0.3600</td>
<td>0.1671</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-3 (0.2)</td>
<td>0.8834</td>
<td>0.1940</td>
<td>0.1068</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-3 (0.3)</td>
<td>0.8975</td>
<td>0.2940</td>
<td>0.1353</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-3 (0.4)</td>
<td>0.8988</td>
<td>0.3485</td>
<td>0.1442</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-4 (0.2)</td>
<td>0.8929</td>
<td>0.1960</td>
<td>0.1089</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-4 (0.3)</td>
<td>0.9017</td>
<td>0.2970</td>
<td>0.1372</td>
<td>0</td>
</tr>
<tr>
<td>SOMST-4 (0.4)</td>
<td>0.9047</td>
<td>0.3525</td>
<td>0.1477</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 3.1: Plot of estimated marginal reliability of ability, $MR(\theta)$, according to maximum conditional exposure rate (MCER) and testing design.
4.1 Introduction

As explained in Chapter 1, the primary objective of computerized adaptive testing (CAT) is to efficiently measure an examinee’s ability (or any latent trait), where efficiency is by and large conceptualized as the degree of measurement precision for a given number of items administered. This is generally accomplished by an algorithm that sequentially selects items according to an information-based optimality criterion. Among the various criteria propagated in literature over the decades, the classic maximum Fisher information criterion (MI; Lord, 1980) remains dominant in current practice due to its straightforward implementation and direct link to measurement precision. Specifically, the asymptotic standard error of the maximum likelihood estimate of ability is the inverse square root of the cumulative Fisher information of scored items. Therefore, in theory, measurement is most precise when selecting items purely based on maximizing Fisher information.

Nevertheless, despite common operationalization, measurement efficiency of CAT should not only be assessed in terms of the number of items administered but also the time it takes to complete the test. To this end, Fan, Wang, Chang, and Douglas (2012) proposed a novel item selection criterion that maximizes the ratio of Fisher information to expected response time (MIT), which can also be interpreted as information per unit of time. In other words, the MIT algorithm selects the next item in the pool with the highest rate of information for the examinee, thus greatly reducing the average completion time for a fixed-length test with only a marginal decrease in the accuracy of ability estimation. In fact, a recent study found that this simple method results in shorter average test times and fewer response time constraint violations compared to imposing explicit constraints or implementing
more complex optimization approaches (Veldkamp, 2016). However, perhaps unsurprisingly, MIT also results in extremely skewed selection of items, since items with both high discrimination and low time-intensity are strongly favored. Given that \(a\)-stratification with \(b\)-blocking (ASB; Chang et al., 2001) is a powerful technique for balancing item exposure, a time-weighted version of it (ASBT) was recommended as a better-balanced alternative to MIT. According to results presented later, however, ASBT comes at substantial costs of increasing both the mean and variance of testing times and estimation error relative to MIT.

Therefore, this chapter investigated the following three alternative techniques for leveraging response times (RTs) in item selection: 1) partitioning the item pool into multiple stages according to time intensities and utilizing MIT within each stage; 2) maximizing the ratio of Fisher information to the absolute difference between item time intensity and examinee latent speed; and 3) maximizing the ratio of Fisher information to an optimally centered and weighted expected RT. Extensive simulations, with item pools and examinees that are both simulated and real, were conducted to evaluate the performances of these methods in controlling both item exposure and testing time distribution while maintaining an adequate level of measurement precision.

4.2 RT Framework

In recent years, there has been a growing interest in using response times in testing. The immense potential of RTs as a rich source of information is certainly not news, but their practical utility could not be realized until the advent of modern computerized test delivery. These days, test delivery software can now store virtually all examinee by task interactions, including RTs for every item, thus greatly facilitating endeavors to harness them via modeling. Some of the more popular models include the lognormal model (van der Linden, 2006), a generalization of the lognormal called the Box-Cox normal model (Klein Entink, van der Linden, & Fox, 2009), a flexible semiparametric approach called the Cox proportional hazards model (C. Wang, Fan, Chang, & Douglas, 2013), and a further generalization called the linear transformation model that subsumes the previous three as special cases.
(C. Wang, Chang, & Douglas, 2013). Each of these RT models was primarily developed as a component in van der Linden’s (2007) two-level hierarchical framework for modeling speed and accuracy. The first level consists of separate measurement models for latent speed and accuracy (e.g., lognormal and 3PL, respectively), and the second level specifies the population and item-domain models (i.e., joint distributions of person and item parameters, respectively). Note that the population model relates speed and accuracy across examinees using a covariance parameter. On the other hand, this modeling framework disregards the within-person speed-accuracy tradeoff, a particularly robust cognitive phenomenon in reaction time tasks. Unless a test is unduly speeded, a reasonable assumption is made that an examinee operates steadily at his or her innate speed, thereby precluding any speed-induced fluctuations in accuracy (van der Linden, Breithaupt, Chuah, & Zhang, 2007).

Among a variety of RT models, the lognormal is perhaps the most recognized due to its relative simplicity and practicability for typical RT data. While it lacks the flexibility of more complex and general models, it is one of the most straightforward to conceptualize and implement, particularly within the hierarchical framework. Specifically, the lognormal model defines the density function of response time for examinee \(i\) on item \(j\) \((T_{ij})\) given the latent speed parameter for the examinee \(\tau_i\) as

\[
f(t_{ij}|\tau_i) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} e^{-[\alpha_j(\log t_{ij} - \beta_j + \tau_i)]^2 / 2}, \tag{4.1}
\]

where \(\alpha_j\) and \(\beta_j\) are the time discrimination and time intensity parameters for item \(j\), respectively, and \(\beta_j\) and \(\tau_i\) are fixed to be on the same scale. Rewriting the density function in standard form,

\[
f(t_{ij}|\tau_i) = \frac{1}{t_{ij}\sqrt{2\pi(1/\alpha_j)^2}} e^{-[\log t_{ij} - (\beta_j - \tau_i)]^2 / [2(1/\alpha_j)^2]}, \tag{4.2}
\]

it becomes clear that \(\mu = \beta_j - \tau_i\) and \(\sigma^2 = (1/\alpha_j)^2\). Thus, the marginal model can be written as

\[
T_{ij}|\tau_i \sim \log-N[\beta_j - \tau_i, 1/\alpha_j^2]. \tag{4.3}
\]

Finally, given that the expected value of a lognormal random variable with
log-mean $\mu$ and log-variance $\sigma^2$ is $e^{\mu+\sigma^2/2}$, an examinee’s expected RT for an item is

$$E(T_{ij}|\tau_i) = e^{\beta_j - \tau_i + 1/(2\alpha_j^2)}.$$  \hspace{1cm} (4.4)

Note that items with low $\beta_j$ and high $\alpha_j$ have low $E(T_{ij}|\tau_i)$.

### 4.3 Motivation

In efforts to increase measurement efficiency in terms of time, Fan et al. (2012) demonstrated the integration of response time into MI by inversely weighting the Fisher information by the examinee’s expected RT for each item. The next item chosen is the one that maximizes the MIT criterion, now formally defined as

$$IT_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)}. \hspace{1cm} (4.5)$$

Here, $\hat{\tau}_i$ is the MLE of $\tau_i$, which is conveniently computed by the closed form expression

$$\hat{\tau}_i = \frac{\sum_{j=1}^{k} \alpha_j^2 (\beta_j - \log t_{ij})}{\sum_{j=1}^{k} \alpha_j^2} \hspace{1cm} (4.6)$$

given an examinee’s RTs $t_{i1}, ..., t_{ik}$ for the $k$ items administered thus far (van der Linden, 2006). The Fisher information function with respect to $\tau_i$ is

$$I^{(k)}(\tau_i) = -E \left( \frac{\partial^2}{\partial \tau_i^2} \log L(\tau_i|t_i) \right) = \sum_{j=1}^{k} \alpha_j^2, \hspace{1cm} (4.7)$$

where

$$L(\tau_i|t_i) = f(t_{i1}, ..., t_{ik}|\tau_i) = \prod_{j=1}^{k} f(t_{ij}|\tau_i) \hspace{1cm} (4.8)$$

(van der Linden, 2008). Therefore, the standard error of $\hat{\tau}_i$ is

$$SE(\hat{\tau}_i) = \sqrt{\frac{1}{I^{(k)}(\hat{\tau}_i)}} = \left( \sum_{j=1}^{k} \alpha_j^2 \right)^{-1/2}, \hspace{1cm} (4.9)$$
which is particularly advantageous since, unlike $\theta_i$, the measurement precision of $\tau_i$ is independent of its current estimate.

Clearly, MIT favors items with high information and low expected RTs, thus attempting to accomplish the two (possibly competing) tasks of accurately estimating ability while reducing the testing time as much as possible. Although quite successful in this regard, Fan et al. (2012) showed that MIT also results in item exposure that is even more skewed than MI. Hence, they introduced ASBT as a compromise that stratifies the item pool as in ASB and inversely weights the $b$-matching criterion by the expected RT. Specifically, this method selects the next item in the present stratum that maximizes the following criterion:

$$BT_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{B_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)},$$

which was shown to balance item exposure rather well, but as shown later, largely by heavily sacrificing the benefits of time-weighting in the first place.

4.4 General Method

In search of alternatives to MIT or ASBT due to their aforementioned drawbacks, this chapter investigated the performance of three new RT-informed criteria for item selection in CAT, all under the hierarchical framework with 3PL and lognormal as the measurement models. In the simulation studies that follow, each of these new methods was directly compared to MIT and ASBT, along with MI as the performance baseline and Random (i.e., completely random item selection) as a reference for ideal item pool usage but worst accuracy.

4.4.1 Proposed Item Selection Procedures

The first method is called $\beta$-partitioned MIT (BMIT), in which $\beta$-partitioning works analogously to the $b$-blocking procedure in ASB. For a given item pool, the items are sorted according to increasing $\beta$ values and evenly partitioned into a specified number of stages as illustrated in Figure 4.1. Items are then selected from each successive stage using MIT, proceeding from the lowest to highest $\beta$-partitions. In this way, BMIT forces a more balanced selection
of items across the entire range of time intensities as opposed to a normally very biased selection of low-β items.

The second method is called MI with β-matching (MIB), which inversely weights Fisher information by the absolute difference between β_j and τ_i in lieu of E(T_{ij}|τ_i):

\[ IB_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{|\beta_j - \hat{\tau}_i|}. \] (4.11)

This method primarily stems from the hypothesis that, compared to MIT, the item exposure skew could be greatly reduced when examinees are administered items in accordance with their latent speed. Provided that the distributions of β_j and τ_i are similar, matching them as closely as possible would be far less restrictive than perpetually selecting items with the lowest β_j and highest α_j values. Moreover, MIB would have the additional benefit of lower RT variability across examinees compared to MIT. This is because MIB strives to achieve β_j = τ_i, in which case the expected RT for item j is reduced to

\[ E(T_{ij}) = e^{1/(2\alpha_j^2)} \] (4.12)

for any examinee regardless of latent speed.

The third method is a generalization of MIT (GMIT), which appends a centering value v and weighting exponent w to the expected RT term:

\[ IT_j^G(\theta_i, \tau_i) = \frac{I_j(\theta_i)}{|E(T_{ij}|\tau_i) - v|^w}, \quad \{v, w\} \in \mathbb{R}^2_{\geq 0}. \] (4.13)

Note that GMIT reduces to MI for w = 0 and MIT for \{v, w\} = \{0, 1\}. The rationale of the generalization is as follows: First, maximizing IT_j^G is in part achieved by minimizing |E(T_{ij}|τ_i) − v|, which occurs when E(T_{ij}|τ_i) = v.

In the case of MIT where v = 0, expected RT of zero is the unattainable lower bound regardless of τ_i, so the effective item pool is severely confined to a handful of the least time-intensive items. This also results in substantial variability of testing times, since much of the same items are being administered to all examinees of varying speeds. In contrast, for a reasonable value of v > 0, the RT-optimal items would vary from person to person depending on τ_i, consequently improving item pool usage. This would also stabilize testing times, because every examinee would generally be administered items that take on average v time units to answer. Second, w allows for varying
the influence of the centered expected RT in item selection. Presumably, decreasing \( w \) would decrease the influence of \(|E(T_{ij}|\tau_i) - v|\), thereby improving item exposure balance at the expense of increasing overall testing time. Third, the absolute value of the centered expected RT is taken since it is of no consequence whether the expected RT is lower or higher than \( v \) (and taking a non-integer exponent of a negative value may result in a complex number). For the simulation studies presented shortly, the sets of \( v \) and \( w \) values were limited to \( V = \{0.0, 0.1, ..., 3.0\} \) and \( W = \{0.50, 0.75, 1.00\} \), respectively, and every \( \{v, w\} \in V \times W \) was run (for a total of \(|V \times W| = 93\) scenarios).

