Summer 2019

Optimizing design of incorporating off-grade items for constrained computerized adaptive testing in K-12 assessment

Xiangdong Liu
University of Iowa

OPTIMIZING DESIGN OF INCORPORATING OFF-GRADE ITEMS FOR CONSTRAINED COMPUTERIZED ADAPTIVE TESTING IN K-12 ASSESSMENT

by

Xiangdong Liu

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations in the
Graduate College of
The University of Iowa

August 2019

Thesis Supervisors: Professor Catherine J. Welch
Professor Stephen B. Dunbar

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Xiangdong Liu

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Psychological and Quantitative Foundations at the August 2019 graduation

Thesis Committee: _____
Catherine J. Welch, Thesis Supervisor

_____
Stephen B. Dunbar, Thesis Supervisor

_____
Robert Ankenmann

_____
Brandon LeBeau

_____
Kung-Sik Chan

To my wife Shannon and my daughters Betsy and Olivia
for your endless love and support.

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my academic advisor and dissertation chair, Dr. Catherine Welch, for providing me guidance and support in my dissertation, and in my academic over the last six years.

I am grateful for my dissertation chair, Dr. Stephen Dunbar, for providing your wisdom, encouragement, and guidance for my dissertation and my research project during my assistantship with Iowa Testing Programs.

In addition, I would like to thank my other committee members – Dr. Robert Ankenmann, Dr. Brandon LeBeau, and Kung-Sik Chan, for their contributions to my dissertation. Their insightful review and suggestions help me greatly in my dissertation work.

I would also like to thank Dr. Timothy Ansley, Dr. Won-Chan Lee, Dr. Michael J. Kolen, Dr. Ariel M. Aloe, and Dr. Walter Vispoel, for providing wonderful courses in the measurement field.

Lastly, I appreciate my friends, Adam Reeger, Orry Susadya, Ye Lin, Keyu Chen, Lida Chen, JinMin Chung, Mingjia Ma, Ye Ma, and Yibo Wang, for their friendship and support during the past seven years.

ABSTRACT

Incorporating off-grade items within an on-grade item pool is often seen in K-12 testing programs. Incorporating off-grade items may provide improvements in measurement precision, test length, and content blueprint fulfillment, especially for high- and low-performing examinees, but it may also identify some concerns when using too many off-grade items on tests that are primarily designed to measure grade-level standards. This dissertation investigates how practical constraints such as the number of on-grade items, the proportion, and range of off-grade items, and the stopping rules affect item pool characteristics and item pool performance in adaptive testing.

This study includes simulation conditions with four study factors: (1) three on-grade pool sizes (150, 300, and 500 items), (2) three proportions of off-grade items in the item pool (small, moderate, and large), (3) two ranges of off-grade items (one grade level and two grade levels), and (4) two stopping rules (variable- and fixed-length stopping rule) with two SE threshold levels. All the results are averaged across 200 replications for each simulation condition.

The item pool characteristics are summarized using descriptive statistics and histograms of item difficulty (the *b*-parameters), descriptive statistics and plots of test information functions (TIFs), and the standard errors of the ability estimate (SEEs). The item pool performance is evaluated based on the descriptive statistics of measurement precision, test length and exposure properties, content blueprint fulfillment, and mean proportion of off-grade items for each test.

The results show that there are some situations in which incorporating off-grade items would be beneficial. For example, a testing organization with a small item pool attempting to improve item pool performance for high- and low-performing examinees. The results also show that practical constraints of incorporating off-grade items, organized here from most impact to

least impact in item pool characteristics and item pool performance, are: 1) incorporating off-grade items into small baseline pool or large baseline pool; 2) broadening the range of off-grade items from one grade level to two grade levels; 3) increasing the proportion of off-grade items in the item pool; and 4) applying variable- or fixed-length CAT. The results indicated that broadening the range of off-grade items yields improvements in measurement precision and content blueprint fulfillment when compared to increasing the proportion of off-grade items. This study could serve as guidance for test organizations when considering the benefits and limitations of incorporating off-grade items into on-grade item pools.

PUBLIC ABSTRACT

Incorporating off-grade items within an on-grade item pool for an adaptive test is an alternative way to improve measurement precision without developing new items. For an item pool confined for examinees from grade 4, items that are restricted to grade 4 content standards are considered on-grade items, and items that expand to below- or above-grade 4 content standards are considered off-grade items. One goal of this study was to investigate whether incorporating off-grade items into an on-grade items pool would be beneficial with respect to item pool characteristics and item pool performance. Another goal of this study was to investigate how different practical constraints such as the number of on-grade items, the proportion, and range of off-grade items, and the stopping rules affect item pool characteristics and item pool performance in adaptive testing.

The results show that there are some situations in which incorporating off-grade items would be beneficial. For example, a testing organization with a small item pool attempting to improve item pool performance for high- and low-performing examinees. The results also show that practical constraints of incorporating off-grade items, organized here from most impact to least impact in item pool characteristics and item pool performance, are: 1) incorporating off-grade items into small baseline pool or large baseline pool; 2) broadening the range of off-grade items from one grade level to two grade levels; 3) increasing the proportion of off-grade items in the item pool; and 4) applying variable- or fixed-length CAT. The results indicated that broadening the range of off-grade items yields improvements in measurement precision and content blueprint fulfillment when compared to increasing the proportion of off-grade items. This study could serve as guidance for test organizations when considering the benefits and limitations of incorporating off-grade items into on-grade item pools.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 INTRODUCTION

1.1 Background

Computerized adaptive testing (CAT) is a form of computer-based testing in which the next question or item selected is based on the examinee's response to the previous item or items. In other words, a correct response leads to a more difficult question and an incorrect response leads to an easier question. As a result, each examinee will receive a unique test based on a pattern of performance. In contrast, in fixed-length, linear testing, every examinee receives a test that is identical. Fixed-length, linear testing can be administered in the form of computer-based testing (CBT) or in the form of paper-and-pencil testing (PPT).

This study focuses on the item-level computerized adaptive testing, which adapts to the examinee's ability level with individual items. CAT is frequently used for accountability testing in K-12 education. Some organizations are currently using CAT, such as Smarter Balanced Assessment Consortium (SBAC) (American Institutes for Research (AIR), 2015), Measures of Academic Progress (MAP) (Northwest Evaluation Association, 2013), and Scantron Performance Series (Scantron, 2004). A CAT can only be administered online.

1.2 The Pros and Cons of CAT

CAT is a form of CBT that provides several advantages over conventional testing (CBT or PPT). First, previous studies concluded that for a unidimensional item pool, a CAT improves measurement precision and reduces test length more than conventional linear testing (Evans, 2010; Jodoin, 2003). This is expected because the CAT selects items that are near an examinee's ability level and maximizes information about an examinee's performance. Jodoin (2003) has argued that a conventional test needs to have double the test length in order to have equivalent reliability compared with CAT. Evans (2010) found that a CAT improves the precision of

measurement relative to conventional testing, especially for the high- and low-performing examinees. Second, CAT relies on Item Response Theory (IRT), which has the capability to locate the item statistics and examinee's score on the same score scale, given that the IRT assumptions are satisfied. As a result, examinees who respond to a very different sequence of items in CAT could have a comparable score, and there is no need for post-test equating to deal with tests of varying difficulty (Stone & Davey, 2011). Third, CAT intends to adjust the item difficulty level to match the examinee's ability level. This will prevent a low-performing examinee from being overwhelmed by overly difficult questions and a high-performing examinee from being bored by overly easy questions (Bong, 2016). Fourth, in CAT, each examinee will receive a unique test, and this practice will reduce the opportunities for cheating (Kantrowitz, Dawson, & Fetzer, 2011)

There are some limitations of CAT to be considered beyond merely the IRT assumptions of unidimensionality (there is one dominant latent trait being measured) and local independence (responses given to the separate items in a test are mutually independent given an individual score on the latent variable(s)). The first limitation of CAT is that item pool characteristics may prevent a CAT from being implemented effectively. Reckase (2003) defined the ideal item pool as one that contains every item requested by the item selection method. However, previous studies concluded that a common characteristic of item pools in the real world is that their item difficulty parameters are not uniformly distributed and there are generally more items for middle-performing examinees than for high- and low-performing examinees (Luecht, 2014; Way, 2006; Yi, Wang, & Ban, 2001; Wang & Vispoel, 1998). For example, Wang and Vispoel (1998) used an item pool with more items at middle difficulty levels than at extreme difficulty levels for their CAT study because it was more similar to many real-world item pools. Yi et al.

(2001) used a real item pool based on the ACT Mathematics Test (PPT) and found it contains more items at the middle difficulty levels than at the two ends of the ability distribution.

One reason this characteristic of item pools is common is that most testing organizations struggle to develop a completely new item pool for CAT, and therefore an item pool developed to support conventional testing may be used to support an initial transition to a CAT. Way (2006) found that the conventional item pool used to assemble a common PPT is typically a peaked item pool (i.e., has more informative items for middle-performing examinees than low- and high-performing examinees) because extremely difficult or easy items are often rejected for statistical reasons in conventional testing. Luecht (2014) concluded that an item pool developed to support the assembly of fixed-form PPT may not have adequate breadth and depth of item difficulty and item discrimination to support CAT. If the test is used for classifying examinees based on a single cut score, this is fine because a precise score is needed only at the cut score used for classifying examinees. However, it might be problematic for K-12 assessment, in that schools or districts aim to obtain high measurement precision for all examinees, including high- and low-performing examinees (Thompson & Weiss, 2011). Therefore, the peaked CAT item pool may potentially result in imprecise measurement for high- and low-performing examinees. There are two approaches to solve this limitation. The first approach is adding items to the item pool. The second approach is extending the on-grade item pool to include some off-grade items. The first approach requires more new items, which is usually very expensive and require extensive time to develop. Therefore, the second approach is the focus of the current study.

The second limitation of CAT is that it may raise issues with respect to content validity (Crotts, Sireci & Zenisky, 2012; Wei & Lin, 2015; the National Center for Research on Evaluation, Standards and Examinee Testing [CRESST], 2016). For conventional testing,

content validity could be confirmed by subject experts before it is implemented. However, for CAT, each examinee receives a unique test based on his or her ability level. It is important to evaluate whether the test administered to different examinees satisfies the content specification or content blueprint, which describes the number of items in each content domain. Luecht, De Champlain, and Nungester (1998) found that ignoring the constraint of content balancing might cause a violation of content alignment if the content blueprint requires items from multiple content domains. For example, if one content domain contains more difficult items while another content domain contains easier items, a low-performing examinee may only see items in one content domain, and a high-performing examinee might only see items from another content domain. Crotts et al., (2012) concluded that even with the constraint of content balancing, whether the content blueprint is satisfied should be investigated. In addition, when an item pool developed to support conventional testing may be used to support a transition to a CAT, the content blueprint fulfillment for high- and low-performing examinees need to be examined.

1.3 CAT in K-12 Assessment

Previous studies in CAT focused primarily on classification tests, which typically classify all examinees into one of two or more categories (e.g., pass/fail) (Thompson, 2009; Xing & Hambleton, 2004; Eggen & Straetmans, 2000). The main goal of these tests is to focus on the better measurement around the cut score, which helps ensure accuracy in decision making. However, recently there has been an increased movement toward CAT in K-12 assessment. The purpose of K-12 assessment is to obtain high measurement precision for all examinees, including high- and low-performing examinees (Thompson & Weiss, 2011). Therefore, the main goal of CAT in K-12 assessment will improve measurement precision across a wide range of ability

levels and ensures content validity for each test. A CAT in K-12 assessment has some special features that will be discussed below.

First, the K-12 assessment usually requires a constrained CAT. A constrained CAT typically refers to a CAT program with constraints of content balancing and item exposure control (He, 2010). Content balancing is used because it is common that there are multiple content areas in K-12 assessment, and this procedure is to help ensure each examinee receives a test that has a similar content composition in the CAT. Item exposure control is used because test security is an important consideration in K-12 summative assessment, and this procedure prevents items from being overused and increases the utilization of items rarely used in the item pool. Since the constraints of content balancing and item exposure control may force the examinees to take items with inappropriate difficulty, the measurement precision and test efficiency of CAT may be affected (Reese, Schnipke & Luebke, 1999).

Second, K-12 assessment has the challenge of measuring accurately for high- and low-performing examinees. In the K-12 setting, some schools or districts aim to identify high- and low-performing examinees in order to provide advanced or remedial courses. This goal could be achieved when the item pool for CAT contains sufficient high-quality items across a wide enough range of ability levels, which would result in high measurement precision for high- and low-performing examinees. SBAC developed new item pools based on Reckase's (2010) "bin" method. Its item pool development aimed to include a wide enough range of item difficulties across ability levels. However, even SBAC needed to expand their on-grade item pool to include some off-grade items in order to measure high- and low-performing examinees sufficiently (SBAC, 2016). For test organizations or school districts that want to implement CAT, if the item pool does not contain sufficient high-quality items for a wide enough range of ability levels, the

problem of imprecise measurement for high- and low-performing examinees maybe more serious.