4.4.2 Evaluation Criteria

The following criteria were used to evaluate the performance of each item selection method given \( n \) examinees:

- root mean squared error (RMSE) for estimation accuracy of examinee parameters,

\[
RMSE(\hat{\theta}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}, \quad (4.14)
\]

\[
RMSE(\hat{\tau}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\tau}_i - \tau_i)^2}; \quad (4.15)
\]

- mean and standard deviation of testing times (\( tt_i \)) across examinees as measures of time efficiency and stability,

\[
\bar{tt} = \frac{1}{n} \sum_{i=1}^{n} tt_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in S_i} t_{ij}, \quad (4.16)
\]

\[
s_{tt} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (tt_i - \bar{tt})^2}, \quad (4.17)
\]

where \( S_i \) is the set of all items administered to examinee \( i \);

- mean and standard deviation of pairwise test overlap rates (\( R_{ii'} \)) be-
tween all possible pairs of examinees $i$ and $i'$ as measures of test security,

$$\overline{R} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} R_{ii'} = \frac{n}{L(n-1)} \sum_{j=1}^{m} \overline{U}_j^2 - \frac{1}{n-1}, \quad (4.18)$$

$$s_R = \sqrt{\left[ \left( \frac{n}{2} \right) - 1 \right]^{-1} \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} (R_{ii'} - \overline{R})^2}, \quad (4.19)$$

where $m$ is the size of the item pool, $L$ is the fixed test length, $R_{ii'}$ is computed as the number of common items between a pair of examinees divided by $L$, and $\overline{U}_j$ is the observed exposure rate for item $j$ calculated as the number of times the item was used divided by $n$ (see Chapter 1 for details). Also, C. Wang, Zheng, and Chang (2014) advocated the use of $s_R$ in addition to the traditional $\overline{R}$, since it is entirely possible to have low $\overline{R}$ overall but very high $R_{ii'}$ among a subgroup of examinees. From this perspective, a relatively constant $R_{ii'}$ but slightly higher $\overline{R}$ is generally preferable to a widely varying $R_{ii'}$ but lower $\overline{R}$. It is worth noting that Fan et al. (2012) used the $G$ statistic to measure the skewness of item exposure rates (originally introduced and coined as $\chi^2$ statistic by Chang and Ying (1999)). Although $G$ and $\overline{R}$ are sometimes reported as two distinct statistics that capture different aspects of item pool usage (e.g., Cheng, Chang, & Yi, 2007; Deng, Ansley, & Chang, 2010), it was shown in Chapter 1 that one is simply a linear transformation of the other as follows:

$$G = \frac{m(n-1)}{n} \overline{R} + \frac{m}{n} - L. \quad (4.20)$$

The present chapter opted to report $\overline{R}$, instead of $G$, for its more intuitive interpretation and wider familiarity.

For easy reference, all item selection methods and evaluation criteria are summarized in Table 4.1.
4.5 Study 1: Simulated Item Pools and Examinees

4.5.1 Method

For this initial study, hundreds of simulations were conducted with a broad range of parameter values in efforts to ensure that the findings are not limited to idiosyncratic data. In the interest of brevity and clarity, just two representative sets of simulated item pools and examinees are presented here to evaluate the item selection criteria under disparate conditions. The first set of data was specified as

**Set 1**

- \((a^*_j, b_j, \beta_j) \sim MVN[\mu_1, \Sigma_1]\),  
  \[\mu_1 = \begin{bmatrix} 0.3 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{bmatrix}\],

  where \(a^*_j = \log a_j\), meaning \(a_j\) has a lognormal distribution;

- \(c_j \sim \beta[2,10]\);

- \(\alpha_j \sim U[2,4]\);

- \((\theta_i, \tau_i) \sim MVN[\mu_2, \Sigma_2]\),  
  \[\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.00 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}\],

and the second set of data was specified as

**Set 2**

- \((a^*_j, b_j, \beta_j) \sim MVN[\mu_1, \Sigma_1]\),  
  \[\mu_1 = \begin{bmatrix} 0.30 \\ 0.00 \\ -0.25 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.20 \\ 0.00 & 0.20 & 0.16 \end{bmatrix}\];

- \(c_j \sim \beta[2,10]\);

- \(\alpha_j \sim U[0.5,2.5]\);

- \((\theta_i, \tau_i) \sim MVN[\mu_2, \Sigma_2]\),  
  \[\mu_2 = \begin{bmatrix} 0.00 \\ 0.25 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.00 & 0.20 \\ 0.20 & 0.16 \end{bmatrix}\).
Note that there are two key differences between the sets: 1) the marginal distributions of $\beta_j$ and $\tau_i$ are identical in Set 1, whereas they are narrower and shifted apart in Set 2, and 2) the mean of $\alpha_j$ is greater in Set 1. Otherwise, the parameter specifications are equivalent.

For each set, $m = 500$ items and $n = 1,000$ examinees were randomly generated, then each examinee’s response and RT were simulated for every item. The test length was fixed at $L = 50$ items, with the first item chosen randomly in order to calculate initial estimates of $\theta_i$ and $\tau_i$. Estimation was performed with a combination approach, in which EAP was used as an interim substitute whenever MLE failed. For ASBT, the item pool was divided into 5 strata of 100 items each, then 10 items were selected in each successive stage. For BMIT, the following $\beta$-partitions were implemented:

- 1 $\beta$-partition: equivalent to no $\beta$-partitioning (i.e., single partition of 500 items).
- 2 $\beta$-partitions: item pool divided into low and high $\beta$-partitions of 250 items each. The first 25 items were selected in the low $\beta$ stage, then the next 25 items were selected in the high $\beta$ stage.
- 3 $\beta$-partitions: item pool divided into low, mid, and high $\beta$-partitions with 167, 167, and 166 items, respectively. The first 17 items were selected in the low $\beta$ stage, the next 17 items were selected in the mid $\beta$ stage, then the final 16 items were selected in the high $\beta$ stage.

4.5.2 Results

Figures 4.2 and 4.3 show the results of BMIT and MIB with Sets 1 and 2, respectively. Each of the evaluation criteria are plotted as a function of the number of $\beta$-partitions, which only applies to BMIT marked as ◦’s. Note that MIT, marked separately as ×, is equivalent to BMIT with one $\beta$-partition. All of the other methods are plotted as horizontal lines representing a single value. The following observations can be made on each set of criteria:

1. Estimation Accuracy: In terms of $RMSE(\hat{\theta})$, BMIT was very close MI regardless of the number of $\beta$-partitions, while MIB was very close to ASBT but still well below Random. $RMSE(\hat{\tau})$ (shown in the shaded plot area) was extremely low and essentially equivalent for all methods.
There were no discernable differences in relative performance between Sets 1 and 2.

2. **Mean and Standard Deviation of Testing Times**: \(\overline{t}\) and \(s_{tt}\) generally increased for BMIT with more \(\beta\)-partitions, the effect being greater with Set 1 than Set 2. For Set 1 in particular, the distribution of testing times for BMIT became worse than that of ASBT and MI beyond 2 \(\beta\)-partitions. MIB performed exceptionally well with Set 1, which was second only to MIT in terms of \(\overline{t}\) and even better than MIT in terms of \(s_{tt}\); on the contrary, MIB performed terribly with Set 2, where both \(\overline{t}\) and \(s_{tt}\) were the worst out of all methods.

3. **Mean and Standard Deviation of Test Overlap Rates**: \(R\) generally decreased for BMIT with more \(\beta\)-partitions, while \(s_R\) generally remained the same regardless. Even with 5 \(\beta\)-partitions, BMIT had higher \(R\) and only slightly lower \(s_R\) than MI. MIB performed more similarly to ASBT, especially with Set 1 where MIB was nearly identical to ASBT and very close to Random in terms of \(R\).

In summary, BMIT is almost as accurate as MI in terms of estimation, but using as few as 2 or 3 \(\beta\)-partitions may inordinately increase the mean and standard deviations of testing times while hardly improving the balance of item exposure compared to MIT. On the other hand, MIB is generally similar to ASBT in terms of estimation and better at controlling test overlap rates than BMIT and MI; however, it may counterproductively increase the mean and variance of testing times if the distributions of \(\beta\) and \(\tau\) are significantly non-overlapping. Therefore, neither BMIT nor MIB prove to be practicable techniques in broader contexts.

Figure 4.4 shows the results of GMIT with Set 1. The corresponding results with Set 2 were very similar in all respects, so they are not presented here. Each of the evaluation criteria are plotted as a function of \(v\), which only applies to GMIT. Note that MIT, explicitly marked with \(\times\), is equivalent to GMIT at \(v = 0\) and \(w = 1\). Also note that MIT, ASBT, MI and Random are all exactly the same as in Figures 4.2 and 4.3. This time, the following observations about GMIT can be made on each set of criteria:

1. **Estimation Accuracy**: \(RMSE(\hat{\theta})\) slowly climbed then leveled-out as \(v\) increased. For \(w = 1\), \(RMSE(\hat{\theta})\) plateaued around the level of ASBT.
At any given $v$, $RMSE(\hat{\theta})$ was always less for smaller $w$, eventually reaching the level of MI as $w$ approaches 0. As before, $RMSE(\hat{\tau})$ was extremely low and essentially equivalent for all methods.

2. Mean and Standard Deviation of Testing Times: Larger $w$ led to lower $\overline{t}$ from $v = 0$ to about 1, at which point $\overline{t}$ equalized for all $w$, then the trend reversed for $v$ beyond 1. On the other hand, larger $w$ always resulted in lower $s_\overline{t}$ at any $v$. For any $w$, $\overline{t}$ and $s_\overline{t}$ was minimized at about $v = 0.3$ and $v = 1.1$, respectively. At these minimum points, GMIT far outperformed all other methods.

3. Mean and Standard Deviation of Test Overlap Rates: Larger $w$ led to higher $\overline{R}$ from $v = 0$ to about 0.5, at which point $\overline{R}$ equalized for all $w$, then the trend reversed for $v$ beyond 0.5. On the other hand, larger $w$ led to lower $s_R$ from $v = 0$ to about 0.3, at which point $\overline{R}$ equalized for all $w$, then the trend reversed for $v$ beyond 0.3. For any $w$, $\overline{R}$ and $s_R$ were both minimized at about $v = 1.1$. At this minimum point, GMIT performed comparably to ASBT.

Several of these observed patterns deserve some elucidation. First, perhaps counterintuitively, $\overline{t}$ was minimized and $\overline{R}$ was maximized not at $v = 0$ but at about $v = 0.3$, which was the approximate minimum of the expected RT at the median of $\tau$: $\text{min}(E[T_j|\text{med}(\tau)]) \approx 0.3$. Since no items can have an expected RT of zero, $E(T_{ij}|\tau_i)$ centered at the representative minimum will generally be less than $E(T_{ij}|\tau_i)$ itself, thereby having greater weight in $IT^G_j$. Second, $\overline{R}$ and $s_R$ were minimized at about $v = 1.1$, which was the approximate median of the expected RT at the median of $\tau$: $\text{med}(E[T_j|\text{med}(\tau)]) \approx 1.1$. A heuristic explanation is that centering the expected RT at its centermost value allows the greatest flexibility in selecting items for examinees at both ends of the $\tau$ spectrum, thereby optimizing item pool usage. Third, $s_\overline{t}$ also happened to be minimized at about $v = 1.1$ for this particular data, but a clear pattern could not be discerned in general. Fourth, $w$ instigated a distinct tradeoff between $RMSE(\hat{\theta})$ and performance on other criteria, specifically $\overline{R}$ for $v > 0.5$ and $s_\overline{t}$. Nevertheless, the effects of $w$ were relatively minor compared to the influence of $v$ on general performance. Therefore, the best performer for this data seemed to be GMIT with $v = 1.1$, with the less important choice of $w$ mostly depending on the minimum accuracy or
maximum average rate of test overlap deemed acceptable.

4.6 Study 2: Real Item Pool and Examinees

4.6.1 Method

To further validate the effectiveness of GMIT, the procedure was next implemented on a set of real data from a high-stakes, large-scale standardized CAT (bestowed by a generous source). The data consisted of raw responses and RTs from about 2,000 examinees, and the item pool contained about 500 multiple-choice items that were pre-calibrated according to 3PLM. The lognormal model item parameters \((\alpha, \beta)\) were estimated using a modified version of van der Linden’s (2007) MCMC routine that fixed the 3PLM item parameters \((a, b, c)\) to the pre-calibrated values, and the distribution of \(\tau\) was set to have a mean of 0. All parameters appeared to converge using 10,000 MCMC draws with a burn-in size of 5,000, and the model seemed to fit well enough for the current application.

For CAT simulation, each examinee’s responses and RTs were generated for all items. The test length was fixed at \(L = 30\), with the first item chosen randomly in order to calculate initial estimates of \(\theta_i\) and \(\tau_i\). As before, estimation was performed using a combination of MLE and EAP. For ASBT, the item pool was divided into 5 strata of about 100 items each, then 6 items were selected in each successive stage.

4.6.2 Results

Figure 4.5 shows the results of GMIT with the real data, which exhibit much of the same patterns as the earlier results with simulated data in Figure 4.4. First, \(\overline{R}\) was minimized and \(\overline{R}\) was maximized at about \(v = 0.6\), which was the approximate minimum of the expected RT at the median of \(\tau\): \(\min(E[T_j | \text{med}(\tau)]) \approx 0.6\). Second, \(s_R\) and \(s_R\) were at their minimum at about \(v = 1.8\), which was the approximate median of the expected RT at the median of \(\tau\): \(\text{med}(E[T_j | \text{med}(\tau)]) \approx 1.8\). Third, \(s_{tt}\) was minimized at about \(v = 1\). Fourth, the tradeoff between \(RMSE(\hat{\theta})\) and performance on other criteria were even less salient than with the simulated data. All things
considered, an optimal combination for this real data could be $v = 1.3$ and $w = 0.5$, which afforded better accuracy than ASBT, kept average testing time close to MIT, drastically reduced the variability of testing times to near minimum, and provided a level of item exposure control comparable to ASBT.

4.7 Discussion

Continual efforts to refine the item selection algorithm in CAT is not only of scholarly interest but also of paramount importance to operational testing. It goes without saying that accurately measuring ability, saving valuable time and resources, minimizing differential speededness among examinees, and strengthening test security are all critical considerations for most high-stakes administrations. In this spirit, the present investigation sought to improve upon the innovative RT-based item selection methods introduced by Fan et al. (2012). The results of extensive simulations, with both real and simulated data, provide strong evidence for the overall superiority of the proposed GMIT over the other evaluated methods. Ultimately, GMIT with carefully chosen centering and weighting values can appreciably increase the validity of test scores, with negligible detriment to measurement precision, in two distinct aspects: curtailing the likelihood of time pressure-induced rapid guessing by markedly reducing the mean and variance of testing times, and decreasing the chances of item preknowledge by dramatically reducing the mean and variance of test overlap rates. The truly remarkable feature of GMIT is that all of these benefits can be realized without imposing explicit item exposure controls or RT constraints (cf. van der Linden, 1999).

The initialization of GMIT for use in practice requires the following steps: 1) calibrating the item pool with appropriate measurement models for ability and speed given responses and response times, respectively, 2) generating examinees based on a reasonable or empirically motivated assumption about the joint distribution of ability and speed of the target population, 3) establishing a set of evaluation criteria, 4) conducting a series of CAT simulations with a range of $v$ and $w$ values, and 5) selecting the optimal $\{v, w\}$ according to performance on the evaluation criteria. If performance is evaluated on two or more criteria that involve tradeoffs, the “optimal” choice ultimately
depends on the minimally acceptable levels on the criteria (e.g., $R \leq 0.20$) or the user’s rational judgment, which can be done via visual inspection of the results as demonstrated.

Alternatively, if a more objective measure is desired to aid in the decision, it is possible to construct an optimality index such as the following:

$$\Omega_{\{v,w\}} = \gamma^T Z_{\{v,w\}}, \quad \{v, w\} \in V \times W,$$

where $\gamma$ is a vector of weights and $Z_{\{v,w\}}$ is a vector of standardized values for each evaluation criterion given $\{v, w\}$. Placing all of the criteria on the same scale through standardization is necessary to ensure that the weighted composite is not influenced by the magnitude and spread of the original scales. Provided that lower values indicate better performance for every criterion, the optimal choice would be $\{v, w\}$ that minimizes $\Omega_{\{v,w\}}$, which could be interpreted as a weighted average of the standardized criteria if the values of $\gamma$ are non-negative and sum to 1. $\gamma$ would be specified according to the importance attributed to each criterion in the overall performance evaluation. As a simple example with the real data results, Table 4.2 shows an excerpt of rank ordered $\Omega_{\{v,w\}}$ values computed using weights of $1/6$ for each of the six evaluation criteria. According to this evenly weighted index, $\{v, w\} = \{1.4, 0.75\}$ was the most optimal whereas the previous choice of $\{v, w\} = \{1.3, 0.50\}$ ranked tenth out of 93. The latter choice placed more emphasis on ability estimation accuracy over the other criteria, but the practical differences between the two choices were relatively slight nonetheless.

Painstaking efforts were taken to assure that the proposed procedure and outcome can be generalized to a broad range of item bank structures and test-taking populations. Although the current investigation was limited to fixed-length CAT with commonly utilized unidimensional 3PL and lognormal models under the hierarchical framework, the flexibility of GMIT allows for easy implementation and evaluation under a wide variety of schemes. For instance, a recent paper reported success in utilizing the original MIT method in computerized classification testing (CTT) with the SPRT stopping rule (Sie, Finkelman, Riley, & Smits, 2015). As a next step, GMIT could be easily tried in the same context with a straightforward modification. Moreover, further scrutiny is certainly warranted to confirm the usefulness of the technique in operational CAT, which is frequently constrained by practical
requirements such as content balancing and ordering. This could not be studied at present because the real data at hand did not contain non-statistical specifications, but there are few compelling reasons to suspect a drastic degradation in GMIT’s efficacy under realistic circumstances. Finally, it would be informative to conduct a separate study comparing GMIT to other RT-based methods not considered in this chapter, including various mathematical optimization approaches (Veldkamp, 2016) and a simplified version of MIT that uses sample-based average log-RTs (in lieu of model-based expected RTs) with randomesque exposure control (Cheng, Diao, & Behrens, 2016).

As a supplemental consideration, although BMIT did not prove to be effective in regards to its originally intended purpose, $\beta$-partitioning may have potential in substantive applications. One such possibility could be abating test anxiety caused by perceived speededness. Conceivably, time-intensive items at the start of a timed test may elicit subpar performance by those who have not properly “warmed up” and harbor legitimate fears of running out of time. The serious underestimation of ability due to such uncharacteristic errors on initial items is well-documented (Chang & Ying, 2008). By $\beta$-partitioning the item pool and selecting items in stages of increasing $\beta$, examinees would start off with short items and gradually progress to longer items, which may help allay time-induced anxiety and thus improve the accuracy of ability estimation. Clearly, empirical studies would need to be conducted to investigate this conjecture.
4.8 Tables and Figures

Table 4.1: Summary of item selection methods and evaluation criteria.

<table>
<thead>
<tr>
<th>Item Selection Methods</th>
<th>Evaluation Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI Maximum Information</td>
<td>$RMSE(\hat{\theta})$ Root Mean Squared Error of $\hat{\theta}$</td>
</tr>
<tr>
<td>MIT MI with Time</td>
<td>$RMSE(\hat{\tau})$ Root Mean Squared Error of $\hat{\tau}$</td>
</tr>
<tr>
<td>ASB $a$-stratification with $b$-blocking</td>
<td>$\bar{t}$ Mean Test Time</td>
</tr>
<tr>
<td>ASBT ASB with Time</td>
<td>$s_{tt}$ Std. Dev. of Test Time</td>
</tr>
<tr>
<td>MIB MI with $\beta$-matching</td>
<td>$\bar{t}_{or}$ Mean Test Overlap Rate</td>
</tr>
<tr>
<td>BMIT $\beta$-partitioned MIT</td>
<td>$s_{tor}$ Std. Dev. of Test Overlap Rate</td>
</tr>
<tr>
<td>GMIT Generalized MIT</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Average of standardized evaluation criteria, $\Omega_{\{v,w\}}$, for GMIT with real data.

<table>
<thead>
<tr>
<th>Rank</th>
<th>${v,w}$</th>
<th>$\Omega_{{v,w}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{1.4, 0.75}</td>
<td>-0.4746</td>
</tr>
<tr>
<td>2</td>
<td>{1.5, 1.00}</td>
<td>-0.4537</td>
</tr>
<tr>
<td>3</td>
<td>{1.5, 0.75}</td>
<td>-0.4436</td>
</tr>
<tr>
<td>4</td>
<td>{1.4, 0.50}</td>
<td>-0.4182</td>
</tr>
<tr>
<td>5</td>
<td>{1.3, 1.00}</td>
<td>-0.4070</td>
</tr>
<tr>
<td>6</td>
<td>{1.6, 0.50}</td>
<td>-0.4027</td>
</tr>
<tr>
<td>7</td>
<td>{1.6, 0.75}</td>
<td>-0.3935</td>
</tr>
<tr>
<td>8</td>
<td>{1.3, 0.75}</td>
<td>-0.3865</td>
</tr>
<tr>
<td>9</td>
<td>{1.9, 0.75}</td>
<td>-0.3758</td>
</tr>
<tr>
<td>10</td>
<td>{1.3, 0.50}</td>
<td>-0.3708</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>93</td>
<td>{3.0, 0.75}</td>
<td>0.5055</td>
</tr>
</tbody>
</table>
Figure 4.1: An illustration of the $\beta$-partitioning process.
Figure 4.2: Performances of BMIT and MIB for simulated data: Set 1. The \( \beta \)-partitions only apply to BMIT.
Figure 4.3: Performances of BMIT and MIB for simulated data: Set 2. The $\beta$-partitions only apply to BMIT.
Figure 4.4: Performances of GMIT for simulated data: Set 1. The centering values $v$ only apply to GMIT.
Figure 4.5: Performances of GMIT for real data. The centering values $v$ only apply to GMIT.
5.1 Introduction

In a typical administration of computerized adaptive testing (CAT), items are sequentially selected in real-time from a large item pool according to the examinee’s current performance. Ideally, this provides each examinee with a unique set of items with minimal overlap, thereby discouraging cheating by copying or sharing answers. In practice, however, item selection algorithms based on maximizing information (or minimizing standard error of measurement) are generally prone to highly unbalanced item exposure. Among other concerns, frequently administered items are at great risk for becoming compromised, thereby undermining the integrity of the test.

To counter such a glaring security issue in CAT, much psychometric research in test security has been focused on preventative measures involving some form of item exposure control while still maintaining the efficiency or accuracy of ability estimation as much as possible. However, even the most successful exposure controls cannot entirely prevent the problem of compromised items in practice, simply because a realistic item pool size is usually much smaller than the number of examinees. Since most items will necessarily be administered multiple times, they are inevitably vulnerable to compromise by unscrupulous test-takers. Therefore, there is a great need for diagnostic measures to spot anomalous behavior of both examinees and items alike.

From the examinee perspective, the strategy is to detect an aberrant pattern of responses or response times (RTs) across all items that have been administered to the test-taker. There is extensive literature on the use of person misfit statistics and related methods for this general purpose, including but certainly not limited to the following: the $l_2$ statistic and variations
thereof (Drasgow, Levine, & Williams, 1985; R. Armstrong, Stoumbos, Kung, & Shi, 2007), caution indices (Tatsuoka, 1984; McLeod & Lewis, 1999), score ratio (Karabatsos, 2003), likelihood ratio (Levine & Drasgow, 1988), KL divergence and K-Index (Belov, Pashley, Lewis, & Armstrong, 2007; Belov & Armstrong, 2010), posterior shift (Belov, 2015), data forensics (Impara & Kingsbury, 2005), effective response time (Meijer & Sotaridona, 2006), Bayesian checks (van der Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008; van der Linden & Lewis, 2015; Marianti, Fox, Marianna, Veldkamp, & Tijmstra, 2014), and CUSUM techniques (van Krimpen-Stoop & Meijer, 2001; Meijer, 2002; R. D. Armstrong & Shi, 2009; Egberink, Meijer, Veldkamp, Schakel, & Smid, 2010; Tendeiro & Meijer, 2012). Additionally, more recent efforts on cheating detection are nicely compiled in handbooks by Kingston and Clark (2014) and Cizek and Wollack (2017b).