Third, K-12 assessment has a need to meet the content blueprint. This is true because K-12 achievement tests primarily focus on evaluating examinee performance on groups of items classified into content knowledge or process skills. The content blueprint fulfillment provides evidence for the content validity of the CAT. However, in the K-12 assessment, some districts or schools may still be using conventional linear testing. In these situations, it is important that the results of the CAT are directly comparable with conventional linear testing (CBT or PPT). A content blueprint provides constraints to help ensure content validity. Comparability of different forms of tests depends on each test satisfying the content blueprint. Content blueprints for a CAT usually include the intended lower- and upper- bounds (the minimum and the maximum number of items) to be administered within each content domain. For an item pool that does not contain sufficient items across a wide range of ability levels for each content domain, the content blueprint fulfillment will likely be a problem in a constrained CAT, especially for high- and low-performing examinees.

1.4 Main CAT Variables Investigated in the Study

An off-grade item refers to an assessment item that was developed for examinees at a higher or lower grade level (Wei & Lin, 2015). Incorporating off-grade items into an on-grade item pool refers to developing an item pool to support a grade level that includes some items developed to support grade levels above and below the target grade. Several studies (Wei & Lin, 2015; AIR, 2015; Way et al., 2010;) have concluded that incorporating off-grade items improved measurement precision, test efficiency, and content blueprint fulfillment for high- and low-

performing examinees. Thus, incorporating off-grade items is an ideal way to address unique challenges in K-12 assessment and are the main interest of the current study.

1.5 Purpose of the Study and Research Questions

The main goal of this study is to investigate how practical constraints such as the number of on-grade items, the proportion and the range of off-grade items, and the stopping rules affect item pool characteristics and item pool performance in adaptive testing.

The following research questions were addressed:

1. How do different numbers of on-grade items in the item pools affect the impact of incorporating off-grade items in terms of item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?
2. How does incorporating different proportions of off-grade items affect the item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?
3. How does incorporating different ranges of off-grade items affect the item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?
4. How does incorporating off-grade items affect item pool performance differently under variable- and fixed-length CAT with two different standard error (SE) threshold levels?

1.6 The Significance of the Study

Wei and Lin (2015) is the only current study that investigated incorporating off-grade items under variable- and fixed-length CAT. The current study extended their work and focusses on other perspectives of incorporating off-grade items by investigating several practical

constraints that affect the impact of incorporating off-grade items. Second, previous studies did not provide a complete picture of the effects of incorporating off-grade items. For example, Wei and Lin (2015) only discussed the benefits of incorporating off-grade items, such as measurement precision, test efficiency, and content blueprint fulfillment. The current study, however, presented the issue of incorporating off-grade items, such as the mean proportion of off-grade items for each test, which is a challenge for tests that are designed to measure grade-level standards. Third, by investigating which practical constraints of incorporating off-grade items has the most impact in item pool characteristics and performance, this study could be a guide for test organizations when considering incorporating off-grade items within an on-grade item pool.

Chapter 2 LITERATURE REVIEW

2.1 Overview

In this chapter, literature related to the design of incorporating off-grade items and termination rules on properties of simulated CAT (such as measurement precision, content blueprint fulfillment, and test efficiency) is discussed. For CAT to work successfully, test developers must consider five basic components: initial ability estimate, item selection method and constraints, ability estimation, termination rule, and item pool (Thompson & Weiss, 2011). However, these components have been well studied and will be fixed in this study. Some basic information and the support for certain choices on those variables are discussed.

2.2 Incorporating Off-Grade Items to the On-Grade Item Pool

Off-grade testing was defined as the representation of an examinee in one grade who is assessed with a test level developed for examinees in another grade (Thurlow, Elliott & Ysseldyke, 1999). Thurlow et al., (1999) indicated that concern for off-grade testing is that it is inappropriate for systematic accountability because off-grade testing does not allow the examinee to demonstrate mastery at the required difficulty level. Minnema, Thurlow, Bielinski, and Scott (2000) indicated two advantages of using off-grade tests. First, off-grade testing may increase measurement precision for high- and low-performing examinees by better matching their ability level to the item difficulty level. Second, off-grade testing may reduce examinees' frustration and chance responding when they are administered an overly difficult or overly easy test. The authors also indicated a concern for off-grade testing is that it requires a common scale across disparate grade levels. However, both concerns from these two studies could be resolved in the CAT, which obtains a common scale through the IRT method.

Way et al., (2010) proposed using off-grade content within CAT. The authors stated that the CAT is a natural vehicle to incorporate off-grade items into the on-grade item pool. Specifically, Way et al., (2010) noted the following:

> in adaptive testing where examinee results and vertical articulated standards are expressed on the same common scale, each examinee can be compared to the same 'on-grade' standard regardless of whether or not off-grade items were used in their particular adaptive test. (p. 5)

The authors stated that using off-grade items in CAT required placing the b-parameters for all items across grade on a common scale through a vertical scaling study. In addition, the authors suggested that measurement of off-grade content in grades 3 - 8 of CAT should be preferred because the content linkage across grade level is less clear at the high school level. Vertical scaling refers to the process of linking test scores that assess similar content areas but at different grade levels onto a common scale, which allows for direct comparison of student test scores across grade levels within a content area (Texas Education Agency, 2013). Thus, this study will use the generated item pool with vertically scaled items across grades. Off-grade items will be realigned to the on-grade blueprints.

2.3 The Number of On-Grade Items in the Item Pool

Item pool size has a great impact in measurement precision, test efficiency, and content blueprint fulfillment (Wei & Lin, 2015; Hembry, 2014; Babcock & Weiss, 2013; Lim, 2010; Xing & Hambleton, 2004; Davey, 2011; Dodd, Koch, & De Ayala, 1993). To illustrate the impact of item pool size, Xing and Hambleton (2004) conducted a simulation study that compared the impact of two item pool size on decision consistency and accuracy. The two item pool sizes are 240 and 480 multiple-choice items. The authors concluded that item pool size had

a practically significant impact on decision consistency and accuracy. A 480-item pool yields improvements in decision consistency and accuracy, test information function, and item exposure rate are more than a 240-item pool. Babcock and Weiss (2013) conducted a simulation study that compared the impact of four stopping rules under two size of item pool (100 and 500 items). The authors concluded that the 500-item pool had a relatively high level of test information over the entire range of θ. The authors also concluded that when using a large item pool, applying a strict (low) standard error threshold for fixed standard error termination rule yielded good measurement precision. Babcock and Weiss (2013) indicated that some testing organizations usually have a few high-volume entry exams with a large item pool along with a variety of advanced exams with low examinee volume and smaller item pool.

Item pool size needs to be sufficient to help ensure desired results in measurement precision, test efficiency, and content blueprint fulfillment. Davey (2011) recommended that the item pool should contain about 10 times as many items as fixed-length CAT. Davey (2011) also indicated that for a high-stakes achievement test, if content balancing method is used, a larger item pool (more than 500 items) might be desirable. However, increasing item pool size to a more than sufficient amount is not necessary. For example, in Wei and Lin's (2015) study, enlarging an item pool by three times only increased content blueprint fulfillment by 2% (from 92% to 94% when SE threshold is 0.25), and produced little improvement in measurement precision and test efficiency.

The current study will examine three on-grade pool sizes (150, 300, and 500 items). A 500-item pool targets high-volume entry exam or summative assessment, while 150- and 300-item pools reflect remedial and advance exams with low-examinee volume or formative assessment. Summative assessment assesses students' mastery of learning goals at the end of the

school year and is usually used for accountability measure. It usually requires a high measurement precision and prefers small or less proportion of off-grade items on each test. Formative assessment is to help teachers conducting in-process evaluations of student learning needs, which are usually developed at the school or district level. It usually does not require a high measurement precision and could tolerance moderate to a large proportion of off-grade items on each test.

Additionally, a 500-item pool is based on Davey's (2011) suggestion that 500 items are more appropriate for summative assessment. A 300-item pool is based on Stocking's (1994) recommendation that the item pool should contain about 6 times as many items as PPT (the PPT test for grade 4 is usually around 50 items). Applying a 150-item pool also considers that the gains in item pool characteristics and item pool performance are more obvious when incorporating off-grade items into a small item pool. The preliminary simulations indicate that the different impacts of these three item pool sizes are evident.

2.4 Appropriate Proportion of Incorporated Off-Grade Items

Previous studies concluded that a common characteristic of item pools in the real world is that their item difficulty parameters are not uniformly distributed and there are generally more items for middle-performance examinees than for high- and low-performing examinees (Luecht, 2014; Way, 2006; Yi et al., 2001; Wang & Vispoel, 1998;). However, He and Reckase (2014) conducted a simulation study that compared the optimal item pool for CAT content-balancing and exposure control procedures. The authors found that for the most desirable item pool designs based on the combination of pool size, test security, and measurement accuracy, their *b*-parameter distributions are closer to a uniform distribution than a normal distribution. To illustrate the impact of different information distribution on item pool, Babcock and Weiss

(2013) conducted a simulation study that compared item pool with uniform *b*-parameter distribution and item pool with a peaked *b*-parameters distribution. The author concluded that for a 500-item pool, uniform *b*-parameter item pool performs better in terms of measurement precision and test length than peaked *b*-parameters item pool under variable-length CAT.

Incorporating off-grade items in on-grade item pool will yield a mixed distribution. With vertical scaling, off-grade items will be realigned to the on-grade blueprints, and the mix of distribution will give the item pool a shape that is closer to a uniform distribution, where more items are available in the two ends of the ability distribution.

Several studies have mentioned the proportion of off-grade items incorporated in the item pool (AIR, 2016; AIR, 2015; Wei & Lin, 2015). However, these studies are not conclusive with respect to the impact of the proportion of off-grade items in the item pool. For example, in Wei and Lin's (2015) study, the ratio of off-grade items to on-grade items is 2 because they combine the item pool for grade 4 with the item pools for grade 3 and 5. This is impractical when designing tests that are tied to grade level standards because the on-grade items should serve as the larger part of the pool. In the Smarter Balanced 2014–15 simulation report, the off-grade items in the item pool have a range of approximately 3% to 7% in grade 3-8 mathematics (AIR, 2015). In the Smarter Balanced 2016–17 simulation report, the off-grade items in the item pool have a range of approximately 2% to 6% in grade 3-8 mathematics (AIR, 2016). But these studies provide no justification as to why the proportion of off-grade items in the item pool is chosen in their study. Therefore, there is limited guidance in the literature of this proportion for test organization on the decision to incorporate off-grade items. The current study will employ three proportions of off-grade items in the item pool: small, moderate, and large (the ratios of the

off-grade item to on-grade items in the item pool are 1/3, 2/3 and 3/3 accordingly). The choice of these three proportions was made based on the results of the preliminary simulation.

2.5 Appropriate Range of Incorporated Off-Grade Items

Several studies have mentioned the range of off-grade items (AIR, 2016; AIR, 2015; Wei & Lin, 2015; Scantron, 2004). However, these studies are not conclusive with respect to the impact of the range of off-grade items. For example, Wei and Lin (2015) incorporated off-grade items from one grade below and one grade above the target grade. Scantron (2004) incorporated off-grade items no more than three grade levels above or below. In a 2014–15 simulation study report, SBAC selected off-grade items, one grade above and one grade below in ELA/L and two grades below in mathematics (AIR, 2015). In a 2016–17 simulation study report, SBAC selected off-grade items—one or two grades above and one to three grades below the on-grade (AIR, 2016). However, these studies provide no justification regarding the ranges of off-grade items were chosen in their study.

Broadening the range of off-grade items from one grade level to two grade levels may provide a wider range of item difficulty for low-and high-performing examinees, but it may also lead to the issue that the middle-performing examinees will be administered items from two grade levels above or below, which may pose a more serious challenge for tests that are designed to measure grade-level standards. Thus, the impact of the range of off-grade items needs to be examined. The current study will provide two conditions for this range of incorporating off-grade items: one grade above and one grade below the target grade, two grade levels above and two grade levels below the target grade. The choices of these two ranges are based on the previous studies' finding.

2.6 Termination Rules

Termination rule is among the most crucial components of a CAT and there are significant differences in results among different termination rules (Wei & Lin, 2015; Babcock & Weiss, 2013;). Typically, there are two main categories for terminating a CAT: variable- and fixed-length termination rules. Variable-length CAT (fixed-standard error (SE) termination) refers to a CAT that ends the test when a predefined estimated standard error level is reached (Wang & Vispoel, 1998). Fixed-length CAT refers to a CAT that ends the test when a predefined number of items is administered to each examinee (Wang & Vispoel, 1998).