From the item perspective, the strategy is to detect an aberrant pattern of responses or RTs across all examinees that have been administered the item. However, literature on this front is surprisingly scarce, the only few examples including a merged information theory and combinatorial optimization algorithm (Belov, 2014), a dual differential person functioning (DPF) and differential item functioning (DIF) approach (O’Leary & Smith, 2013), and a log-odds ratio index of item fit (McLeod & Schnipke, 1999). These particular methods can be effective in detecting compromised items, but only after a group of aberrant (or a larger set of potentially aberrant) examinees have been identified at the end of a testing cycle. The possibility of detecting compromised items in real-time was explored in two studies, both of which utilized CUSUM to sequentially monitor item parameter drift. A pioneering study by Veerkamp and Glas (2000) employed a standardized CUSUM statistic for detecting drift in the restricted 3PLM (i.e., fixed c parameter), and a recent study by Kang and Chang (2016) extended the technique by using a log-likelihood CUSUM statistic for detecting overall drift in both the unrestricted 3PLM and the lognormal model of RTs within the hierarchical framework (van der Linden, 2007). Although these methods demonstrated great promise, their major drawback is the enormous computational burden of estimating item parameters at each sequential step. Consequently, for practical implementation, CUSUM in this context can only be performed at intervals throughout the usage lifetime of an item (e.g., every 100 times the item is exposed).
To the best of our knowledge, the only published methods for true real-time detection are sequential analysis procedures introduced by Zhang (2014) and Zhang and Li (2016). The essential idea in both of these papers is that, for a CAT administration period with a set item bank, each item can be continuously monitored after every exposure for any significant increase in the proportion of correct responses. The various implementations of this technique are illustrated shortly after explaining the requisite theoretical framework in the next section. In brief, the procedures were shown to be capable of detecting compromised items quickly with relatively high accuracy under certain conditions, albeit with room for improvement.

Therefore, in efforts to build upon this promising work, the current study proposes the use of RTs in addition to responses. More specifically, examinees’ RTs are incorporated into the process by simultaneously monitoring any significant decrease in the average RT of each item over repeated exposure. By evaluating abnormal changes in both the number of correct responses as well as the average RTs for items, the procedure has the potential to provide even greater statistical power for detecting compromise as well as stronger substantive evidence that an item is indeed compromised. The efficacy of this enhanced method is investigated in detail.

5.2 Sequential Monitoring Procedures

5.2.1 Using Responses

The goal is to detect a significant increase in the number of correct responses to an item over time, which can be accomplished by periodically comparing the sum of recent responses to a benchmark value that is expected when the item is not compromised. To this end, define a moving sample to be the most recent \( m \) examinees to item \( j \), which gradually isolates potentially compromised responses after a leak. The sum of responses in the moving sample is then calculated as,

\[
Y_j^{(m)} = \sum_{i=n-m+1}^{n} X_{ij}, \tag{5.1}
\]
where the superscript \((m)\) denotes moving sample, \(m\) is the moving sample size, and \(n > m\) is the updated total sample size for item \(j\). Under the null hypothesis that the item is not compromised, \(X_{ij}\) is a Bernoulli random variable with the following expectation and variance:

\[
E(X_{ij}) = P_j(\theta_i), \quad Var(X_{ij}) = P_j(\theta_i)(1 - P_j(\theta_i)). \quad (5.2)
\]

Since \(X_{ij}\)'s are independently but not identically distributed, \(Y_j^{(m)}\) is a Poisson-binomial random variable with the following expectation and variance:

\[
E(Y_j^{(m)}) = \sum_{i=n-m+1}^{n} P_j(\theta_i), \quad Var(Y_j^{(m)}) = \sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i)). \quad (5.3)
\]

Hence, under the null assumption, the following test statistic has an asymptotic standard normal distribution:

\[
\frac{Y_j^{(m)} - \sum_{i=n-m+1}^{n} P_j(\theta_i)}{\sqrt{\sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))}} \to N(0, 1) \text{ under } H_0, \quad (5.4)
\]

Noting that \(\hat{p}_j^{(m)} = Y_j^{(m)}/m\) is a sample proportion, the test statistic can be equivalently expressed as

\[
\frac{\hat{p}_j^{(m)} - \sum_{i=n-m+1}^{n} P_j(\theta_i)/m}{\sqrt{\sum_{i=n-m+1}^{n} P_j(\theta_i)(1 - P_j(\theta_i))/m^2}} \overset{d}{\to} \mathcal{N}(0, 1) \text{ under } H_0, \quad (5.5)
\]

where the null hypothesis, \(H_0 : \hat{p}_j^{(m)} = \sum_{i=n-m+1}^{n} P_j(\theta_i)/m\), is tested against the one-sided alternative hypothesis, \(H_1 : \hat{p}_j^{(m)} > \sum_{i=n-m+1}^{n} P_j(\theta_i)/m\). However, true \(\theta_i\) is never known in reality, so Zhang and Li (2016) approximated the test statistic by substituting with \(\hat{\theta}_i\). This method was shown to be very powerful, but only when ability estimation was relatively uncorrupted by item preknowledge. As an item pool becomes progressively compromised, an examinee would likely have preknowledge of a greater number of administered items. In effect, \(\hat{\theta}_i\)'s would become increasingly positively biased,
thereby inflating $E(Y_j^{(m)})$ and diminishing the power to detect compromise.

As an alternative approach, Zhang (2014) proposed framing the problem as a comparison of two sample proportions. Specifically, the moving sample is compared to a reference sample, which is defined as the first $n - m$ examinees to item $j$. The proportion of correct responses in this complementary sample is then computed as

$$\hat{p}_j^{(r)} = \frac{\sum_{i=1}^{n-m} X_{ij}}{n - m}, \quad (5.6)$$

where the superscript $(r)$ denotes reference sample. $\hat{p}_j^{(r)}$ serves as an appropriate empirical benchmark as long as the item has not been compromised before $n - m$. Thus, the test statistic for two sample proportions can be constructed as

$$\hat{p}_j^{(m)} - \hat{p}_j^{(r)} \quad \sqrt{\hat{p}_j(1 - \hat{p}_j) \left( \frac{1}{m} + \frac{1}{n - m} \right)} \quad \overset{d}{\longrightarrow} \mathcal{N}(0, 1) \quad \text{under } H_0, \quad (5.7)$$

where $H_0 : p_j^{(m)} = p_j^{(r)}$ is tested against $H_1 : p_j^{(m)} > p_j^{(r)}$. Since the true $p_j$ is unknown, the original study substituted with $\hat{p}_j^{(r)}$ in the denominator to approximate the test statistic. Nevertheless, the present study opts to use the more conventional method of estimating $p_j$ by pooling $\hat{p}_j^{(r)}$ and $\hat{p}_j^{(m)}$ as

$$\hat{p}_j = \frac{(n - m)\hat{p}_j^{(r)} + m\hat{p}_j^{(m)}}{(n - m) + m} = \frac{\sum_{i=1}^{n} X_{ij}}{n}, \quad (5.8)$$

which is simply the proportion correct out of all $n$ responses for item $j$. Ultimately, the approximated test statistic is given as

$$Z_j = \frac{\hat{p}_j^{(m)} - \hat{p}_j^{(r)}}{\sqrt{\hat{p}_j(1 - \hat{p}_j) \left( \frac{1}{m} + \frac{1}{n - m} \right)}} = \frac{\hat{p}_j^{(m)} - \hat{p}_j^{(r)}}{\sqrt{\hat{p}_j(1 - \hat{p}_j)/m}} \sqrt{\frac{n - m}{n}}, \quad (5.9)$$

which is used to conduct the test each time the item is administered to a new examinee by comparing it to a chosen critical value, $z_c$. In other words, if $Z_j > z_c$, then $H_0$ is rejected and the item is flagged as compromised since there is evidence that the number of correct responses has increased signif-
icantly. Figure 5.1 illustrates the sequential process of monitoring an item starting at a predesignated exposure point followed by three possible decision scenarios: 1) type I error of flagging an uncompromised item; 2) correct decision of flagging a compromised item, where the number of exposures from the point of compromise (also known as the change point) to point of flag is called the lag; 3) type II error of failing to flag a compromised item by the end of the CAT cycle.

The choice of $z_c$ depends on the desired rate of type I error, $\alpha$, which is complicated by the fact that many items are each being tested over repeated occasions. In other words, multiplicity occurs both between and within items, resulting in different interpretations of $\alpha$ depending on how we define the “family” of tests for which type I error should be controlled. In the simplest case, a “family” consists of a single monitored item on a single occurrence, so $\alpha$ is the probability of incorrectly flagging a given item on any given exposure. In other words, there is a(n) $100(\alpha)\%$ chance of flagging an uncompromised item every time it is tested. This level of error is easily controlled by setting $z_c = \Phi^{-1}(1-\alpha)$, where $\Phi$ is the standard normal CDF. On the other extreme, a “family” could be defined as all monitored items on all occurrences, in which case $\alpha$ is the probability of incorrectly flagging at least once across all items and their exposures for the duration of a given CAT cycle. In other words, we can be $100(1-\alpha)\%$ confident that none of items in the bank will be incorrectly flagged. Determining a precise $z_c$ to control for this level of error is much more difficult due to an unknown degree of dependence between items as well as heavy dependence within items without prior knowledge of exposure counts. Note that the strongest dependence within an item occurs on two consecutive tests, since the latter shares all of the same data with the former except for a single new observation added to the moving sample and the oldest observation in the moving sample transferred to the reference sample. In this study, a “family” is defined more moderately as a single monitored item across all occurrences, so $\alpha$ is the probability of incorrectly flagging an item across all of its exposures. In other words, for a given CAT cycle, we are content with flagging $100(\alpha)\%$ of uncompromised items in the bank. Lacking more convenient analytic methods, Monte Carlo simulations can be conducted to determine $z_c$ for desired values of $\alpha$. 

90
5.2.2 Using Response Times

In general, examinees with preknowledge of an item would be expected to respond quicker than usual. Thus, the goal is to detect a significant decrease in RTs to an item over repeated administrations, which can be accomplished by periodically comparing the average of recent RTs to a benchmark value that is expected when the item is not compromised. This requires a model of RTs that, at the very least, parameterizes the speededness of individual items across examinees. Among a variety of options, the lognormal model (van der Linden, 2006) remains a popular choice for its relative simplicity and practicability for typical RT data.

Recall from Chapter 4 that given the latent speed of the $i$th examinee ($\tau_i$), the density function of RT for the $j$th item ($T_{ij}$) is defined as

$$f(t_{ij}|\tau_i) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}}e^{-\left[\alpha_j(\log t_{ij} - \beta_j + \tau_i)\right]^2/2},$$

(5.10)

where $\alpha_j$ (not to be confused with type I error rate) and $\beta_j$ are respectively the time discrimination and time intensity parameters, and $\tau_i$ and $\beta_j$ are scaled on the same metric. Rewriting the density function in standard form for a lognormal random variable,

$$f(t_{ij}|\tau_i) = \frac{1}{t_{ij}\sqrt{2\pi(1/\alpha_j)^2}}e^{-\left[\log t_{ij} - (\beta_j - \tau_i)\right]^2/[2(1/\alpha_j)^2]},$$

(5.11)

it becomes clear that $\mu_{ij} = \beta_j - \tau_i$ and $\sigma_j^2 = (1/\alpha_j)^2$. In other words, conditional on examinee speed, the log of RT is normally distributed as follows:

$$\log T_{ij}|\tau_i \sim \mathcal{N}[\beta_j - \tau_i, 1/(1/\alpha_j)^2].$$

(5.12)

Hence, a moving sample technique is proposed in which the average log RT of the last $m$ examinees for item $j$ is first computed as

$$\hat{\mu}_j^{(m)} = \frac{1}{m} \sum_{i=n-m+1}^{n} \log T_{ij}.$$  

(5.13)
The expectation and variance of $\hat{\mu}_j^{(m)}$ under the null are

$$E(\hat{\mu}_j^{(m)}) = \frac{1}{m} \sum_{i=n-m+1}^{n} (\beta_j - \tau_i), \quad Var(\hat{\mu}_j^{(m)}) = \frac{1}{m\alpha_j^2}, \quad (5.14)$$

so the following test statistic can be constructed:

$$\frac{\hat{\mu}_j^{(m)} - \sum_{i=n-m+1}^{n} (\beta_j - \tau_i)/m}{(1/\alpha_j)/\sqrt{m}} \xrightarrow{d} N(0, 1) \text{ under } H_0, \quad (5.15)$$

where $H_0 : \mu_j^{(m)} = \sum_{i=n-m+1}^{n} (\beta_j - \tau_i)/m$ is tested against $H_1 : \mu_j^{(m)} < \sum_{i=n-m+1}^{n} (\beta_j - \tau_i)/m$. Although log $T_{ij}$'s are independently but not identically distributed, the asymptotic normality of the test statistic is assured by Lyapunov's CLT (see Appendix B for proof). But since the true $\tau_i$'s are unknown, the test statistic could be approximated by substituting with the MLE's of $\tau_i$ as given in Chapter 4. Just as with ability, however, speed would be routinely overestimated for those with preknowledge of administered items, thereby reducing the power of the test.

To avoid having to determine specific $\tau_i$'s for each item, a general assumption could be made that $\tau_i$ follows a standard normal distribution for every item $j$. Defining $g(\tau_i)$ to be the standard normal density function, it can be shown that the marginal density of RT for item $j$ is

$$f(t_j) = \int_{-\infty}^{\infty} f(t_j | \tau_i) g(\tau_i) d\tau_i = \frac{1}{t_j \sqrt{2\pi(1 + 1/\alpha_j^2)}} e^{-[\log t_j - \beta_j]^2/[2(1+1/\alpha_j^2)]}, \quad (5.16)$$

which is lognormal with $\mu_j = \beta_j$ and $\sigma_j^2 = 1 + 1/\alpha_j^2$. In other words, the marginal distribution of log RT is as follows:

$$\log T_j \sim N[\beta_j, 1 + 1/\alpha_j^2], \quad (5.17)$$

which simplifies the null expectation and variance of $\hat{\mu}_j^{(m)}$ to,

$$E(\hat{\mu}_j^{(m)}) = \beta_j, \quad Var(\hat{\mu}_j^{(m)}) = (1 + 1/\alpha_j^2)/m. \quad (5.18)$$
As a result, the following test statistic can be constructed:

\[
\frac{\hat{\mu}_j^{(m)} - \beta_j}{\sqrt{(1 + 1/\alpha_j^2)/m}} \sim \mathcal{N}(0, 1) \text{ under } H_0,
\]  

(5.19)

where \( H_0 : \mu_j^{(m)} = \beta_j \) is tested against \( H_1 : \mu_j^{(m)} < \beta_j \). Nevertheless, even if it is true that \( \tau_i \) is standard normal in the general population, this convenient formulation only holds when \( \theta_i \) and \( \tau_i \) are independent. Otherwise, BAS would indirectly influence the distribution of \( \tau_i \)'s for an item. For instance, if \( \theta_i \) and \( \tau_i \) are positively correlated, an item with high \( b_j \) would generally be selected for examinees with high \( \theta_i \)'s and in turn higher \( \tau_i \)'s. Consequently, \( \tau_i \)'s for this item would no longer be distributed as standard normal, rendering the above test statistic inaccurate.

Alternatively, an empirical route can be taken in which the moving sample is compared to the reference sample via a two sample means \( t \)-test. The mean of log RTs for the reference sample is

\[
\hat{\mu}_j^{(r)} = \frac{1}{n - m} \sum_{i=1}^{n-m} \log T_{ij},
\]

(5.20)

and the variances of log RTs for the moving and reference samples are

\[
\hat{\sigma}_j^2(m) = \frac{\sum_{i=n-m+1}^{n} (\log T_{ij} - \hat{\mu}_j^{(m)})^2}{m - 1} \quad \text{and} \quad \hat{\sigma}_j^2(r) = \frac{\sum_{i=1}^{n-m} (\log T_{ij} - \hat{\mu}_j^{(r)})^2}{n - m - 1},
\]

(5.21)

respectively. Assuming that \( \sigma_j^2(m) = \sigma_j^2(r) \), the pooled sample variance is

\[
\hat{\sigma}_j^2 = \frac{(m - 1)\hat{\sigma}_j^2(m) + (n - m - 1)\hat{\sigma}_j^2(r)}{n - 2}.
\]

(5.22)

Therefore, the test statistic is given as

\[
W_j = \frac{\hat{\mu}_j^{(m)} - \hat{\mu}_j^{(r)}}{\sqrt{\hat{\sigma}_j^2(1/m + 1/(n-m))}} = \frac{\hat{\mu}_j^{(m)} - \hat{\mu}_j^{(r)}}{\hat{\sigma}_j/\sqrt{m}} \sqrt{\frac{n-m}{m}} \sim T(n-2) \text{ under } H_0,
\]

(5.23)

where \( H_0 : \mu_j^{(m)} = \mu_j^{(r)} \) is tested against \( H_1 : \mu_j^{(m)} < \mu_j^{(r)} \) each time the item
is administered to a new examinee by comparing $W_j$ to a specified critical value, $t_c$. In other words, if $W_j < t_c$, then $H_0$ is rejected and the item is flagged as compromised since there is evidence that the average log RT has dropped significantly. As with $z_c$ when testing proportions, $t_c$ for desired levels of $\alpha$ can be found via Monte Carlo.

5.2.3 Using Responses and Response Times Jointly

The sequential monitoring of responses and RTs, as described above, can be run concurrently but independently as dual univariate (DU) procedures. Within this scheme, define two ways to deem an item compromised:

- **DU-1**: Flag item $j$ if $[(Z_j > z_c) \cap (W_j < 0)] \cup [(Z_j > 0) \cap (W_j < t_c)];$
- **DU-2**: Flag item $j$ if $(Z_j > z_c) \cap (W_j < t_c).$

DU-1 presumes that a significant result for either responses or RTs is sufficient evidence for compromise, as long as the insignificant result is in the direction of $H_1$. On the other hand, DU-2 presumes that significant results for both responses and RTs are necessary to make an informed decision. To avoid the complication of having to determine separate critical values for the response and RT processes, the latter can just be set as $t_c = -z_c$.

Alternatively, responses and RTs can be monitored simultaneously within a single multivariate (SM) framework, which accounts for the possible dependence between responses and RTs. Dropping the subscript $j$ to reduce notational clutter, define the following moving sample statistics for item $j$: $\hat{\mu}_1^{(m)} = \hat{p}^{(m)}$ is the mean of responses (i.e., proportion of correct responses), $\hat{\mu}_2^{(m)}$ is the mean of log RTs, $\hat{\sigma}_1^{2(m)}$ is the variance of responses, $\hat{\sigma}_2^{2(m)}$ is the variance of log RTs, and $\hat{\sigma}_{12}^{(m)}$ is the covariance between responses and log RTs. Unbiased estimators are used in all cases, including the sample variance of responses: $\hat{\sigma}_1^{2(m)} = \hat{p}^{(m)}(1 - \hat{p}^{(m)})(m/(m - 1))$. Thus, the estimated mean vector and covariance matrix for a moving sample can be specified as

$$\hat{\mu}^{(m)} = \begin{bmatrix} \hat{\mu}_1^{(m)} \\ \hat{\mu}_2^{(m)} \end{bmatrix} \quad \text{and} \quad \hat{\Sigma}^{(m)} = \begin{bmatrix} \hat{\sigma}_1^{2(m)} & \hat{\sigma}_{12}^{(m)} \\ \hat{\sigma}_{12}^{(m)} & \hat{\sigma}_2^{2(m)} \end{bmatrix},$$

(5.24)
respectively. Likewise, for the reference sample,

\[
\hat{\mu}^{(r)} = \begin{bmatrix} \hat{\mu}_1^{(r)} \\ \hat{\mu}_2^{(r)} \end{bmatrix} \quad \text{and} \quad \hat{\Sigma}^{(r)} = \begin{bmatrix} \hat{\sigma}_1^{2(r)} & \hat{\sigma}_{12}^{(r)} \\ \hat{\sigma}_{12}^{(r)} & \hat{\sigma}_2^{2(r)} \end{bmatrix}.
\] (5.25)

Although the joint distribution of responses and RTs is clearly not normal, the asymptotic bivariate normality of the mean vectors can be inferred by the multivariate CLT. Therefore, computing the unbiased pooled covariance matrix as

\[
\hat{\Sigma} = \frac{m-1}{n-2} \hat{\Sigma}^{(m)} + \frac{n-m-1}{n-2} \hat{\Sigma}^{(r)},
\] (5.26)

the two-sample Hotelling’s \( T^2 \) statistic can be constructed as

\[
T^2 = \left[ \hat{\mu}^{(m)} - \hat{\mu}^{(r)} \right]' \left[ \frac{1}{m} + \frac{1}{n-m} \right]^{-1} \left[ \hat{\mu}^{(m)} - \hat{\mu}^{(r)} \right],
\] (5.27)

which is approximately related to the \( F \)-distribution as follows:

\[
F_{\text{obs}} = \frac{n-3}{2(n-2)} T^2 \sim F(2, n-3) \text{ under } H_0.
\] (5.28)

The null hypothesis, \( H_0 : \mu^{(m)} = \mu^{(r)} \), is tested against the directional alternative hypothesis, \( H_1 : \mu_1^{(m)} > \mu_1^{(r)} \& \mu_2^{(m)} < \mu_2^{(r)} \), after each item exposure until significance is reached. In other words, an item is flagged as compromised if \( F_{\text{obs}} > F_c \), provided that \( \hat{p}^{(m)} > \hat{p}^{(r)} \) and \( \hat{\mu}_2^{(m)} < \hat{\mu}_2^{(r)} \). The imposed constraints ensure that the specific directionality of the test is achieved, and the critical value \( F_c \) can be determined for any level of \( \alpha \) through Monte Carlo. Note that the conventional Hotelling’s \( T^2 \) test with the non-directional alternative, \( H_1 : \mu^{(m)} \neq \mu^{(r)} \), would be inefficient in this context.

5.3 Method

5.3.1 Data

The sequential monitoring procedures were evaluated through simulations based on real data from a high-stakes, large-scale standardized CAT. The data consisted of raw responses and RTs (in minutes) from about 2,000 ex-
aminees with an item pool of about 500 items whose 3PLM parameters were already estimated. The lognormal model parameters were calibrated under the two-level hierarchical framework (van der Linden, 2007), which accounts for the relationship between accuracy and speed. The first level consisted of the 3PL and lognormal models, and the second level specified the covariance structure between the person parameters ($\theta_i$, $\tau_i$) and among the item parameters ($a_j$, $b_j$, $c_j$, $\alpha_j$, $\beta_j$). Note that this modeling framework disregards the classic within-person speed-accuracy tradeoff, or in other words, the compromise between $\theta_i$ and $\tau_i$ within an individual examinee during the course of the test. Instead, a reasonable assumption is made that an individual’s latent parameters remain constant as long as the test is not unduly speeded. Ultimately, $\alpha_j$, $\beta_j$, $\theta_i$, and $\tau_i$ were estimated using a modified version of van der Linden’s (2008) MCMC routine that fixed $a_j$, $b_j$, and $c_j$ to the pre-calibrated values and centered the distribution of $\tau_i$ at 0. Every parameter estimate appeared to converge using 10,000 MCMC draws with a burn-in size of 5,000, and the overall hierarchical framework seemed to fit well enough for the current application. Therefore, all estimates from this calibration step were regarded as the true parameter values when simulating CAT.

5.3.2 Simulation Design

Therefore, in the interest of stronger test security and better item pool usage, the BAS method was employed in the present investigation of utilizing RTs in the sequential detection of compromised items. Additionally, ability estimation was performed with a combination approach, in which EAP was used as a provisional fail-safe whenever an infeasibility occurred with MLE. In contrast, the original sequential detection study by Zhang (2014) implemented MFI-SH with exclusive EAP estimation, and the follow-up study by Zhang and Li (2016) used a shadow test engine with all interim estimates in EAP and the final estimates in MLE. The shadow test methodology is not discussed for brevity.

The CAT system was built upon the BAS item selection algorithm, with the item pool divided into 5 strata of about 100 items each. Fixing the test length at 30 items, the first 5 were chosen randomly from each stratum in
order to calculate initial estimates of $\theta_i$ and $\tau_i$, then subsequent items were selected using the $b$-matching criterion at each of the five stages. Additionally, the maximum exposure rate was set at 0.2 to ensure a relatively balanced usage of items even under extreme simulation conditions. Based on the true parameters, the $i$th examinee’s response to the $j$th administered item was randomly generated in real-time from the Bernoulli distribution with success probability $P_j(\theta_i)$; likewise, response time was randomly generated from $\log N(\beta_j - \tau_i, 1/\alpha_j^2)$.