Previous studies concluded that fixed-standard error termination rule performed better than its fixed-length counterparts in measurement precision, while fixed-length CAT yielded better content blueprint fulfillment than variable-length CAT (Cohen & Albright, 2014; Babcock & Weiss, 2013). For example, Babcock and Weiss (2013) concluded that fixed-standard error termination rule performed better than its fixed-length counterparts in terms of measurement precision. Cohen and Albright (2014) concluded that a fixed-standard error termination rule may result in content blueprint violation because of the inconsistencies between the blueprint content specifications and the information criteria. Wei and Lin (2015) concluded that fixed-length CAT results in better content blueprint fulfillment than variable-length termination. Wei and Lin found that fixed-length CAT produces results in 100% content blueprint fulfillment while variable-length (fixed-SE) CAT produces results in content blueprint fulfillment from 93% to 100% depending on the threshold of standard error.

Several studies conducted CAT studies with both fixed- and variable-length CATs (Wei & Lin, 2015; Shin, Chien & Way, 2012; Wang & Vispoel, 1998). For example, Wang and Vispoel (1998) compared four ability estimation methods with fixed- and variable-length (fixed-SE) CAT. Several studies have recommended that 0.32 should be used for a fixed-SE

termination rule for accurate θ (person ability) estimation (Babcock & Weiss, 2013; Wang & Vispoel, 1998). For example, Babcock and Weiss (2013) concluded that a fixed-SE termination rule should use a standard error equal to or smaller than 0.315 (reliability of 0.90) for accurate θ estimation. The authors also conclude that using a higher SE threshold such 0.385 will yield a test with too few items and has a negative effect in measurement precision (high Bias and RMSE). Wei and Lin (2015) examined the impact of different SE threshold values of fixed-SE termination rule and found that different SE thresholds yield different content blueprint fulfillment. For on-grade item pool, the SE level of 0.2, 0.25, 0.3, and 0.35 results in 73%, 87%, 92%, and 98% respectively of examinees' tests meeting the requirement of the content blueprint. The average ability estimates for grade 4 on mathematics are 0.26 at the overall level for SBAC (AIR, 2016). To examine whether SE threshold will have an impact on incorporating off-grade items, this study will employ two SE threshold levels, 0.25 and 0.32.

To prevent some extreme high- or low- performing examinees from receiving tests that are either too long or short, a fixed minimum and/or a maximum number of items could also be set even under a variable-length termination rule. Yi et al. (2001) used a 30-item fixed-length CAT termination rule and set a 60-item maximum for fixed-SE termination rule. Babcock and Weiss (2013) established a 100-item maximum for fixed-SE termination rule and a 10-15 item minimum for dichotomously scored items. Wang and Vispoel (1998) established a 50-item maximum for fixed-SE termination rule.

Thus, the impact of incorporating off-grade items will be examined in both variable- and fixed-length constrained CATs in the current study. For the variable-length CATs, two SE threshold levels (0.32 and 0.25) will be used for fixed-SE termination rule and a 15-item

minimum and 80-item maximum will be set. For the fixed-length CAT, 40 items will be set for fixed-length termination rule.

2.7 Evaluation Criteria

The impact of incorporating off-grade items will be examined based on item pool characteristics and item pool performance. Item pool characteristic will be introduced first since it will clarify why item pool performances are different across item pool compositions.

2.7.1 The Item Pool Characteristics

Item pool characteristics were widely applied and discussed in several studies (Yang, 2016; Mao, 2014; Lee, & Dodd, 2012). Item pool characteristics in combination with item pool performance will provide a full and accurate picture of the impact of incorporating off-grade items. The item pool characteristics in this study will involve descriptive statistics and histograms of the *b*-parameters, descriptive statistics, and plots of test information functions (TIFs) and the standard errors of the estimate (SEEs).

Item information function is an important feature of the IRT model, which is a mathematical function of the ability level theta and the item parameters that indicate how informative the item is at any given theta level (Magis, 2013). For the IRT 1PL model, the following is the item information function.

$$I_i(\theta) = \frac{e^{(\theta - b_i)}}{(1+e^{(\theta - b_i)})^2}$$

Where theta is the estimated abilities and b_i are the difficulty parameter value for item i. Test information function is the sum of the administered item information functions. SEE is defined as the inverse of the TIF. The lower the SEE, the more precision of ability estimation (Baker, 2001).

Luecht (2015) suggested that item pool development should consider targeted test information (specific prescribed amount of measurement precision along the scale). Item pool characteristics such as the figures of the TIFs and SEEs provide powerful tools to view the precision of item pools that incorporate off-grade items. For example, an increase in the ranges of *b*-parameters indicates more items at the two ends of the θ scale (for low-and high-performing simulees). The increase of the average value of the TIFs indicates more test information, which the test could measure ability more precisely. The decrease in the average value of the SEEs indicates more measurement precision.

2.7.2 Measurement Precision

One goal of a test is to achieve the desired level of precision. Bias (a measure of the systematic deviation of the estimated ability from the true ability); RMSE (a measure of absolute accuracy of parameter recovery, taking into account Bias and SE); correlation between true and estimated ability, and SE (a measure of standard error of ability estimate) are commonly used statistics for evaluating ability estimation precision (Wang & Zhang, 2017; Yang, 2016; Risk, 2015; Piromsombat, 2014; Babcock & Weiss, 2013; Wang & Vispoel, 1998). For example, Wang and Vispoel (1998) compared different ability estimation methods and used the indices of Bias, RMSE, correlation, and SE to evaluate the precision of ability estimation. Wang and Zhang (2017) examined the impact of information type, information zone size and item bank size on item selection methods and used the indices of Bias, RMSE, and SE to evaluate the precision of ability estimation. As a result, these four indices will be used to provide descriptive information in measurement precision for the current study. This study will examine measurement precision at the overall level, and for high- and low-performing simulees.

(1) Bias: a measure of the systematic deviation of the estimated ability from the true ability (Risk, 2015). It indicates whether the estimate parameters are over- or under-estimating true parameters. A positive value of Bias means an under-estimate, whereas a negative value of Bias means an over-estimate. The lower Bias value indicates a more accurate estimate close to zero. Lei and Zhao (2012) defined Bias as the difference between the average of the parameter estimates over 100 replications and the corresponding true parameter. Wang and Vispoel (1998) defined Bias as a systematic error.

(2) Root mean squared error (RMSE): a measure of the absolute accuracy of parameter recovery, taking into account the Bias and SE (Risk, 2015). It is a function of both bias and the variability of the sample parameter (standard error). The lower the RMSE value, the closer the CAT estimated ability is to the examinee's true ability. Lei and Zhao (2012) defined the RMSE as the square root of the sum of the squared Bias and variance of the estimates over 100 replications. Wang and Vispoel (1998) defined RMSE as the measure of the total error of estimation that has two components: the squared Bias and the squared SE of the estimates.

(3) Correlation between true and estimated ability: Pearson product-moment correlations between true and estimated thetas describe the accuracy of the estimation, the higher the correlations the more accurate the estimation.

(4) Standard error (SE): indicates the standard error of ability estimate. Wang and Vispoel (1998) defined SE as a random or sampling error.

Equations used to calculate Bias, RMSE and SE are shown by the following equations (Wang & Zhang, 2017):

$$Bias = \frac{1}{N}\sum_{j=1}^{N}(\widehat{\theta_J} - \theta_J) \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\widehat{\theta_J} - \theta_J)^2} \qquad (2)$$

$$SE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\,\widehat{\theta} - \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta_J}\,)^2} \quad (3)$$

Where $\widehat{\theta_J}$and θ_J are the estimated and true abilities of the J examinee. N is the number of examinees.

2.7.3 Test Efficiency

Test efficiency usually involves test length and exposure properties. Average or mean test length is commonly used for the index of the test efficiency of the CAT (Wei & Lin, 2015; Babcock & Weiss, 2013). One advantage of a variable-length CAT over fixed-length CAT and PPT is that it reduces the average test length. Thompson (2009) indicated that the constraints of item exposure and content balancing will increase average test length (ATL). As expected, Thompson (2009) stated that setting a maximum test length will decrease ATL, whereas setting a minimum test length will increase ATL. The ideal test length of CAT is half that of traditional PPT. The shorter the test length, the greater the test efficiency.

This study will also use exposure control properties, such as the item overlap rate and maximum exposure rate. The item overlap rate refers to the average number of cases in which the same items were administered to two randomly selected examinees (Moyer, Galindo, & Dodd, 2012). A maximum exposure rate of 0.3 means an item would be seen by no more than 30% of all examinees. Based on previous studies (He & Reckase, 2014; Moyer et al., 2012; Grossman, 2010; Chang, & Ansley, 2003), the ideal item overlap rate and maximum exposure rate would be less than 0.3. Thus, the closer the item overlap rate and maximum exposure rate to 0.3, the more test security the test will have.

2.7.4 Content Blueprint Fulfillment

The second goal of a test is to achieve alignment with the content blueprint. Test content blueprint includes the expected number of items and the lower and upper bounds of the numbers of items administered to examinees in each content domain. The content blueprint fulfillment refers to the proportions of examinees who took tests with a desirable number of items within the lower and upper bounds for each content domain in the blueprints.

Several studies have evaluated content blueprint fulfillment (Wang & Zhang, 2017; CRESST, 2016; AIR, 2015; He, Diao & Hauser, 2014). CRESST (2016) provided two indices of content blueprint fulfillment for each subject and grade in SBAC: (1) the blueprint violation that only failed to meet the lower-bound of blueprints, and (2) the blueprint violation that exceeded the upper-bound of the blueprint. CRESST (2016) concluded that an administered test might not meet the content blueprint for two reasons: First, in some cases, item difficulty distribution and the content coverage maybe different. Second, there maybe no remaining eligible items in the item pool that satisfy the content blueprint. He et al., (2014) indicated three indices used to evaluate content blueprint fulfillment: (1) overall content blueprint violation (the percentage of tests that violated both the intended lower- and/or upper-bounds of blueprints) (2) the average number of content violation across all examinees, and (c) lower-bound only content blueprint violation (the percentage of tests that violated the intended lower-bound of blueprint). He et al., (2014) indicated that lower-bound only content blueprint violation is a less restrictive criterion than upper-bound only content blueprint violation because a test usually can be considered valid if the requirement of the minimum number of item is met. Wei and Lin's (2015) study focused on content blueprint fulfillment for overall examinees, and for high- and low-performing examinees.

Based on the literature, the current study will provide two indices: (1) lower-bound content blueprint fulfillment (the percentage of tests that meet the intended lower bound of blueprint for content coverage), and (2) upper-bound content blueprint fulfillment (the percentage of tests that meet the intended upper-bound of blueprint for content coverage). For example, if all the tests that examinees responded to meet the predefined lower and upper bounds of the blueprint, then the content blueprint fulfillment would be 100%. The current study will examine these three indices at the overall level, and for high- and low-performing simulees.

Content-balancing control needs to be taken into consideration for content blueprint fulfillment. Content balancing is used when a test is comprised of multiple content domains and CAT algorithms need to help ensure each test meets the desired content distribution. For example, Kingsbury and Zara's (1989) constrained CAT (CCAT) procedure restricts the number of items eligible for selection within each content domain each time an item is selected (Leung, Chang, & Hau, 2003). Shin et al., (2012) compared three content balancing methods and found that for tests with mutually exclusive content constraints, CCAT methods had 100% on target rate (meaning the restrictions are 100% successfully implemented).

2.7.5 Mean proportion of off-grade Items for Each Test

SBAC limited off-grade items administered only to low- and high-performing examinees through their unique CAT algorithm. Through their CAT algorithm, after examinees respond to two-thirds of the operational items, if an examinee is classified into the novice level or the advanced level, then off-grade items will be administered to the examinee. SBAC evaluated this criterion by comparing: (1) the number of examinees who responded to off-grade items, (2) the number of low-performing examinees who took below-grade items, and (3) the number of high-performing examinees who took above-grade items (AIR, 2015). However, their unique CAT

algorithm may introduce complexity and produce more classification error. According to Wainer et al.,'s (1990) conclusion, CAT is fair as long as the off-grade items are realigned to the on-grade blueprints, and as long as the *b*-parameters for all items across grades are on a common scale through vertical scaling. Then the limitation of the off-grade item in SBAC is not necessary. The current study incorporates off-grade items to the on-grade item pool without a limitation of when to incorporate. This study uses three indices to evaluate mean proportion of off-grade items for each test: (1) mean proportion of off-grade items (includes grades 2, 3, 5 and 6) for each test, (2) mean proportion of off-grade items from two grade levels only (grades 2 and 6) for each test, and (3) the number of tests that include off-grade items. The reason to use an index (2) is that compared to off-grade items from one grade level (grades 3 and 5), it would be a more serious challenge for tests that are designed to measure grade-level standards.

Two studies defined high- and low-performing examinees as the top and bottom 10% of the ability distribution (Wei & Lin, 2015; Evans, 2010). Thus, to make the simulation more realistic, the current study will define low-performing examinees as the bottom 10% of the overall simulees according to their true abilities within a particular grade, and high-performing examinees as the top 10%.

2.8 Basic CAT Components

There are several basic CAT components that an optimal CAT design needs to take into consideration: (1) the IRT model, (2) item selection method, (3) the constraints of content balancing and item exposure control, (4) ability estimation method, and (5) initial ability estimate. However, these variables have been well studied and will be fixed in this study. Some basic information and the support for selecting certain values among those variables are provided below.

2.8.1 IRT Model

The IRT model is the building block for other CAT variables. Three IRT models are commonly used for dichotomous score items: one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL). With a 2PL or 3PL model, items with a high discrimination parameter will have more chances to be selected; when item exposure control is implemented, those highly discriminating items will quickly run out, which may result in unexpected item pool performance (Gonulates, 2015). To prevent overuse of highly discriminating items and equalize item exposure, SBAC employed the 2PL IRT model but fixed the item discrimination parameters with a constant for each grade (Cohen & Albright, 2014). However, one may argue that SBAC does not apply a real 2PL IRT model. In addition, Wang and Zhang (2017) examined the effects of ignoring the discrimination parameter in CAT item selection in SBAC and found that it will lead to differences in the accuracy of ability estimation and item exposure. Therefore, the current study will use the unidimensional 1PL model for multiple-choice items, as many previous studies did with their CAT simulation and item response generation (Gonulates, 2015; Risk, 2015; Wei & Lin, 2015; He & Reckase, 2014). The 1PL model assumes that all items have identical discrimination and guessing has a negligible effect. The items in the item pool will differ only by their difficulty parameter. Below is the 1PL model.

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1+\exp(\theta - b_i)} \qquad (4)$$

where P_i indicates the probability of a correct response for an examinee with a latent ability θ. b_i indicates item difficulty parameter.

2.8.2 Item Selection Method

There are two main item selection methods: the maximum information (MI) method and the Bayesian approach. The MI method selects the items with maximum Fisher information at the current ability estimate or the cut score. The Bayesian method is based on the entire posterior ability distribution and selects the item with the most information on average across the high-density region of the posterior distribution (Keng, 2008). Chen, Ankenmann, and Chang (2000) compared the MI and Bayesian methods and found no significant differences between these two methods with respect to Bias, RMSE, SE, and item overlap when test length is over 10 items. Therefore, as MI at the cut score is more conceptually appropriate for classification purposes, MI at the current ability estimate method will be chosen for the current study.

2.8.3 Constraints of Content Balancing and Item Exposure Control

The constraints of content balancing and item exposure control are important factors to help ensure content validity and test security objectives. For content balancing procedures, the two most popular are Kingsbury and Zara's (1989) constrained CAT (CCAT) procedure and Stocking and Swanson's (1993) weighted deviations model (WDM). The CCAT method selects the most optimal item from the content domain that is farthest below its predefined administration percentage for each examinee (Leung et al., 2003). The WDM method selects the most optimal item that can minimize the weighted sum of deviations from the target value (it requires assigned weights for each constraint) (He et al., 2014). Shin et al., (2012) found that for tests with mutually exclusive content constraints, The CCAT method (Kingsbury & Zara, 1989) performs better than the WDM method and produces a result with 100% on-target rate. Unlike the WDM method, which requires weights to be assigned and more time to adjust, the CCAT

method is easy to understand and simple to implement, and therefore will be applied in the current study.

Item exposure control procedures prevent items from being overused and increase the utilization of items rarely used in the item pool. The exposure rate of an item refers to the proportion of tests in which the item was administered. There are two commonly used item exposure control procedures: randomization selection and conditional selection. Randomesque exposure control approach first chose a group of items which is close to a person's performance level and randomly selects one from the group of items (Leroux, Lopez, Hembry, & Dodd, 2013). Conditional selection procedure constrains the target maximum exposure rate to a pre-specified level, such as Sympson-Hetter method (Leroux et al., 2013). Leroux et al., (2013) reviewed several item exposure control methods and found randomesque exposure control procedures perform better in measurement precision than other exposure control methods when minimal item exposure control is needed. This study will apply a randomesque exposure control approach.

2.8.4 Ability Estimation

There are four options for ability estimation methods: maximum likelihood estimation (MLE), expected a posteriori (EAP), maximum a posteriori (MAP), and weighted likelihood estimation (WLE). MLE and EAP are the most commonly used among these four options. EAP and MAP are Bayesian methods. Previous studies indicated that there were no significant differences in measurement precision between MLE and Bayesian method for test length over 30 items (Kolen & Tong, 2010; van der Linden & Pashley, 2009; Wang & Vispoel, 1998). Wang and Vispoel (1998) concluded that MLE performs better than Bayesian methods for unbiased estimates and less bias, especially at the ability extremes. However, Wang and Vispoel (1998)

also indicated that MLE will result in an infinite ability estimate when the response is all correct or incorrect. There are three options that could resolve this problem. One way is to use the step size method to increase or decrease the item difficulty level until the mix response pattern has happened (Weiss & Guyer, 2012). A second way is to use the EAP method until a mixed response pattern happens (Weiss & Guyer, 2012). A final option is used in SBAC in which under MLE condition, in which all correct or all incorrect responses will be assigned the highest or lowest obtainable scale score (CRESST, 2016). This study will use the EAP method at the beginning of the test, then MLE at the final estimation method.

2.8.5 Initial Ability Estimate

There are usually three options for determining the initial ability estimate: the fixed initial ability estimate, the variable initial ability estimate, and the prior initial ability estimate (Weiss & Guyer, 2012). In the first option, all examinees typically begin the CAT with the same item difficulty (medium) as the initial ability estimate. In the second option, an initial ability estimate is randomly assigned each examinee within the specified interval. This will reduce the item exposure rate for the first items. The third option allows prior information of examinee's ability to be used. This happens especially when several tests are taken at the same time. This will increase the efficiency of the CAT. Bergstrom, Lunz, and Gershon (1992) concluded that under a long test length, different initial entry levels do not significantly impact the test results. Wang and Vispoel (1998) indicated that using a variable initial ability estimate will reduce the bias of Bayesian estimation methods but will have only negligible effects on the MLE. The current study will use 0 as the initial ability estimate, then use randomesque exposure control to select one item from a group of five items that are close to the ability estimate of 0.

2.9 CAT Simulation Tools

There are two commonly used computer programs specifically designed for simulation studies on CAT: CATSim (Weiss & Guyer, 2012) and R package catR (Magis, Raiche, & Barrada, 2017). Piromsombat (2014) compared the CATSim with the R package catR and found the overall performance is similar. This study will use catR because catR has recently been updated by the authors, it provides a function to draw TIFs and SEEs, and it is easy to do replications. However, CATSim does not allow the user to modify its syntax. There is no option for multiple replications. In addition, catR provides desired results of dependent variables; for example, Bias, RMSE, correlations between true ability and estimated ability, mean test length, item overlap rate and maximum exposure rate at the overall level, and Bias, RMSE, mean standard error, mean test length at conditional on deciles of the true ability level. However, for CATSim, these statistics need to be calculated manually. For these reasons, the R package catR will be used in this study.

There are two main ways to conduct a CAT simulation study: Monte Carlo simulation and post-hoc simulation or real-data simulation. According to Thompson and Weiss (2011), the difference is that Monte Carlo simulation generates the response of each examinee to each item, while post-hoc simulation utilizes real data. Thompson and Weiss (2011) concluded that one substantial drawback with post-hoc simulation is the presence of missing responses since the examinees see a small proportion of the items in the item pool. However, the Monte Carlo simulation will avoid the issue of missing responses and generate item parameters and true ability level according to the researcher's specifications. In addition, previous studies (Evans, 2010; Hol, Vorst, & Mellenbergh, 2007) found that simulated examinee responses provided similar results for the CAT when compared to real examinee responses. For these reasons, a simulated item response dataset will be used in this study.

Chapter 3 METHODS

The main goal of this study was to investigate how practical constraints such as the number of on-grade items, the proportion and the range of off-grade items, and the stopping rules affect item pool characteristics and item pool performance in adaptive testing.

3.1 Overview

As shown in Table 3.1, this study involved several independent variables: (1) three on-grade pool size (150, 300, and 500 items), (2) the proportion of off-grade items in the item pool (small, moderate, and large), (3) the range of off-grade items (one grade level and two grade levels), and (4) the stopping rules (variable- and fixed-length stopping rule). The dependent variables included item pool characteristics, which included descriptive statistics and histograms of the *b*-parameters, descriptive statistics, and plots of the TIFs and SEEs, and item pool performance, which included measurement precision, test length, and exposure properties, content blueprint fulfillment, and the mean proportion of off-grade items for each test. The details for each independent and dependent variables were discussed below.

Table 3. 1 The research design of the study

Dependent Variables	Independent Variables			
At the overall level, and for high- and low-performing simulees	On-grade items in the item pool	The proportion of off-grade items in the item pool	The range of off-grade items	Termination rules
Item pool characteristics Measurement precision Test length and exposure properties Content blueprint fulfillment Mean proportion of off-grade items for each test	150 items 300 items 500 items	Small (1/3) Moderate (2/3) Large (3/3)	One grade below and one above Two grade levels below and two above	Variable-length CAT (0.32 and 0.25) Fixed-length CAT

3.1.1 The Number of On-Grade Items in the Item Pool

The number of on-grade items was constructed for the current study. Probably the most commonly cited item pool sizes were 300 and 500 items per grade (Wang & Vispoel, 1998; Babcock & Weiss, 2013; Xing & Hambleton, 2004; Davey, 2011), where the 500-item pool targeted high-stakes summative assessment, and the 300-item pool targeted low-stakes formative assessment. Because the current study was primarily focused on the impact of incorporating off-grade items, a 150-item pool was considered. The preliminary simulation results also indicated that incorporating off-grade items has a more evident impact on a 150-item pool. These three on-grade pool sizes (150, 300, and 500 items) referred to small, moderate, and large baseline item pools. As shown in Table 3.2, three item pools were used to examine the impact of different

numbers of on-grade items on incorporating off-grade items: item pools 7, 9 and 11. These three item pools were chosen because they incorporate a similar number of items (150, 150, and 170) and the same ranges of off-grade items (two grade levels). The ratios of off-grade items to on-grade items in the item pool were 3/3, 2/3 and 1/3 for item pools 7, 9 and 11 respectively (item pool 11 incorporates 20 more off-grade items than item pools 7 and 9). The proportions of off-grade items for item pools 7, 9 and 11 reflected reality in practice, where a test organization with a small item pool could incorporate a large proportion of off-grade items and a test organization with a large item pool could incorporate a small proportion of off-grade items. The current study compared these three item pools to see whether any particular pool yields greater improvements in item pool characteristics and item pool performance compared with their baseline pools.

Table 3. 2. Item pool compositions

Item pool	grade 2	grade 3	grade 4	grade 5	grade 6
Pool 1 (SB)			150		
Pool 2 (1/3)		25	150	25	
Pool 3 (1/3)	10	15	150	15	10
Pool 4 (2/3)		50	150	50	
Pool 5 (2/3)	20	30	150	30	20
Pool 6 (3/3)		75	150	75	
Pool 7 (3/3)	30	45	150	45	30
Pool 8 (MB)			300		
Pool 9 (1/2)	30	45	300	45	30
Pool 10 (LB)			500		
Pool 11 (1/3)	35	50	500	50	35

Note. SB refers to small baseline pool; MB refers to moderate baseline pool. LB refers to large baseline pool. (1/3) refers to the ratio of off-grade items to on-grade items within the item pool.

3.1.2 Appropriate Proportion of Incorporated Off-Grade Items

An appropriate proportion of off-grade items was a variable of interest for the current study because it studied the benefit of incorporating off-grade item gains, such as measurement precision, and the issues of incorporating off-grade yield, such as increasing the mean proportion of off-grade items for each test. The current study employed three proportions of off-grade items in the item pool: small proportion, moderate proportion, and a large proportion (the ratios of the off-grade item to on-grade items in the item pool are 1/3, 2/3 and 3/3 accordingly). No studies have compared the impact of different proportions of off-grade items. As shown in Table 3.2, this study compared among small proportion (item pools 2 and 3), the moderate proportion (item pools 4 and 5), and a large proportion (item pools 6 and 7). The reason to explore the impact of different proportions of off-grade items on a small baseline item pool (150-item pool) was that the impact of incorporating off-grade items maybe more significant than that on moderate or large baseline item pool (300- or 500-item pool).