There are two broad manifestations of item compromise in CAT: 1) a particular subset of examinees gaining preknowledge of a particular subset of the item pool (e.g., a group of colluders sharing stolen items); and 2) certain items leaking to the general public (e.g., overexposed items spreading through word of mouth or discussions in online forums). In the former situation, all conspirators would be expected to have preknowledge of every compromised item they are administered. The latter situation is the focus of this study, in which every examinee is assumed to be a potential beneficiary of a compromised item with stationary probability $\psi$. The preknowledge distribution of responses to any compromised item was modeled as

$$P^*(X = x) = 0.999^x \cdot 0.001^{(1-x)} \iff X \sim \text{Bernoulli}(0.999), \quad (5.29)$$

which specifies a correct response with near but not absolute certainty to allow for inadvertent mistakes by even those with preknowledge. Also, the preknowledge distribution of RTs (in minutes) on any compromised item was modeled as

$$f^*(t_{ij}) = \frac{3.5}{t_{ij}\sqrt{2\pi}} e^{-3.5^2(\log t_{ij} + 2)/2} \iff \log T \sim N(-2, 1/3.5^2), \quad (5.30)$$

which specifies a reasonable range from about 2 to 30 seconds with a mean of about 8.5 seconds. Therefore, responses and RTs to an item, from the point of compromise onward, follow the preknowledge distributions with probability $\psi$ and the regular distributions with probability $1 - \psi$, which can be expressed in terms of mixture distributions as follows:

$$\tilde{P}_j(\theta_i) = \psi P^*(X_{ij} = 1) + (1 - \psi) P_j(\theta_i), \quad (5.31)$$
\[
\tilde{f}_j(t_{ij}|\tau_i) = \psi f^*(t_{ij}) + (1 - \psi)f_j(t_{ij}|\tau_i).
\]

The monitoring process was set to start for every item at the 40th exposure using a moving sample size of \(m\). For instance, using \(m = 10\), the moving and reference samples of the initial test would consist of the last 10 and first 30 examinees to have been administered the item, respectively. A random quarter of the item pool (about 125 items) were queued to be compromised, each starting at a randomized exposure count between 40 and 100. Any examinee administered a compromised item had preknowledge with a designated probability of \(\psi\). Defining \(C\) as the set of all compromised items and \(F\) as the set of all flagged items, type I error rate and power were estimated as

\[
P(\text{Type I Error}) \approx P(F|C') = \frac{P(F \cap C')}{P(C')} = \frac{|F \cap C'|}{|C'|},
\]

\[
\text{Power} \approx P(F|C) = \frac{P(F \cap C)}{P(C)} = \frac{|F \cap C|}{|C|}.
\]

If an item in \(C\) was prematurely flagged before the designated change point, it was moved to the uncompromised set \(C'\) and counted as a type I error. Any flagged item, whether or not in error, was recorded but otherwise kept operational in the item pool. Additionally, the average lag \(\bar{L}\) from the change point \(l_j\) to flag point \(n_j\) for the set of correctly flagged items \((F \cap C)\) was calculated as

\[
\bar{L} = \frac{\sum_{j \in F \cap C} (n_j - l_j)}{|F \cap C|}
\]

to evaluate how quickly compromised items could be detected on average.

The performances of the sequential monitoring procedures were comparatively evaluated on these three criteria instead of the average run length (ARL) that is commonly utilized in conventional change-point detection. The reason is simply that ARL assumes that the sequential process continues ad infinitum until a significant change is detected, which is clearly not the case in CAT due to a finite number of examinees. Higher \(\psi\) is expected to yield greater power at a given type I error rate, since a higher prevalence of preknowledge makes it easier to detect. Likewise, smaller \(m\) is expected to yield shorter lag at a given type I error rate, since a smaller moving sample retains less older data that may act as dead weight.
5.4 Results

The first set of simulations compared the performances of the five monitoring schemes: responses alone (R), RTs alone (T), dual univariate 1 (DU-1), dual univariate 2 (DU-2), and single multivariate (SM). Every technique was evaluated on each of 2 sample sizes ($m = 5, 20$) at each of 3 preknowledge probabilities ($\psi = 0.15, 0.25, 0.35$) for a total of 6 conditions. The results, which were averaged across 100 replications, are presented as receiver operating characteristic (ROC) curves in Figure 5.2 and lag plots in Figure 5.3. The most salient observation is that the performances of T, DU-1, and SM were all nearly identical with the highest power and lowest lag at any given type I error rate. On the contrary, R was worst by far and DU-2 was somewhere in the middle in terms of general performance. In other words, R and T were effectively the lower and upper performance baselines, respectively, indicating that preknowledge RT’s were much easier to detect than preknowledge responses. Consequently, DU-1 and SM were overwhelmingly dominated by RTs, while DU-2 was evenly influenced by both responses and RTs. Moreover, for every procedure, lag was shorter for higher $\psi$ and smaller $m$, and power was greater for higher $\psi$ regardless of $m$ as expected. However, a closer look at the ROC curves reveals an interesting pattern: power was greater for larger $m$ at $\psi = 0.35$, very similar for both $m = 5$ and 20 at $\psi = 0.25$, and actually greater for smaller $m$ at $\psi = 0.15$. This suggested an interaction between $\psi$ and $m$, which warranted a follow-up study.

The second set of simulations compared the performances of 5 moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of 6 preknowledge probabilities ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$) exclusively for SM. As before, the results were averaged across 100 replications and presented as ROC curves in Figure 5.4 and lag plots in Figure 5.5. The particular interaction effect becomes quite noticeable here: $\psi$ strongly moderated the effect of $m$ on power at any given type I error rate. For $\psi < 0.25$, smaller $m$ resulted in greater power, with larger differences in effect for lower $\psi$; at $\psi = 0.25$, $m$ had no appreciable effect on power; for $\psi > 0.25$, larger $m$ resulted in greater power, with larger differences in effect for higher $\psi$. This phenomenon occurs because when $\psi$ is very low, there is a dearth of preknowledge responses and RTs. As a result, a smaller moving sample can more easily isolate them, thereby increasing power even at the cost of larger sampling error. In the
current context, $\psi = 0.25$ happened to be the point of equilibrium at which the opposing forces of preknowledge isolation and sampling error balanced out to the same power for every $m$. Also, for $\psi > 0.25$, there were negligible improvements in power for $m$ greater 5, most likely due to the ceiling effect. On the other hand, moderator effects were not observed for lag. Just as in the earlier results, lag was always shorter for smaller $m$ and higher $\psi$.

5.5 Discussion

The promising results demonstrate that response times can be effectively utilized in conjunction with responses to improve the sequential detection of compromised items. Both DU-1 and SM were shown to be equally superior over DU-2 in detection accuracy and speed. Nevertheless, SM has two distinct advantages over DU-1: First, SM is easier to implement since only a single process needs to be tracked as opposed to two separate streams. Second, SM can be seen as a more holistic approach that combines all information into a single evidentiary criterion instead of cherry-picking the favorable outcome. Choosing an appropriate moving sample size is a trickier matter, since it depends on the unknown probability that a random examinee has preknowledge of any given compromised item. Because the optimal $m$ is most likely unique for every CAT, it must be determined by the user through a series of simulations. This can be accomplished by first finding the equilibrium point, $\psi_e$. If true $\psi$ is believed to be less than $\psi_e$, use $m = 2$ for best results; otherwise, choose the largest $m$ beyond which there seem to be insubstantial improvements in power. Once $m$ is determined, an item can be monitored as soon as $n = m + 2$.

At this point, a word of caution regarding the interpretation of power would be prudent. It may be tempting to interpret power as the probability that a flagged item is compromised, or $P(C|F)$, which would be of primary interest in practice. However, doing so would be committing an inverse fallacy, recalling that power is actually the probability that a compromised item is flagged, or $P(F|C)$. Succinctly, Power = $P(F|C) \neq P(C|F)$; instead,
we properly apply Bayes’ theorem to obtain

\[ P(C|F) = \frac{P(F|C)P(C)}{P(F|C)P(C) + P(F|C')P(C')} = \frac{\text{Power} \times P(C)}{[\text{Power} \times P(C)] + [\alpha \times (1 - P(C))]} \]  

(5.36)

Note that \( P(C) \) is the base rate of item compromise, which is typically unknown. Nevertheless, to illustrate the substantial impact of the base rate, say we have 90% power at 5% type I error but the base rate is relatively low at 5.5%. Then, there is only about a 50% chance that a flagged item is actually compromised even with such high power. Although somewhat discouraging, this is a typical phenomenon in diagnostic testing in general, such as in medical screening for a rare disease. As with any such tool, the sequential detection procedures should be utilized responsibly, preferably with corroborating evidence of compromise.

There are several issues that have not been addressed in this study. First, only one form of item compromise was considered, namely a set of items leaking to the general public in which any given test-taker may gain pre-knowledge of an item. As mentioned earlier, another form of compromise is a particular group of conspirators sharing a specific set of items, leading to their preknowledge of these stolen items with certainty. It is possible that either or both forms of compromise can occur during a CAT cycle. Second, the particular lognormal distribution used to model preknowledge RTs is certainly plausible and suitable for the purposes of this study, but it is admittedly an uninformed choice. Currently, no empirically supported alternatives have been proposed in literature, most likely due to the difficulty of obtaining real RT data from examinees verified to have item preknowledge. Third, sequential monitoring assumes that the general characteristics of the examinee population are consistent over the course of item usage. As such, the simulations did not consider scenarios of either responses or RTs becoming aberrant for reasons unrelated to item compromise. For example, significant changes in response patterns may occur due to a benign cause of item parameter drift over time, or there could be a sudden shift in the demographics of test-takers leading to potential DIF. Acknowledging these various circumstances, it would be worthwhile to extend the simulations to reflect the more complex reality. Moreover, empirical studies need to be conducted to assess the applicability and efficacy of the procedures in practice.
5.6 Figures

Figure 5.1: An illustration of the sequential monitoring process with three possible decision scenarios: type I error, correct flag, and type II error.
Figure 5.2: ROC curves for each of the five sequential procedures (R, T, DU-1, DU-2, SM) across six conditions ($m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 5.3: Lag plots for each of the five sequential procedures (R, T, DU-1, DU-2, SM) across six conditions \((m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\})\). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 5.4: ROC curves for the SM procedure with five moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of six levels of item preknowledge ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 5.5: Lag plots for the SM procedure with five moving sample sizes ($m = 2, 5, 10, 20, 30$) at each of six levels of item preknowledge ($\psi = 0.05, 0.10, 0.15, 0.25, 0.35, 0.45$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
CHAPTER 6

SEQUENTIAL DETECTION OF COMPROMISED ITEMS USING RESPONSE TIMES IN CAT (PART 2)

6.1 Introduction

In the previous chapter, the detection of item compromise was formulated as a sequence of Wald tests comparing the moving and reference groups. In particular, the Hotelling’s $T^2$ statistic was highlighted as a way to combine both response and RT data into a single multivariate test. As a direct continuation of the study, three additional statistical methods are examined:

1) extension of comparing two proportions, including binomial and Fisher’s exact tests;
2) generalized likelihood ratio test (GLRT);
3) nonparametric techniques comparing empirical distribution functions (EDFs), specifically the Kolmogorov-Smirnov (KS) and Kuiper’s tests.

6.2 Additional Sequential Methods

6.2.1 Two-Proportion Tests

Recall that the original method by Zhang (2014) compares the proportion correct of the moving sample ($p_j^{(m)}$) to that of the reference sample ($p_j^{(r)}$), which is formulated as an approximate $z$-test given in Equation 5.9. The basic premise is that if item $j$ becomes compromised, then $p_j^{(m)} > p_j^{(r)}$ since the moving sample would be contaminated with preknowledge responses. Extending this premise to incorporate RTs, define “quicker than average” to be a log RT less than the reference mean: $\log T_{ij} < \hat{\mu}_j^{(r)}$. Next, define $X_{ij}^*$ as follows:

$$X_{ij}^* = I(T_{ij})X_{ij},$$

(6.1)
where \( I \) is an indicator function that maps \( T_{ij} \) to 1 if \( \log T_{ij} < \hat{\mu}_j^{(r)} \), 0 otherwise. Then the number of correct responses that are quicker than average in the moving and reference samples are, respectively,

\[
K_j^{(m)} = \sum_{i=n-m+1}^{n} X_{ij}^* \quad \text{and} \quad K_j^{(r)} = \sum_{i=1}^{n-m} X_{ij}^*,
\]

(6.2)

and the corresponding proportions are

\[
\hat{q}_j^{(m)} = \frac{K_j^{(m)}}{m} \quad \text{and} \quad \hat{q}_j^{(r)} = \frac{K_j^{(r)}}{n-m}.
\]

(6.3)

The idea is that \( \hat{q}_j^{(m)} > \hat{q}_j^{(r)} \) if item \( j \) becomes compromised, which can be sequentially tested against \( H_0 : \hat{q}_j^{(m)} = \hat{q}_j^{(r)} \) using different approaches.