3.1.3 Appropriate Range of Incorporated Off-Grade Items

Two ranges of incorporating off-grade items were used in this study: one grade above and one grade below the target grade, two grade levels above and two grade levels below the target grade. These two ranges were selected based on previous studies (AIR, 2016; AIR, 2015; Wei & Lin, 2015; Scantron, 2004). No studies have compared the impact of different ranges of off-grade items. The current study compared item pools that incorporated off-grade items from one grade level (item pools 2, 4 and 6) with item pools that incorporated off-grade items from two grade levels (item pools 3, 5 and 7). The reason to explore the impact of different ranges of off-grade items on a small baseline item pool (150-item pool) was that the impact of incorporating off-grade items maybe more significant.

3.1.4 Termination Rules

Two stopping rules were used in this study: variable-length (fixed-SE) and fixed-length CAT. Wei and Lin (2015) investigated the impact of incorporating off-grade items under variable- and fixed-length CAT. However, the number of on-grade items and the proportion and range of off-grade items were very limited and inappropriate. Therefore, this study investigated the impact of different stopping rules on incorporating off-grade items again. This study compared the results of the same item pool under variable- and fixed-length CAT. In addition, the impact of two different SE thresholds (0.32 and 0.25) was examined.

3.2 Item Pool Compositions

There were 11 item pools in this study: (1) item pools 1-7 have 150 on-grade items, (2) item pools 8-9 have 300 on-grade items, and (3) item pools 10-11 have 500 on-grade items. Item pools 1, 8 and 11 referred to small, moderate and large baseline item pools accordingly (150, 300, 500 on-grade items) which only include on-grade items. These item pool sizes were set based on previous studies (Gonulates, 2015; Babcock & Weiss, 2013; Chajewski, 2011; Yi et al., 2001) and preliminary simulation results. Table 3.2 shows the item pool compositions in the current study. As shown, item pools 2, 4, and 6 included the same range (one grade above and one below) of off-grade items but different proportions of off-grade items in the item pool (small, moderate, and large proportions). Item pools 3, 5, 7, 9 and 11 included the same range (two grade levels above and two grade levels below) of off-grade items but different proportions of off-grade items in the item pool. The grades closest to 4th grade (grades 3 and 5) contained more items than the grades farthest from 4th grade (grades 2 and 6) for the current study. The same item pool composition was employed in both variable- and fixed-length constrained CATs.

This study used generated item pools with vertically scaled items across grades. Vertical scaling referred to the method of linking test scores that assess similar content areas but at different grade levels onto a common scale (Texas Education Agency, 2013). The simulees for CAT simulation were targeted grade 4, items restricted to grade 4 content standards were considered on-grade items, and items restricted to below-or above-grade 4 content standards were considered off-grade items. Off-grade items were realigned to the on-grade blueprints. The CAT simulation included multiple-choice items exclusively. Table 3.3 listed the means and standard deviations of the item difficulty for each grade. The gap between two grades and standard deviations of item difficulty parameters was the average result of four studies with vertical scaling (CRESST, 2017; AIR, 2016; AIR, 2015; Texas Education Agency, 2013). The mean and SD of the item difficulty parameters (*b*-parameters) for grade 2 was based on estimation since the previous studies only have grade 3 as the lowest.

Table 3. 3. Mean and Standard Deviation (SD) for item difficulty and ability, Mathematics

Grade	*b*-parameter		Ability	
	Mean	SD	Mean	SD
2	-1.50*	1.00*		
3	-0.70	1.05		
4	0.00	1.10	0.00	1.00
5	0.60	1.15		
6	1.00	1.20		

Note. * The mean and SD of item difficulty parameters for grade 2 is based on an estimation

These item pools have several notable features. First, the items in the item pool differed only by their difficulty parameter. This study adopted the 1PL model, which assumes that all items have identical discriminations and that guessing has a negligible effect. The reason for the 1PL model was to prevent overuse of highly discriminating items and equalize item exposure.

Second, the item difficulty parameters for item pools were generated randomly from a normal distribution, as many previous studies have done (Gonulates, 2015; Tay, 2015; Lim, 2010; Xiong, 2010; Xing & Hambleton, 2004). The difficulty parameters of items in the item pool were vertically scaled across grade levels. Third, except for three baseline item pools, each item pool came from a mixed distribution. For example, item pool 2 came from 25 3^{rd}-grade items from N (-0.7, 1.05), 150 4^{th}-grade items from N (0, 1.1), and 25 5^{th}-grade items from N (0.6, 1.15). Compared with item pool 1 (150 on-grade items only), item pool 2 had more test information at the two ends of the ability distribution.

3.3 Data Generation

A sample of 1,000 simulees with θ randomly generated from the standard normal distribution N (0, 1) was drawn, as established by many previous studies for their ability distribution (Kim, Moses & Yoo, 2015; Hembry, 2014; Barrada, Ponsoda, & Abad, 2010; Glas & van der Linden, 2003). The current study used R package catR (Magis et al., 2017), a package specifically designed for simulation studies on CAT, as previous studies used this for CAT simulation (Piromsombat, 2014). The current study used simulated examinee responses. The first step was obtaining item response data through Monte-Carlo simulation, where catR generate 1,000 simulees true θ from a standard normal distribution, N (0, 1) and item difficulty parameters using the mean and SD of Table 3.3. The response probability of each examinee was computed from the generated simulee's true θ value and from the generated item difficulty parameter for each item using the 1PL model in formula (4). Then this response probability was compared with a random number drawn from the standard uniform distribution, U (0, 1). If the probability was less than the random number, then a 0 would be assigned to the examinee, otherwise, a 1 would be assigned (Piromsombat, 2014; Hembry, 2014). This procedure was used to generate an item

response matrix to all items for all 1,000 simulees. Once the item response matrix was created in this manner, it was used as simulated datasets. The second step was implementing CAT simulation, where catR used those item responses data to produce item pool characteristics and performance using the CAT specifications that this study had chosen.

3.4 CAT Simulation

1. Starting point. All simulees started with an initial θ of 0 (the mean of the population). The starting rules also involved Kingsbury and Zara's (1989) CCAT method and a randomesque exposure control procedure which involved randomly selecting one of a set of five items with difficulty parameters closest to the current ability estimate. The exposure control method avoided every simulee receiving the same first items (Thompson & Weiss, 2011).

2. Item selection method. This study applied the MI method to select the next item which is the most informative item(s) at the current θ estimate. For IRT 1PL model, the MI criteria were limited to using *b*-parameters only. Kingsbury and Zara's CCAT method was utilized to help ensure each CAT consisted of the appropriate balance for all five content subdomains, as specified in Table 3.4. The item selection method involved a randomesque exposure control procedure which involved randomly selecting one of a set of five items with difficulty parameters closest to the current ability estimate. The exposure control method ensured that item overlap rate and maximum exposure rate are not too high.

3. Ability estimation. This study used expected a posteriori (EAP) method as provision ability estimation method with a normal distribution as its prior distribution, and maximum likelihood (ML) as final ability estimation (Magis et al., 2017). This study applied EAP as provision ability estimation because previous studies have concluded that ML resulted in an infinite ability estimate when the response is all correct or incorrect (Wang & Vispoel, 1998).

4. Stopping rules. For each item pool's data sets, the same simulated item response matrix ran 200 replications. Two stopping rules were simulated: fixed- and variable-length CAT. The process of ability estimation and item selection were repeated for each examinee until the termination criteria are satisfied. For example, the test ended when a 0.32 estimated standard error level is reached or a maximum of 80 items is administered in the variable-length CAT. The test ended when 40 items are administered to each examinee in the fixed-length CAT.

To alleviate the effects of the sampling error, many previous studies use 100 replications for simulation study (Chang, 2016; Yang, 2016; Kang & Tay, 2015; Risk, 2015; Luo, 2015; Hembry, 2014; Evans, 2010). To test the adequacy of the chosen number of replications, Piromsombat (2014) compared results of 100, 200, 300, 400, and 500 replications and found the results were stable after 100 replications. Thus, for one simulation condition, the results of 100, 200, 300, 400 and 500 replications were compared and found no significant differences between 100 and 500 replications. Going forward 200 replications for each simulation condition were used, and all the evaluation indices were averaged across the 200 replications.

3.5 Analyses and Comparisons

The primary purpose of this study is to investigate the optimal design of incorporating off-grade items, which results in desired results in measurement precision, test efficiency, content blueprint fulfillment, and mean proportion of off-grade items for each test in a constrained CAT. First, for the impact of different numbers of on-grade items in the item pool, as shown in Table 3.2, this study compared the baseline condition with item pools of different size (item pools 7, 9 and 11), because they incorporate a similar number and the same range of off-grade items within the small, moderate and large baseline pools respectively. This study compared the baseline item pools 1, 8 and 10 with item pools 7, 9 and 11, and examine whether item pools 7, 9 and 11 yield

improvements in item pool characteristics and item pool performance. Second, for the impact of different proportions of off-grade items, this study compared among a small proportion (item pools 2 and 3), a moderate proportion (item pools 4 and 5), and a large proportion (item pools 6 and 7). The comparison focused on whether increasing the proportions of off-grade items would yield more improvement in item pool characteristics and performance. Third, for the impact of different ranges of off-grade items, this study compared item pools that incorporated off-grade items from one grade level (item pools 2, 4 and 6) with item pools that incorporated off-grade items from two grade levels (item pools 3, 5 and 7). The comparison focused on whether broadening the range of off-grade items would yield more improvement in item pool characteristics and performance. Fourth, for the impact of different stopping rules on incorporating off-grade items, this study compared the results of the same item pool under variable- and fixed-length CAT. The comparison focused on whether variable-length or fixed-length CAT would yield more improvement in item pool performance. This study employed two SE threshold levels (0.32 and 0.25). A SE threshold of 0.32 was used for all item pools 1-11, and a SE threshold of 0.25 only was used for item pools 10 and 11 to illustrate the impact of incorporating off-grade items with a high measurement precision.

Wei and Lin (2015) found that incorporating off-grade items improved item pool performance at the overall level, and for low- and high-performing examinees. Thus, this study examined item pool characteristics and performance at the overall level, and for low- and high-performing simulees separately. One reason to examine both the overall level and for low- and high-performing simulees, was to be consistent with previous studies. Another reason is to confirm that the improvements in item pool performance of high- and low-performing simulees were not at the cost of the middle-performing simulees. This study assumed that incorporating off-grade

items would yield more improvements in item pool performance for low- and high-performing simulees than at the overall level. Thus, analysis of item pool performance started with low- and high-performing simulees then discussed at the overall level.

This study presented a benefits analysis for each factor. For example, some factors may yield the benefit of improvement in measurement precision and content blueprint fulfillment and reduce test length and maximum exposure rate. However, some factors may lead to increasing test length, decreasing content blueprint fulfillment, and increasing the mean proportion of off-grade items. Since there would be a conflict of the issue and benefit for each combination of incorporating off-grade items, it is not possible to draw any conclusions on the optimal design of incorporating off-grade items. The determination of optimizing the design of incorporating off-grade items depends on which benefit the test organization prefers and which issue they can tolerate; for example, whether they aim to prefer high measurement precision for low- and high-performing examinees over middle-performing examinees, or whether they prefer measurement precision over item exposure issue. The final results, as well as the benefit and the issue analysis for each factor related to incorporating off-grade items, would provide guidance for test organizations when they want to incorporate off-grade items into an on-grade item pool.

3.6 Evaluation Criteria

The evaluation criteria included item pool characteristics and item pool performance. The item pool characteristics included descriptive statistics and histograms of the *b*-parameters, descriptive statistics, and plots of the TIFs and SEEs. The item pool performance included measurement precision, test efficiency, content blueprint fulfillment and the mean proportion of off-grade items for each test.

3.6.1 The Item Pool Characteristics

This study summarized the item pool characteristics based on the descriptive statistics and histograms of the *b*-parameters, descriptive statistics, and plots of the TIFs and SEEs. CatR 2.6 provides a function to calculate descriptive statistics and draw histograms of the *b*-parameters, plots of the TIFs and SEEs. The variable- and fixed-length CAT were using the same generated item pools; therefore, their item pool characteristics were the same. The increasing the ranges of the *b*-parameters indicated more items appropriate for low-and high-performing simulees, and the decreasing of the average value of the SEEs indicated more measurement precision.

3.6.2 Measurement Precision

At the overall level, this study used Bias (a measure of the systematic deviation of the estimated ability from the true ability); RMSE (a measure of the absolute accuracy of parameter recovery, taking into account the Bias and SE); and the correlations between true and estimated ability for evaluating ability estimation precision. For high- and low-performing simulees, this study used Bias, RMSE, and mean standard error (a measure of the standard error of ability estimate) for evaluating ability estimation precision. All the results were averaged across 200 replications in each simulation condition.

3.6.3 Test Efficiency

At the overall level, this study used mean test length, item overlap rate, and maximum exposure rate to evaluate test efficiency. The item overlap rate referred to the average number of cases in which the same items were administered to two randomly selected examinees (Moyer, et al., 2012). Based on previous studies (He & Reckase, 2014; Moyer et al., 2012; Grossman, 2010; Chang, & Ansley, 2003), the ideal item overlap rate and maximum exposure rate would be less

than 0.3. For high- and low-performing simulees, catR did not provide exposure properties; only mean test length would be used for evaluating test efficiency.

3.6.4 Content Blueprint Fulfillment

Table 3.4 presented the expected number of items and the lower and upper bounds by the domain that was used in this study. The distribution of items in each domain was based on content standards in 4th grade. For grade 4 mathematics, items cover five content domains. For each content domain, the expected number of items was 8, and the lower and upper bounds of items were 7 and 9. For a fixed-length CAT, test length was 40 items because it matches the expected content blueprint in Table 3.4.

Table 3. 4. Math domains assessed-distribution of items drawn for grade 4

Math Domain	Domain #	The expected proportion of items	Lower-bound	Upper-bound
Operations and Algebraic Thinking	1	20%	7	9
Number and Operations in Base Ten	2	20%	7	9
Number and Operations -Fractions	3	20%	7	9
Measurement and data	4	20%	7	9
Geometry	5	20%	7	9
Total	5	100%	35	45

This study provided two indices: (1) lower-bound content blueprint fulfillment (the percentage of tests that meet the intended lower bound of blueprint for content coverage), and (2) upper-bound content blueprint fulfillment (the percentage of tests that meet the intended upper-bound of blueprint for content coverage). The current study used Kingsbury and Zara's (1989) constrained CAT (CCAT) as the content balancing method. CCAT procedure restricted the number of items eligible for selection within each content domain each time an item is selected (Leung et al., 2003). Shin et al., (2012) found that for tests with mutually exclusive content

constraints, CCAT methods had a 100% on-target rate. The preliminary simulation results also confirmed that CCAT method is successfully implemented with 100% on-target rate. Thus, this study used the test length to evaluate content blueprint fulfillment. With CCAT methods implemented with mutually exclusive content constraints, this study defined meeting lower-bound content blueprint fulfillment as a test length is longer than 35 items and meeting upper-bound content blueprint fulfillment as a test length is shorter than 45 items. The lower and upper bounds of content blueprint fulfillment were evaluated for the overall level, and for high- and low-performing simulees separately.

3.6.5 Mean proportion of off-grade items for each test

For an item pool confined for examinees from grade 4, items that are restricted to grade 4 content standards were considered on-grade items, and items that are restricted to below-or above-grade 4 content standards are considered off-grade items. This study used three indices to evaluate the mean proportion of off-grade items for each test: (1) mean proportion of off-grade items (includes grades 2, 3, 5 and 6) for each test, and (2) mean proportion of off-grade items from two grade levels only (grades 2 and 6) for each test, and (3) the number of tests that include off-grade items. The index (2) mean proportion of off-grade items from two grade levels is only available for item pools 3, 5, 7, 9, and 11. These indices were evaluated at the overall level, and for high- and low-performing simulees separately.

3.6.6 Simulation condition

A total of 11 conditions were manipulated with 200 replications simulated for each condition. All combinations of the condition were summarized in Table 3.5. Each condition was labeled to represent the simulation condition. For example, item pool 1, the small baseline pool, had the simulation condition with zero off-grade items, the ratio of the off-grade item to on-grade

item is 0, the range of off-grade items is not available, and the base form is 150 on-grade items. Item pool 2 had the simulation condition with 50 off-grade item, the ratio of the off-grade item to on-grade item is 1/3, the range of off-grade items is one grade level (grades 3 and 5) and the base form is 150 on-grade items. Item pool 9 had the simulation condition with 150 off-grade item, the ratio of the off-grade item to on-grade item is 1/2, the range of off-grade items is two grade levels (grades 2, 3, 5 and 6) and the base form is 300 on-grade items.

Table 3.5 The comparison of item pools for each studied factor

Item Pool number	Condition	Number of off-grade items	The ratio of off-grade items to on-grade items	The ranges of off-grade items	Number of on-grade items in the baseline pool
1	Baseline pool	0	0	N/A	150
2		50	(1/3)	One grade level	150
3		50	(1/3)	Two grade levels	150
4		100	(2/3)	One grade level	150
5		100	(2/3)	Two grade levels	150
6		150	(1/1)	One grade level	150
7		150	(1/1)	Two grade levels	150
8	Baseline pool	0	0	N/A	300
9		150	(1/2)	Two grade levels	300
10	Baseline pool	0	0	N/A	500
11		170	(1/3)	Two grade levels	500

Chapter 4 RESULTS

This study examines the impact of the practical constraints of incorporating off-grade items in item pool characteristics and item pool performance with respect to four factors: (1) three on-grade pool sizes (150, 300, and 500 items), (2) the proportions of off-grade items in the item pool (three ratios of off-grade items to on-grade items in the item pool are 1/3, 2/3, 3/3 respectively), (3) the ranges of off-grade items (one grade level and two grade levels), and (4) the stopping rules (variable- and fixed-length stopping rule) with two SE threshold levels (0.32 and 0.25).

This chapter first presents a summary of the CAT simulation results for each item pool. Within each of the study factors, the item pool characteristics are examined first. The item pool characteristics are summarized based on descriptive statistics and histograms of the *b*-parameters, descriptive statistics, plots of the test information functions (TIFs), and the standard errors of the estimate (SEEs). Then, the item pool performance for each studied factor is examined. The evaluation criteria for item pool performance included measurement precision (Bias, RMSE, correlation between true and estimated ability, mean standard error), test length, exposure control properties (item overlap rate and maximum exposure rates), content blueprint fulfillment, and mean proportion of off-grade items for each test. Results of item pool performance were averaged across 200 replications in each condition. This study employed two SE threshold levels (0.32 and 0.25) as stopping rules to examine stopping rule effects. All other study factors were examined with the SE threshold of 0.32 only-reliability of 0.9 based on Babcock and Weiss (2013)'s study.

Tables 4.1 through 4.13 are presented at the end of this chapter. A general introduction for each table is presented here and specific details were discussed later in the chapter. For the impact of the practical constraints of incorporating off-grade items in item pool characteristics,

Table 4.1 presents summary statistics of the *b*-parameters, the TIFs and SEEs across different baseline pools and item pools that incorporate off-grade items. Tables 4.2 to 4.13 provide summaries of performance by the pool (see Table 3.5) for each studied factor. For the impact of the practical constraints of incorporating off-grade items on item pool performance, Tables 4.2 through 4.4 indicate item pool performance of item pools 1 through 11 under variable-length CAT (SE level of 0.32) at the overall level (Table 4.2) and for low-and high-performing simulees respectively (Tables 4.3 and 4.4). Table 4.5 through 4.7 present item pool performance of item pools 1 through 9 under fixed-length CAT (40 items) at the overall level and for low-and high-performing simulees respectively. Tables 4.8 through 4.10 describe item pool performance of item pools 10 and 11 under variable-length CAT (SE level of 0.32) at the overall level and for low-and high-performing simulees respectively. Tables 4.11 through 4.13 describe item pool performance of item pools 10 and 11 under variable-length CAT (SE level of 0.25) at the overall level and for low-and high-performing simulees respectively. Item pool 10 contains 500 on-grade items and item pool 11 incorporates 170 off-grade items. Item pools 10 and 11 are used to examine the impact of different SE threshold levels (0.32 and 0.25). The details for each table and figure are discussed later for each studied factor in the chapter.

Item pools 1, 8, and 10 serve as small, moderate and large baseline pool with 150, 300, and 500 on-grade items only respectively. Item pools 7, 9, and 11 incorporate a similar number and the same range of off-grade items (two grade levels) within item pools 1, 8 and 10 respectively. Item pools 1 and 7, item pools 8 and 9, item pools 10 and 11 are used to compare the impact of incorporating off-grade items into small, moderate and large baseline pools respectively. Item pools 2 and 3 incorporate the small proportions of off-grade items. Item pools 4 and 5 incorporate the moderate proportions of off-grade items. Item pools 6 and 7 incorporate

the large proportions of off-grade items. Item pools 2, 4, and 6, and Item pools 3, 5, and 7 are used to compare the impact of different proportions of off-grade items. Item pools 2, 4, and 6 incorporate off-grade items from one grade level, while item pools 3, 5, and 7 incorporate off-grade items from two grade levels. Item pools 2, 4, and 6 and item pools 3, 5, and 7 are used to compare the impact of different ranges of off-grade items.

Table 4.2 presents a summary of the CAT simulation results for overall simulees across item pools 1 through 11 under variable-length CAT (SE threshold of 0.32). As shown in Table 4.2, the correlations between true and estimated ability are very similar (range from 0.948 to 0.952). Wang and Vispoel (1998) defined fidelity as the correlation between estimated ability and true ability and higher correlations reflect higher fidelity. The Bias across all item pools 1 to 11 are range from -0.006 to 0.013, which indicates that the true score is recovered from the estimated score to within at least 2 decimal places. The item overlap rates across item pools range from 0.07 to 0.30, which are positive findings given the ideal item overlap rate (0.3) based on previous studies. The maximum exposure rates across item pools range from 0.28 to 0.53, which are above the ideal maximum exposure rate (0.3) based on previous studies. The lower-bound blueprint fulfillment (satisfied the minimum number of items (7) specified for each content domain) across item pools are all 100%, while the upper-bound blueprint fulfillment (satisfied the maximum number of items (9) specified for each content domain) range from 92% to 98%. Higher content blueprint fulfillment reflects evidence of content comparability across forms. However, as shown in Table 4.2, the mean proportions of off-grade items for each test range from 0.22 to 0.48 for those pools that contained off-grade items. All tests included off-grade items. This may raise concerns when including too many off-grade items on tests that are designed to measure grade-level standards.

4.1. Different Size of Baseline Pools and Item Pools that Incorporate off-Grade Items

For the impact of different numbers of on-grade items in the item pool, this study compared item pools 1 and 7, item pools 8 and 9, and item pools 10 and 11 for the impact of incorporating off-grade items into small, moderate and large baseline pools respectively. Item pool 7 incorporated off-grade items within the small baseline pool (150 on-grade items); item pool 9 incorporated off-grade items within the moderate baseline pool (300 on-grade items); and item pool 11 incorporated off-grade items within the large baseline pool (500 on-grade items). Item pools 7, 9 and 11 were chosen because they incorporate approximately similar numbers of off-grade items (150) and the same ranges of off-grade items from two grade levels. In addition, the proportions of off-grade items for each item pool may reflect reality in practice, where a testing organization with a small item pool may incorporate a large proportion of off-grade items, and a testing organization with a large item pool may incorporate a small proportion of off-grade items.

4.1.1 Item Pool Characteristics

Table 4.1 and Figures 4.1 through 4.9 present the descriptive statistics and histograms of the *b*-parameters, descriptive statistics, and plots of the TIFs and SEEs across item pools. In the figures, two vertical lines at -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees. One horizontal line is used for the convenience of visual comparisons. The variable- and fixed-length CATs use the same item pools; therefore, their item pool characteristics are discussed together.

As shown in Table 4.1, the results suggest that at the overall level, compared with their baseline item pool, item pool 7 (incorporate off-grade items within the small baseline pool) results in the smaller average value of the SEE than item pools 9 and 11, which provide more

measurement precision across the θ scale. For example, Table 4.1 shows that compared with their baseline pools (item pools 1, 8 and 10), the decrease in the average value of the SEEs is 0.08 (from 0.27 to 0.19), 0.04 (0.19 to 0.15) and 0.02 (0.15 to 0.13) for item pools 7, 9 and 11 respectively. The results suggest that compared with their baseline item pools, item pool 7 yields a wider range of *b*-parameter than item pools 9 and 11, which provide more easy or difficult items at the two ends of the θ scale. More specifically, Table 4.1 shows that compared with their baseline pools (item pools 1, 8 and 10), the improvements in the ranges of the *b*-parameters are 1.54 (from 6.62 to 5.08), 0.53 (6.62 to 6.09) and 0.40 (7.90 to 7.50) for item pools 7, 9 and 11 respectively.

Additionally, the results suggest that at the overall level, compared with their baseline pools, the improvements in the maximum value of test information is similar among item pools 7, 9, and 11. The test information function indicates how well the test could estimate ability at the overall level. For example, Table 4.1 shows that compared with their baseline pools (item pools 1, 8 and 10), the improvements in the average value of the TIFs are 17.31 (from 35.00 to 17.69), 17.31 (52.56 to 35.25) and 19.62 (78.19 to 58.57) for item pools 7, 9 and 11 respectively. The improvements for item pool 11 are slightly better than those of item pools 7 and 9 likely because item pool 11 incorporated 20 additional off-grade items.