First, analogous to the original method, the normal approximation is considered as follows:

\[
Z_j^* = \frac{\hat{q}_j^{(m)} - \hat{q}_j^{(r)}}{\sqrt{\hat{q}_j^{(1)} \left( \frac{1}{m} + \frac{1}{n-m} \right) / n}},
\]

(6.4)

where

\[
\hat{q}_j = \frac{(n-m)\hat{q}_j^{(r)} + m\hat{q}_j^{(m)}}{(n-m) + m} = \frac{\sum_{i=1}^{n} X_{ij}^*}{n}.
\]

(6.5)

Second, the binomial test is considered, in which the \( p \)-value is formulated as:

\[
P(K \geq k) = \sum_{k=K_j^{(m)}}^{m} \binom{m}{k} \left( \hat{q}_j^{(r)} \right)^k \left( 1 - \hat{q}_j^{(r)} \right)^{m-k}.
\]

(6.6)

Note that \( \hat{q}_j^{(r)} \) is being used as an estimate for the population proportion under the null assumption of no compromise. Third, Fisher’s exact test is considered by conceptualizing the data as the following \( 2 \times 2 \) contingency table:

<table>
<thead>
<tr>
<th>( X_{ij}^* = 1 )</th>
<th>Moving</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>( K_j^{(m)} + K_j^{(r)} - k )</td>
<td>( K_j^{(m)} + K_j^{(r)} )</td>
</tr>
<tr>
<td>( m - k )</td>
<td>( n - m - K_j^{(m)} - K_j^{(r)} + k )</td>
<td>( n - K_j^{(m)} - K_j^{(r)} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X_{ij}^* = 0 )</th>
<th>Moving</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
<td>( n - m - K_j^{(m)} - K_j^{(r)} + k )</td>
<td>( n - K_j^{(m)} - K_j^{(r)} )</td>
</tr>
</tbody>
</table>
Note that all values are constant except for $k$, meaning the marginal counts are assumed to be fixed (as opposed to unconditional tests such as Barnard’s or Boschloo’s), which allows for computing the $p$-value as

$$P(K \geq k) = \sum_{k=k_{j}^{(m)}}^{m} \frac{(K_{j}^{(m)} + K_{j}^{(r)})}{k} \left( \frac{n - K_{j}^{(m)} - K_{j}^{(r)}}{m - k} \right). \quad (6.7)$$

As before, using the nominal $\alpha$ would lead to an inflated Type I error rate due to multiplicity, so the critical value or cutoff $p$-value for a desired $\alpha$ must be determined via simulations.

### 6.2.2 Generalized Likelihood Ratio Test

In order to utilize GLRT, a modeling paradigm that relates RTs to responses is required. Taking advantage of the fact that responses are simply Bernoulli random variables, the general distribution of RTs for item $j$ across examinees can be modeled as conditional on either a correct or incorrect response: $f(t_{ij}|X_{ij} = 1)$ and $f(t_{ij}|X_{ij} = 0)$, respectively. Therefore, invoking the law of total probability, the marginal distribution of RTs is

$$f(t_{ij}) = f(t_{ij}|X_{ij} = 1)P(X_{ij} = 1) + f(t_{ij}|X_{ij} = 0)P(X_{ij} = 0). \quad (6.8)$$

In other words, $f(t_{ij})$ can be specified as a mixture density in which $f_{1}(t_{ij}) = f(t_{ij}|X_{ij} = 1)$ and $f_{0}(t_{ij}) = f(t_{ij}|X_{ij} = 0)$ are the mixture components and $p_{j} = P(X_{ij} = 1)$ and $(1 - p_{j}) = P(X_{ij} = 0)$ are the corresponding mixture weights.

Next, assuming a lognormal distribution for each of the conditional RTs (i.e., $f_{1}(t_{ij})$ and $f_{0}(t_{ij})$ are lognormal density functions):

$$\log(T_{ij}|X_{ij} = 1) \sim \mathcal{N}(\mu_{1j}, \sigma_{1j}^{2}); \quad (6.9)$$
$$\log(T_{ij}|X_{ij} = 0) \sim \mathcal{N}(\mu_{0j}, \sigma_{0j}^{2}), \quad (6.10)$$

the expectation and variance of the marginal RTs are, respectively,

$$E(\log T_{ij}) = \mu_{j} = p_{j}\mu_{1j} + (1 - p_{j})\mu_{0j}; \quad (6.11)$$
\[ \text{Var}(\log T_{ij}) = \sigma_j^2 = p_j \left( (\mu_{1j} - \mu_j)^2 + \sigma_{1j}^2 \right) + (1 - p_j) \left( (\mu_{0j} - \mu_j)^2 + \sigma_{0j}^2 \right). \] 

(6.12)

Thus, given \( \mathbf{v}_j = \{x_{1j}, t_{1j}\}, \ldots, \{x_{nj}, t_{nj}\} \), the likelihood function is expressed as

\[
L(p_j, \mu_j, \sigma_j^2 | \mathbf{v}_j) = L(p_j, \mu_{1j}, \mu_{0j}, \sigma_{1j}^2, \sigma_{0j}^2 | \mathbf{v}_j) = \prod_{i=1}^{n} \left[ p_j f_1(t_{ij} | \mu_{1j}, \sigma_{1j}^2) + (1 - p_j) f_0(t_{ij} | \mu_{0j}, \sigma_{0j}^2) \right].
\]

(6.13)

The goal is to sequentially test \( H_0 : p_j^{(m)} = p_j^{(r)} \) & \( \mu_j^{(m)} = \mu_j^{(r)} \) against \( H_1 : p_j^{(m)} > p_j^{(r)} \) & \( \mu_j^{(m)} < \mu_j^{(r)} \), where the respective parameter spaces are:

\[
\Omega_{H0} = \left\{ (p_j, \mu_{1j}, \mu_{0j}, \sigma_{1j}^2, \sigma_{0j}^2) : 0 \leq p_j \leq 1, (\mu_{1j}, \mu_{0j}) \in \mathbb{R}, (\sigma_{0j}^2, \sigma_{1j}^2) > 0 \right\};
\]

(6.14)

\[
\Omega_{H1} = \left\{ (p_j^{(m)}, \mu_{1j}^{(m)}, \mu_{0j}^{(m)}, \sigma_{1j}^{2(m)}, \sigma_{0j}^{2(m)}, p_j^{(r)}, \mu_{1j}^{(r)}, \mu_{0j}^{(r)}, \sigma_{1j}^{2(r)}, \sigma_{0j}^{2(r)}) : 0 \leq p_j^{(r)} < p_j^{(m)} \leq 1, \mu_j^{(m)} < \mu_j^{(r)}, (\sigma_j^{2(m)}, \sigma_j^{2(r)}) > 0 \right\}.
\]

(6.15)

Therefore, the generalized likelihood ratio can be constructed as

\[
\Lambda(\mathbf{v}_j) = \frac{\sup\{L(\omega | \mathbf{v}_j) : \omega \in \Omega_{H0}\}}{\sup\{L(\omega' | \mathbf{v}_j) : \omega' \in \Omega_{H1}\}} = \frac{L(\hat{\mu}_{1j}, \hat{\mu}_{0j}, \hat{\sigma}_{1j}^2, \hat{\sigma}_{0j}^2 | \mathbf{v}_j)}{L(p_j^{(m)}, \hat{\mu}_{1j}^{(m)}, \hat{\mu}_{0j}^{(m)}, \hat{\sigma}_{1j}^{2(m)}, \hat{\sigma}_{0j}^{2(m)} | \mathbf{v}_j^{(m)}) L(p_j^{(r)}, \hat{\mu}_{1j}^{(r)}, \hat{\mu}_{0j}^{(r)}, \hat{\sigma}_{1j}^{2(r)}, \hat{\sigma}_{0j}^{2(r)} | \mathbf{v}_j^{(r)})}.
\]

(6.16)

in which the maximum likelihoods are generally computed as

\[
L(\hat{\omega} | \mathbf{v}_j) = \prod_{i=1}^{n} [\hat{p}_j f_1(t_{ij} | \hat{\omega}) + (1 - \hat{p}_j) f_0(t_{ij} | \hat{\omega})].
\]

(6.17)

The MLE’s of the parameters are given in Table 6.1. Also, note that if \( p_j^{(m)} \leq p_j^{(r)} \) or \( \mu_j^{(m)} \geq \mu_j^{(r)} \), then \( \Lambda(\mathbf{v}_j) = 1 \). In the conventional non-sequential
context, GLRT can be carried out by leveraging Wilk’s theorem,

$$-2 \log \Lambda(v_j) \xrightarrow{d} \chi^2(\nu),$$  \hspace{1cm} (6.18)

where $\nu = 5$ since five extra parameters are estimated in the alternative model. In the current sequential application, however, the critical value for a desired $\alpha$ must be determined by simulation.

6.2.3 Empirical Distribution Function Tests

The Wald and LR-based tests are parametric methods that assume an underlying model for the observed data whose parameters can be estimated. Especially in the case of GLRT, the RTs were strictly assumed to follow a mixture lognormal distribution, which may not be the case in reality. Hence, it may be beneficial to consider nonparametric techniques as alternatives, particularly tests based on the EDF of the observed data.

Given the general mixture density of RTs in Equation 6.8, the corresponding CDF is

$$F(t_{ij}) = pF_1(t_{ij}) + (1 - p)F_0(t_{ij})$$ \hspace{1cm} (6.19)

where $F_1(t_{ij}) = F(t_{ij}|X_{ij} = 1)$ and $F_0(t_{ij}) = F(t_{ij}|X_{ij} = 0)$. Defining the following sets:

$$\mathcal{R}_1 = \{i : X_{ij} = 1, 1 \leq i \leq (n - m)\};$$ \hspace{1cm} (6.20)
$$\mathcal{R}_0 = \{i : X_{ij} = 0, 1 \leq i \leq (n - m)\};$$ \hspace{1cm} (6.21)
$$\mathcal{M}_1 = \{i : X_{ij} = 1, (n - m + 1) \leq i \leq n\};$$ \hspace{1cm} (6.22)
$$\mathcal{M}_0 = \{i : X_{ij} = 0, (n - m + 1) \leq i \leq n\},$$ \hspace{1cm} (6.23)

the EDFs of the moving and reference samples can be constructed as, respec-
tively,

\[
\hat{F}(r)(t) = \frac{\hat{p}(r)}{|R_1|} \sum_{i \in R_1} I_{(\leq t)}(T_{ij}) + \frac{1 - \hat{p}(r)}{|R_0|} \sum_{i \in R_0} I_{(\leq t)}(T_{ij}) \\
= \frac{|R_1|/(n - m)}{|R_1|} \sum_{i \in R_1} I_{(\leq t)}(T_{ij}) + \frac{|R_0|/(n - m)}{|R_0|} \sum_{i \in R_0} I_{(\leq t)}(T_{ij}) \tag{6.24}
\]

\[
= \frac{1}{n - m} \left( \sum_{i \in R_1} I_{(\leq t)}(T_{ij}) + \sum_{i \in R_0} I_{(\leq t)}(T_{ij}) \right)
\]

\[
= \frac{1}{n - m} \sum_{i=1}^{n-m} I_{(\leq t)}(T_{ij}),
\]

\[
\hat{F}(m)(t) = \frac{\hat{p}(m)}{|M_1|} \sum_{i \in M_1} I_{(\leq t)}(T_{ij}) + \frac{1 - \hat{p}(m)}{|M_0|} \sum_{i \in M_0} I_{(\leq t)}(T_{ij}) \\
= \frac{|M_1|/m}{|M_1|} \sum_{i \in M_1} I_{(\leq t)}(T_{ij}) + \frac{|M_0|/m}{|M_0|} \sum_{i \in M_0} I_{(\leq t)}(T_{ij}) \tag{6.25}
\]

\[
= \frac{1}{m} \left( \sum_{i \in M_1} I_{(\leq t)}(T_{ij}) + \sum_{i \in M_0} I_{(\leq t)}(T_{ij}) \right)
\]

\[
= \frac{1}{m} \sum_{i=n-m+1}^{n} I_{(\leq t)}(T_{ij})
\]

where \( I_{(\leq t)}(T_{ij}) \) is an indicator function that equals 1 if \( T_{ij} \leq t \) and 0 otherwise. In other words, \( \hat{F}(r)(t) \) and \( \hat{F}(m)(t) \) reduce to the EDFs of RTs independent of the responses, so the responses must be modeled separately:

\[
\hat{F}(r)(x) = \frac{1}{n - m} \sum_{i=1}^{n-m} I_{(\leq x)}(X_{ij}); \tag{6.26}
\]

\[
\hat{F}(m)(x) = \frac{1}{m} \sum_{i=n-m+1}^{n} I_{(\leq x)}(X_{ij}). \tag{6.27}
\]