Figures 4.1 to 4.3 provide similar results to Table 4.1 on the distributions of the *b*-parameters, TIFs and SEEs at the overall level. However, the most notable characteristic of these figures demonstrates the improvements in item pool characteristics on low- and high-performing simulees. Specifically, looking at the two vertical lines at -1.8 and 1.8 on the θ scale (the approximate mean θ of low- and high-performing simulees), compared with their baseline item

pool, improvements in the TIFs and SEEs are more evident in item pool 7 than in item pools 9 and 11 for low- and high-performing simulees.

4.1.2 Item Pool Performance

The results of item pool performance for variable- and fixed-length CATs have similar patterns and are discussed together.

In terms of measurement precision, the index of measurement precision at the overall level includes Bias, RMSE, and the correlation between true and estimated ability, while that for low- and high-performing simulees are Bias, RMSE, and mean standard error.

For low- and high-performing simulees, the results suggest that compared with the baseline item pool, item pool 7 provides more improvements with respect to RMSE and mean standard error when compared to item pools 9 and 11. More specifically, Table 4.3 shows that for low-performing simulees under variable-length CAT, compared with their baseline pools (item pools 1 and 10), RMSE decreased by 0.038 (from 0.376 to 0.338) and 0.011 (0.344 to 0.333) for item pools 7 and 11 respectively. Compared with their baseline pools (item pools 1, 8 and 10), the mean standard error decreased by 0.008 (from 0.346 to 0.338), 0.002 (0.34 to 0.338) and 0 (0.337 to 0.337) for item pools 7, 9 and 11 respectively. However, at the overall level, the results suggest that compared with their baseline pools, the improvements of item pools 7, 9 and 11 on Bias, RMSE, and the correlations between true and estimated ability are negligible.

The index of test efficiency at the overall level includes the mean test length, item overlap rate, and maximum exposure rate which are shown in Table 4.2, while that for low- and high-performing simulees is mean test length only which are shown in Tables 4.3 and 4.4.

For low- and high-performing simulees, the results suggest that compared with their baseline pools, item pool 7 provides more improvements in the mean test length than those of

item pools 9 and 11. More specifically, Table 4.3 shows that for low-performing simulees, compared with their baseline pools, the mean test length decreased by 7 item (from 47 to 40), 2 items (42 to 40) and 0 item (39 to 39) items for item pools 7, 9 and 11 respectively. However, at the overall level, the results suggest that compared to their baseline pools, item pool 7 provides a decreased item overlap rate and maximum exposure rate when compared to those of item pools 9 and 11. For example, Table 4.2 shows that compared with their baseline pools (item pools 1, 8 and 10), the item overlap rate decreased by 0.14 (from 0.30 to 0.16), 0.04 (0.15 to 0.11) and 0.02 (0.09 to 0.07), and the maximum exposure rate decreased by 0.16 (from 0.53 to 0.37), no change (0.33 to 0.33), and 0.03 (0.31 to 0.28) for item pools 7, 9 and 11 respectively.

In terms of content blueprint fulfillment, the index of content blueprint fulfillment includes the lower and upper bounds of content blueprints fulfillment. The lower and upper bounds of content blueprints are 100% satisfied with fixed-length CAT and the lower-bound of content blueprints are 100% satisfied under variable-length CAT. Thus, the discussion only focuses on the fulfillment of upper-bound of content blueprints under variable-length CAT.

For low-and high-performing simulees, the results suggest that compared with their baseline item pools, item pool 7 provides more improvements in the upper-bound content blueprint fulfillment than those of item pools 9 and 11. More specifically, Table 4.4 shows that for high-performing simulees, compared with their baseline pools, the improvements in upper-bound content blueprint fulfillment are 28% (from 66% to 94%), 6% (91% to 97%), and 1% (97% to 98%) for item pools 7, 9 and 11 respectively. However, at the overall level, the results suggest that compared to their baseline pools, item pool 7 provides more improvements in upper-bound content blueprint fulfillment than those of item pools 9 and 11. More specifically, Table 4.2 shows that compared with their baseline pools, the improvements in upper-bound content

blueprint fulfillment are 7% (from 92% to 99%) for item pool 7. For item poos 9 and 11, due to the strong fulfillment, there was little room for improvement (1% (98% to 99%), and no change (100% to 100%) for item pools 9 and 11 respectively).

In terms of mean proportion of off-grade items for each test, the index includes: (1) mean proportion of off-grade items (includes grades 2, 3, 5 and 6) for each test, (2) mean proportion of off-grade items from two grade levels only (grades 2 and 6) for each test, and (3) the number of tests that include off-grade items. The reason to use an index (2) is that compared to off-grade items from one grade level (grades 3 and 5), it would be a more serious challenge for tests that are designed to measure grade-level standards.

For low- and high-performing simulees, the results suggest compared with their baseline pools, item pool 7 provides a higher mean proportion of off-grade items for each test and a higher mean proportion of off-grade items from two grade levels than those of item pools 9 and 11. More specifically, Table 4.3 shows that for low-performing simulees under variable-length CAT, the mean proportions of off-grade items for each test are 46%, 32% and 25% for item pools 7, 9 and 11 respectively. The mean proportions of off-grade items from two grade levels are 26%, 18% and 13% for item pools 7, 9 and 11 respectively.

4.2. Item Pools that Incorporated Different Proportions of Off-Grade Items

To better understand the impact of different proportions of off-grade items, this study compared three proportions off-grade items: small proportion (item pools 2 and 3), moderate proportion (item pools 4 and 5), and large proportion (item pools 6 and 7)(three ratios of off-grade items to on-grade items in the item pool are 1/3, 2/3, 3/3 respectively). As shown in Figures 4.4 through 4.6, item pools 2, 4 and 6, and item pools 3, 5 and 7 are used to compare the impact of different proportions of off-grade items. The item pool characteristics and performance

of item pools 2, 4 and 6 are similar to item pools 3, 5 and 7 since they incorporate the same proportions of off-grade items. Thus, the current study only chose item pools 3, 5 and 7 for illustration.

4.2.1 Item Pool Characteristics

As shown in Table 4.1, the results suggest that at the overall level, increasing the proportions of off-grade items provides improvements in the average value of the TIFs and the SEEs, and improvements in the range of the *b*-parameters. More specifically, Table 4.1 shows that the average value of the TIFs (across theta value -4 to +4) for item pools 3, 5 and 7 are 23.45, 29.21, and 35.00 respectively, while the average value of the SEEs for item pools 3, 5 and 7 are 0.23, 0.21, and 0.19 respectively. Table 4.1 shows that the range of the *b*-parameters for item pools 3, 5 and 7 are 5.94, 6.62 and 6.62 respectively. Even the range of the b-parameter for Item 5 and 7 are same, but item pool 7 have more items at the two ends of the distribution (appropriated for low- and high-performing simulee) than item pool 5.

Figures 4.4 to 4.6 provide similar results to Table 4.1. For low- and high-performing simulees, increasing the proportions of off-grade items provides improvements in the TIFs and SEEs and improvements in the distribution of the *b*-parameters. The improvements are more obvious for low- and high-performing simulees than those of the overall level.

4.2.2 Item Pool Performance

The results of item pool performance from variable- and fixed-length CAT have similar patterns and are discussed together.

In terms of measurement precision, for low- and high-performing simulees, the results suggest that increasing the proportions of off-grade items shows consistent improvements in mean standard errors. For example, Table 4.4 shows that for high-performing simulees under

variable-length CAT, the mean standard errors are 0.341, 0.339 and 0.338 for item pools 3, 5 and 7 respectively. At the overall level, as shown in Table 4.5, the results suggest that increasing the proportions of off-grade items does not show consistent improvements in Bias, RMSE, and correlation between true and estimated ability.

In terms of test efficiency, for low- and high-performing simulees, the results suggest that increasing the proportions of off-grade items shows negligible improvements in mean test length, and the improvements are 1-2 items. At the overall level, the results suggest that increasing the proportions of off-grade items shows consistent improvements in the item overlap rate and maximum exposure rate, but not in mean test length. For example, Table 4.2 shows that the item overlap rates are 0.23, 0.18 and 0.16 for item pools 3, 5 and 7 respectively, and the maximum exposure rates are 0.47, 0.41 and 0.37 for item pools 3, 5 and 7 respectively.

In terms of content blueprint fulfillment, for low- and high-performing simulees, the results suggest that increasing the proportions of off-grade items shows consistent improvements in upper-bound content blueprint fulfillment. For example, Table 4.4 shows that for high-performing simulees, the upper-bound content blueprint fulfillments are 87%, 92% and 94% for item pools 3, 5 and 7 respectively. At the overall level, the results suggest that increasing the proportions of off-grade items does not show consistent improvements in upper-bound content blueprint fulfillment.

Table 4.3 shows that for low-performing simulees, the mean proportions of off-grade items are 24%, 38%, and 46% for each test for item pools 3, 5 and 7 respectively, and the mean proportions of off-grade items from two grade levels are 13%, 21%, and 26% for item pools 3, 5 and 7 respectively. For example, Table 4.2 shows that the mean proportions of off-grade items are 22%, 37%, and 46% for each test for item pools 3, 5 and 7 respectively, and the mean

proportions of off-grade items from two grade levels are 4%, 7%, and 8% for each test for item pools 3, 5 and 7 respectively. All tests have off-grade items.

4.3. Item Pools that Incorporated Different Ranges of Off-Grade Items

For the impact of different proportions of off-grade items, this study compared item pools that incorporated off-grade items from one grade level (item pools 2, 4 and 6) with item pools that incorporated off-grade items from two grade levels (item pools 3, 5 and 7).

4.3.1 Item Pool Characteristics

As shown in Table 4.1, for the overall level, broadening the range of off-grade items (from one grade level to two grade levels) provides improvements in the range of the *b*-parameters, but not on the average value of the TIFs and the SEEs. For example, Table 4.1 shows that compared with item pools 2, 4, 6 (one grade level off-grade items), item pools 3, 5 and 7 (two grade levels off-grade items) broaden the range of *b*-parameters by 0.27 (from 5.67 to 5.94), 0.95 (5.67 to 6.62), and 0.23 (6.39 to 6.62). The improvements in the range of the *b*-parameters indicate additional items are appropriate for low- and high-performing simulees.

Figures 4.7 to 4.9 provide similar results to Table 4.1. For low- and high-performing simulees, broadening the range of off-grade items provides moderate improvements in the range of the *b*-parameters, but negligible improvements in the TIFs and SEEs. As shown in Table 4.2, the improvements in the *b*-parameters, the TIFs and SEEs are negligible at the overall level.

4.3.2 Item Pool Performance

The results of item pool performance from variable- and fixed-length CAT have similar patterns and are discussed together.

In terms of measurement precision, for low- and high-performing simulees, the results suggest that broadening the range of off-grade items shows slight improvements in mean

standard errors at the second decimal place. For example, Table 4.6 shows that for low-performing simulees under fixed-length CAT, compared with item pools 2, 4, 6 (incorporate one grade level off-grade items), item pools 3, 5 and 7 (two grade levels off-grade items) decrease mean standard error by 0.18 (from 0.364 to 0.346), 0.12 (0.357 to 0.345) and 0.01 (0.342 to 0.341) respectively. However, the results suggest that at the overall level, as shown in Table 4.2, broadening the range of off-grade items does not show consistent improvements in Bias, RMSE, or the correlations between true and estimated ability. The comparison between item pools 2 and 3 is a good example.

In terms of test efficiency, for low- and high-performing simulees, the results suggest that broadening the range of off-grade items does not show consistent improvements in mean test length. At the overall level, Table 4.2 shows that increasing the proportions of off-grade items does not show consistent improvements in mean test length and exposure properties.

In terms of content blueprint fulfillment, for low- and high-performing simulees under variable-length CAT, the results suggest that broadening the range of off-grade items shows consistent improvements in upper-bound content blueprint fulfillment. More specifically, Table 4.3 shows that for lower-performing simulees under variable-length CAT, compared with item pools 2, 4, and 6, the improvements in upper-bound content blueprint fulfillment are 18% (from 71% to 89%), 14% (76% to 90%) and 0% (93% to 93%) for item pools 3, 5 and 7 respectively. However, at the overall level, the results suggest that broadening the range of off-grade items does not show consistent improvements in upper-bound content blueprint fulfillment.

Table 4.3 shows that for low-performing simulees under variable-length CAT, compared with item pools 2, 4, and 6, the mean proportions of off-grade items for each test increased 8% (16% to 24%) , 3% (35% to 38%) and 2% (44% to 46%) for item pools 3, 5 and 7 respectively.