Note that since \( X_{ij} \) is binary, the EDFs take on just the following values:

\[
\hat{F}(r)(0) = 1 - \hat{p}(r), \quad \hat{F}(r)(1) = 1; \tag{6.28}
\]

\[
\hat{F}(m)(0) = 1 - \hat{p}(m), \quad \hat{F}(m)(1) = 1. \tag{6.29}
\]
The two-sample KS test statistic for RTs is computed as
\[ D_t = \sup_t \left| \hat{F}^{(m)}(t) - \hat{F}^{(r)}(t) \right|, \] (6.30)
which is essentially the maximum vertical distance between the EDFs. Although KS is technically for continuous distributions, the corresponding statistic for responses is simply the absolute difference in sample proportions:
\[ D_x = \sup_x \left| \hat{F}^{(m)}(x) - \hat{F}^{(r)}(x) \right| = \left| \hat{p}^{(r)} - \hat{p}^{(m)} \right|. \] (6.31)
Therefore, the following combined KS-type statistic is proposed:
\[ D = D_t + D_x. \] (6.32)

Additionally, the Kuiper’s statistic is a slight modification of the KS that takes the sum of the maximum positive and negative distances between the EDFs. For RTs, this is given as
\[ V_t = D_t^+ + D_t^- = \sup_t \left( \hat{F}^{(m)}(t) - \hat{F}^{(r)}(t) \right) + \sup_t \left( \hat{F}^{(r)}(t) - \hat{F}^{(m)}(t) \right), \] (6.33)
whereas for responses, the Kuiper’s is equivalent to KS:
\[ V_x = D_x^+ + D_x^- = D_x. \] (6.34)
Therefore, the following combined Kuiper’s-type statistic is also proposed:
\[ V = V_t + V_x. \] (6.35)

Both \( D \) and \( V \) are compared against respective critical values that correspond to a desired \( \alpha \), with directional conditions of \( \hat{p}_j^{(m)} > \hat{p}_j^{(r)} \) and \( \hat{\mu}_j^{(m)} < \hat{\mu}_j^{(r)} \) in order reject \( H_0 \). In general, the KS test should be more adept at detecting shifts in distributions, which is most noticeable at the median of the EDFs. On the other hand, the Kuiper’s test should be more suitable for detecting differences in distributional spread, which is most noticeable at the tails of the EDFs.
6.3 Data and Simulation

The data and general simulation setup are identical to the studies in Chapter 5, including the CAT design, examinees and item pool, response and RT models, and number of replications. The focus is on comparing the new methods with the R and SM procedures at predetermined moving sample sizes \((m = 5, 20)\) and probabilities of item preknowledge \((\psi = 0.15, 0.25, 0.35)\).

6.4 Results

The first set of simulations compared the performances of the two-proportion tests: binomial, Fisher, and normal. The results are displayed as ROC curves and lag plots in Figures 6.1 and 6.2, respectively. For all three tests, three general observations can be made at any given type I error rate \((\alpha)\): 1) greater power for larger \(\psi\) and \(m\), 2) shorter lag for larger \(\psi\), and 3) no substantial change in lag for different \(m\). More interestingly, Fisher and normal had nearly identical power and lag across all conditions. On the other hand, compared to Fisher and normal, binomial always had lower power at a given \(\alpha\), but it also had slightly shorter lag under certain limited conditions. Therefore, the Fisher and normal seem to be better methods than the binomial.

The second, broader set of simulations compared the performances of R, SM, Fisher, GLRT, KS, and Kuiper. The results are presented as ROC curves and lag plots in Figures 6.3 and 6.4, respectively. Overall, at any given \(\alpha\), KS outperformed Kuiper, Fisher, and R across most conditions. Nevertheless, KS came at a distant third compared to SM and GLRT, which were the clear standouts in terms of both power and lag. Specifically, at any given \(\alpha\) and \(\psi\), SM had the greatest power and shortest lag with \(m = 5\), while GLRT had the greatest power and shortest lag with \(m = 20\). Figure 6.5 shows the direct comparison of ROC curves and lag plots between these two best performers. At any given \(\alpha\) and \(\psi\), GLRT with \(m = 20\) had slightly greater power and nearly equivalent lag compared to SM with \(m = 5\).
6.5 Discussion

Considering all of the proposed sequential techniques for detecting compromised items, GLRT with a larger moving sample performed the best, while SM with a smaller moving sample came a close second. In particular, these two techniques were demonstrated to be quite capable of achieving high power with low type I error and short lag, even with only a small chance of an examinee having preknowledge of a compromised item.

Great care was taken to design a reasonably realistic and broad simulation, so the results and implications are expected to be fairly generalizable. However, the strong performances of SM and GLRT might be attributable, at least in part, to generating log RTs from the normal distribution and correctly modeling them as such for the likelihood ratios. Therefore, a future study should examine the robustness of these techniques to non-normality and model misspecification, particularly in comparison to the nonparametric techniques such as Fisher and KS. Additionally, the two-sample proportion tests could possibly be improved by defining “quick” RT differently. For example, “quicker than median” could be used instead of “quicker than average”, so that the indicator function $I$ is redefined as mapping $T_{ij}$ to 1 if $\log T_{ij}$ is less than the median of the reference sample, 0 otherwise. Lastly, all of the limitations discussed in Chapter 5 still apply here since the core study design remained the same.
Table 6.1: Maximum likelihood estimates (MLE’s) of model parameters for GLRT.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Moving Sample (m)</th>
<th>Reference Sample (r)</th>
<th>Combined Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_j$</td>
<td>$\hat{p}^{(m)}<em>j = \frac{1}{m} \sum</em>{i=n-m+1}^n X_{ij}$</td>
<td>$\hat{p}^{(r)}<em>j = \frac{1}{n-m} \sum</em>{i=1}^{n-m} X_{ij}$</td>
<td>$\hat{p}<em>j = \frac{1}{n} \sum</em>{i=1}^n X_{ij}$</td>
</tr>
<tr>
<td>$\mu_{1j}$</td>
<td>$\hat{\mu}_{1j}^{(m)} = \frac{1}{</td>
<td>\mathcal{M}_1</td>
<td>} \sum_{i \in \mathcal{M}<em>1} \log T</em>{ij}$</td>
</tr>
<tr>
<td>$\sigma_{1j}^2$</td>
<td>$\hat{\sigma}_{1j}^{2(m)} = \frac{1}{</td>
<td>\mathcal{M}_1</td>
<td>} \sum_{i \in \mathcal{M}<em>1} \left( \log T</em>{ij} - \hat{\mu}_{1j}^{(m)} \right)^2$</td>
</tr>
<tr>
<td>$\mu_{0j}$</td>
<td>$\hat{\mu}_{0j}^{(m)} = \frac{1}{</td>
<td>\mathcal{M}_0</td>
<td>} \sum_{i \in \mathcal{M}<em>0} \log T</em>{ij}$</td>
</tr>
<tr>
<td>$\sigma_{0j}^2$</td>
<td>$\hat{\sigma}_{0j}^{2(m)} = \frac{1}{</td>
<td>\mathcal{M}_0</td>
<td>} \sum_{i \in \mathcal{M}<em>0} \left( \log T</em>{ij} - \hat{\mu}_{0j}^{(m)} \right)^2$</td>
</tr>
</tbody>
</table>

$\mathcal{M}_1 = \{ i : X_{ij} = 1, (n-m+1) \leq i \leq n \}$
$\mathcal{M}_0 = \{ i : X_{ij} = 0, (n-m+1) \leq i \leq n \}$
$\mathcal{R}_1 = \{ i : X_{ij} = 1, 1 \leq i \leq (n-m) \}$
$\mathcal{R}_0 = \{ i : X_{ij} = 0, 1 \leq i \leq (n-m) \}$
$\mathcal{C}_1 = \mathcal{M}_1 \cup \mathcal{R}_1 = \{ i : X_{ij} = 1, 1 \leq i \leq n \}$
$\mathcal{C}_0 = \mathcal{M}_0 \cup \mathcal{R}_0 = \{ i : X_{ij} = 0, 1 \leq i \leq n \}$
Figure 6.1: ROC curves for binomial, Fisher, and normal tests across six conditions ($m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 6.2: Lag plots for binomial, Fisher, and normal tests across six conditions ($m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 6.3: ROC curves for R, SM, Fisher, GLRT, KS, and Kuiper tests across six conditions ($m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 6.4: Lag plots for R, SM, Fisher, GLRT, KS, and Kuiper tests across six conditions ($m = \{5, 20\} \times \psi = \{0.15, 0.25, 0.35\}$). Results are averaged across 100 replications with about 500 items and 2,000 examinees.
Figure 6.5: ROC curves and lag plots for SM ($m = 5$) and GLRT ($m = 20$) at $\psi = 0.15, 0.25, 0.35$. Results are averaged across 100 replications with about 500 items and 2,000 examinees.
CHAPTER 7
CONCLUSION

Test security is ultimately a matter of test validity, which is unequivocally considered to be the most fundamental aspect of developing and evaluating tests (AERA, APA, & NCME, 2014). Understanding validity as the extent to which a test measures what it intends to measure, it is easy to appreciate the serious threat posed by security breaches to the integrity of testing programs (Cizek & Wollack, 2017a). Thus, in the interest of protecting test validity, this thesis focused on both preventive and diagnostic methodologies for improving test security from a psychometric perspective.

On the preventive end, simple yet effective techniques were introduced for controlling item exposure and test overlap in a variety of testing modes, including linear, on-the-fly MST, and CAT. Additionally, the asymptotic distribution of test overlap rate was derived for completely random item selection in CAT, which serves as the baseline comparison for assessing the security of adaptive algorithms. On the diagnostic end, powerful real-time procedures were introduced to detect item compromise in CAT. Specifically, it was demonstrated that response times can be successfully leveraged to find compromised items faster and more accurately than using responses alone.

Specific limitations of individual studies nonwithstanding, it is the author’s sincere hope that the presented body of research makes an incremental contribution to the relatively young and growing psychometric literature on test security. In ongoing efforts to better inform practice, much work still remains in advancing the field to maturity. In that respect, the prospect of continued engagement in such meaningful and consequential research is incredibly exciting and rewarding.
APPENDIX A

APPLICATION OF LYAPUNOV’S CLT

Assume that log RT is normally distributed as follows: \( \log T_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_j^2) \), where \( \mu_{ij} = \beta_j - \tau_i \) and \( \sigma_j^2 = 1/\alpha_j^2 \). The mean log RT of the moving sample for item \( j \) is then given as \( \hat{\mu}_j^{(m)} = \frac{1}{m} \sum_{i=n-m+1}^{n} \log T_{ij} \). Also, define the following:

\[
s_m^2 = \sum_{i=n-m+1}^{n} \sigma_j^2 = m\sigma_j^2.
\]

In this context, Lyapunov’s central limit theorem (CLT) states that

\[
\frac{1}{s_m^2} \sum_{i=n-m+1}^{n} (\log T_{ij} - \mu_{ij}) = \hat{\mu}_j^{(m)} - \frac{\sum_{i=n-m+1}^{n} \mu_{ij} / m}{\sigma_j / \sqrt{m}} \xrightarrow{d} \mathcal{N}(0, 1) \quad (A.1)
\]

if, for any \( \delta > 0 \), the following condition is met:

\[
\lim_{m \to \infty} \frac{1}{s_m^{2+\delta}} \sum_{i=n-m+1}^{n} E\left( |\log T_{ij} - \mu_{ij}|^{2+\delta} \right) = 0. \quad (A.2)
\]

Recognizing that the expectation term is a central absolute moment of log \( T_{ij} \),

\[
E\left( |\log T_{ij} - \mu_{ij}|^{2+\delta} \right) = \sigma_j^{2+\delta}(1+\delta)!! \cdot \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if } 2+\delta \text{ is odd} \\ 1 & \text{if } 2+\delta \text{ is even} \end{cases}. \quad (A.3)
\]

Therefore, using \( \delta = 2 \) for simplicity,

\[
\lim_{m \to \infty} \frac{1}{s_m^4} \sum_{i=n-m+1}^{n} E\left( |\log T_{ij} - \mu_{ij}|^4 \right)
\]

\[
= \lim_{m \to \infty} \frac{1}{m^2\sigma_j^4} \sum_{i=n-m+1}^{n} 3\sigma_j^4
\]

\[
= \lim_{m \to \infty} \frac{m(3\sigma_j^4)}{m^2\sigma_j^4}
\]
\[
= \lim_{m \to \infty} \frac{3}{m}
= 0,
\]

thereby meeting Lyapunov’s condition for the asymptotic normality of the test statistic.
REFERENCES


McLeod, L. D., & Schnipke, D. L. (1999, April). Detecting items that have been memorized in the computerized adaptive testing environment. In _Paper presented at the annual meeting of the national council on measurement in education_. Montreal, Canada.