The mean proportions of off-grade items from two grade levels increased for item pools 3, 5 and 7 are 13%, 21%, and 26% respectively. At the overall level, the results suggest that broadening the range of off-grade items does not show consistent improvements in the mean proportion of off-grade items for each test. Table 4.2 shows that the mean proportions of off-grade items from two grade levels for item pools 3, 5 and 7 are 4%, 7%, and 8% respectively. All tests have off-grade items.

4.4. Two Stopping Rules with Two SE Threshold Levels

For the impact of different stopping rules on incorporating off-grade items, the comparison focuses only on whether variable-length or fixed-length CAT yields improvements in item pool performance. In addition, the impact of two different SE threshold levels (0.32 and 0.25) under variable-length CAT was examined.

4.4.1 Item Pool Characteristics

Figure 4.10 illustrates the conditional test length and the conditional standard error of ability estimation of the deciles of the true ability levels with SE threshold of 0.25 and 0.32. As shown in Figure 4.10, the conditional test length is longer at the two ends of the ability distribution than at the middle of the ability distribution. The conditional standard error of ability estimation is smaller at the two ends of the ability distribution than at the middle of the ability distribution. These results are similar to the results in Tables 4.8 through 4.13.

4.4.2 Item Pool Performance

In terms of measurement precision, for low- and high-performing simulees, the results suggest that compared with a fixed-length CAT, a variable-length CAT shows consistent improvements in RMSE and mean standard errors. For example, Tables 4.3 and 4.6 show that for low-performing simulees, compared with the fixed-length CATs, the variable-length CATs

decreases RMSE by 0.042 (from 0.409 to 0.367), 0.027 (from 0.364 to 0.337) and 0.005 (from 0.342 to 0.337), mean standard error by 0.021 (0.364 to 0.343), 0.015 (0.357 to 0.342), and 0.003 (0.342 to 0.339) for item pools 2, 4 and 6 respectively. However, as shown in Tables 4.2 and 4.5, at the overall level, the results suggest that compared with a fixed-length CAT, variable-length CAT does not show consistent improvements in Bias, RMSE, nor correlations between true and estimated ability.

Additionally, the results suggest that for low- and high-performing simulees, compared with a SE threshold level of 0.32, a SE threshold level of 0.25 shows slight improvements in Bias, RMSE and mean standard error. More specifically, Tables 4.9 and 4.12 show that for low-performing simulees, compared to a SE threshold level of 0.32, a SE threshold level of 0.25 decreases Bias, RMSE and mean standard error by 0.004 (from 0.012 to 0.008), 0.07 (0.333 to 0.263) and 0.079 (0.337 to 0.258) respectively for item pool 11. At the overall level, as shown in Tables 4.8 and 4.11, the results suggest that compared to a SE threshold level of 0.32, a SE threshold level of 0.25 shows slight improvements in RMSE and correlations between true and estimated ability. For example, Tables 4.8 and 4.11 show that compared to a SE threshold level of 0.32, a SE threshold level of 0.25 decreases RMSE by 0.078 (0.331 to 0.253) and increases the correlations between true and estimated ability by 0.02 (0.950 to 0.970) respectively for item pool 11.

In terms of test efficiency, for low- and high-performing simulees, the results suggest that fixed-length CAT shows consistent improvements in mean test length. For example, Table 4.3 shows that for low-performing simulees, compared to a fixed-length CAT, the mean test length of variable-length CAT increases by 5 items (40 to 45), 3 items (40 to 43), and 1 item (40 to 41) for item pools 2, 4 and 6 respectively. At the overall level, the results suggest that variable-length CAT

shows negligible improvements in mean test length and exposure properties. More specifically, Tables 4.2 and 4.5 show that variable-length CAT decreases the mean test length by 1-2 items, item overlap rate by 0.01 and maximum exposure rate by 0.01-0.02 for item pools 2 through 7.

Additionally, the results suggest that for low- and high-performing simulees, compared with a SE threshold level of 0.32, a SE threshold level of 0.25 shows an increase in mean test length. More specifically, Tables 4.9 and 4.12 show that for low-performing simulees, compared to a SE threshold level of 0.32, a SE threshold level of 0.25 increases mean test length by 27 items (from 39 to 66 items) for item pool 11. At the overall level, the results suggest that compared with a SE threshold level of 0.32, a SE threshold level of 0.25 shows an increase in mean test length and exposure properties. For example, Tables 4.8 and 4.11 show that compared to a SE threshold level of 0.32, a SE threshold level of 0.25 increases mean test length, item overlap rate and maximum exposure rate by 25 items (from 38 to 63), 0.04 (0.07 to 0.11) and 0.04 (0.28 to 0.32) respectively for item pool 11.

In terms of content blueprint fulfillment, the fixed-length CATs yields 100% lower- and upper- bound content blueprint fulfillment. As shown in Tables 4.2 through 4.4, under a SE threshold of 0.32, the variable-length CATs yields 100% lower-bound content blueprint fulfillment and 58% to 98% upper-bound content blueprint fulfillment across item pools 1 to 11 for low- and high-performing simulees, and 92% to 99% upper-bound content blueprint fulfillment at the overall level. As shown in Tables 4.11 through 4.13, under a SE threshold of 0.25, the variable-length CATs yields 100% lower-bound content blueprint fulfillment and 0% upper-bound content blueprint fulfillment for item pools 10 and 11 for low- and high-performing simulees and at the overall level. This is expected as the mean test length is 63 to 67 items for the overall level, and for low- and high-performing simulees under a SE threshold of 0.25. Thus, the results suggest that

for low- and high-performing simulees and at the overall level, compared with a SE threshold level of 0.32, a SE threshold level of 0.25 shows a decrease on upper-bound content blueprint fulfillment for item pools 10 and 11.

In terms of the mean proportion of off-grade items for each test, the difference between fixed- and variable-length CAT in both mean proportion of off-grade items and mean proportion of off-grade items from two grade levels is negligible at the overall level, and for low- and high-performing simulees. All tests have off-grade items.

Additionally, for low- and high-performing simulees, the results suggest that compared with a SE threshold level of 0.32, a SE threshold level of 0.25 shows an increase in both the mean proportion of off-grade items and mean proportion of off-grade items from two grade levels only. More specifically, Tables 4.10 and 4.13 indicate that for high-performing simulees, compared to a SE threshold level of 0.32, a SE threshold level of 0.25 increases 3% (from 26% to 29%) of off-grade for each test, and 1% (12% to 13%) of off-grade items from two grade levels only. As shown in Tables 4.8 and 4. 11, at the overall level, the difference between SE threshold levels of 0.25 and 0.32 in both mean proportion of off-grade items and mean proportion of off-grade items from two grade levels is negligible.

Table 4. 1. Summary descriptive statistics of the *b*-parameters, TIFs, and SEEs

	The *b*-parameters					TIF					SEE				
	Mean	SD	Min	Max	Range	Average	SD	Min	Max	Range	Average	SD	Min	Max	Range
Pool 1	0.02	1.00	-2.44	2.64	5.08	17.69	9.16	3.82	31.10	30.10	0.27	0.09	0.18	0.51	0.33
Pool 2	0.05	1.06	-3.03	2.64	5.67	23.51	11.86	5.25	40.72	39.72	0.24	0.08	0.16	0.44	0.28
Pool 3	0.02	1.10	-3.03	2.91	5.94	23.45	11.56	5.54	40.12	39.12	0.23	0.07	0.16	0.42	0.26
Pool 4	0.04	1.05	-3.03	2.64	5.67	29.40	14.87	6.58	50.97	49.97	0.21	0.07	0.14	0.39	0.25
Pool 5	0.02	1.16	-3.71	2.91	6.62	29.21	14.02	7.25	49.22	48.22	0.21	0.06	0.14	0.37	0.23
Pool 6	0.05	1.11	-3.03	3.36	6.39	35.16	17.34	8.05	60.14	59.14	0.19	0.06	0.13	0.35	0.22
Pool 7	-0.01	1.19	-3.71	2.91	6.62	35.00	16.58	8.87	58.52	57.52	0.19	0.06	0.13	0.34	0.21
Pool 8	-0.04	1.06	-3.18	2.91	6.09	35.25	17.77	7.95	61.07	60.07	0.19	0.06	0.13	0.35	0.22
Pool 9	0.01	1.17	-3.71	2.91	6.62	52.56	25.19	13.19	88.49	87.49	0.15	0.05	0.11	0.28	0.17
Pool 10	0.03	1.11	-3.31	4.19	7.50	58.57	28.87	13.89	100.19	99.19	0.15	0.05	0.10	0.27	0.17
Pool 11	0.01	1.18	-3.71	4.19	7.90	78.19	37.31	19.79	131.30	130.30	0.13	0.04	0.09	0.22	0.13

Note. The *b*-parameters describe item pool characteristics, while TIFs and SEEs present the item pool performance. These statistics are based on the theta range from -4 to +4. Figures 4.1 to 4.9 are based on this table.

Table 4. 2. Means (SDs) of item pool performance under variable-length CAT: at the overall level (SE threshold of 0.32)

All simulees (V)	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9	**Pool 10**	Pool 11
Bias	-0.006 (0.007)	0.012 (0.008)	-0.001 (0.008)	0.004 (0.008)	0.006 (0.009)	0.013 (0.008)	0.001 (0.009)	-0.004 (0.009)	-0.005 (0.010)	0.007 (0.008)	0.001 (0.010)
RMSE	0.336 (0.006)	0.335 (0.006)	0.332 (0.006)	0.339 (0.006)	0.333 (0.006)	0.331 (0.008)	0.336 (0.006)	0.337 (0.008)	0.333 (0.007)	0.337 (0.007)	0.331 (0.008)
Correlation	0.952 (0.002)	0.951 (0.002)	0.948 (0.002)	0.952 (0.002)	0.951 (0.002)	0.951 (0.002)	0.949 (0.002)	0.950 (0.002)	0.950 (0.002)	0.950 (0.002)	0.950 (0.002)
Mean Test Length	40 (0.0)	39 (0.0)	39 (0.0)	39 (0.0)	38 (0.0)	38 (0.0)	38 (0.0)	38 (0.0)	38 (0.0)	38 (0.0)	38 (0.0)
Item Overlap Rate	0.30 (0.00)	0.23 (0.00)	0.23 (0.00)	0.18 (0.00)	0.18 (0.00)	0.15 (0.00)	0.16 (0.00)	0.15 (0.00)	0.11 (0.00)	0.09 (0.00)	0.07 (0.00)
Maximum Exposure Rate	0.53 (0.01)	0.45 (0.01)	0.47 (0.01)	0.40 (0.01)	0.41 (0.02)	0.35 (0.01)	0.37 (0.01)	0.33 (0.01)	0.33 (0.01)	0.31 (0.01)	0.28 (0.01)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	0.92 (0.00)	1.00 (0.00)	0.98 (0.00)	0.96 (0.00)	0.98 (0.00)	0.99 (0.00)	0.99 (0.00)	0.98 (0.00)	0.99 (0.00)	1 (0)	1 (0)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.22 (0.00)	0.22 (0.00)	0.39 (0.00)	0.37 (0.00)	0.48 (0.00)	0.46 (0.00)	0 (0)	0.31 (0.00)	0 (0)	0.24 (0.00)
Mean Proportions of Off-grade Items from Two Grade Levels Only	0 (0)	0 (0)	0.04 (0.00)	0 (0)	0.07 (0.00)	0 (0)	0.08 (0.00)	0 (0)	0.06 (0.00)	0 (0)	0.04 (0.00)

Number of Tests that Include Off-Grade Items	0 (0)	1000 (0.07)	1000 (0)	1000 (0)	1000 (0)	1000 (0)	1000 (0)	0 (0)	1000 (0)	0 (0)	1000 (0.3)

Table 4. 3. Means (SDs) of item pool performance under variable-length CAT: for low-performing simulees (SE threshold of 0.32)

Low-Performing Simulees	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9	**Pool 10**	Pool 11
Bias	0.017 (0.018)	0.011 (0.020)	0.055 (0.022)	-0.023 (0.018)	0.053 (0.022)	0.037 (0.025)	-0.023 (0.026)	-0.031 (0.022)	-0.015 (0.029)	0.064 (0.024)	0.012 (0.032)
RMSE	0.376 (0.016)	0.367 (0.017)	0.323 (0.019)	0.337 (0.018)	0.334 (0.018)	0.337 (0.021)	0.338 (0.021)	0.320 (0.017)	0.342 (0.021)	0.344 (0.019)	0.333 (0.024)
Mean Standard Error	0.346 (0.001)	0.343 (0.001)	0.339 (0.000)	0.342 (0.001)	0.340 (0.001)	0.339 (0.001)	0.338 (0.001)	0.340 (0.001)	0.338 (0.001)	0.337 (0.000)	0.337 (0.000)
Mean Test Length	47 (0.3)	45 (0.2)	42 (0.2)	43 (0.2)	41 (0.2)	41 (0.2)	40 (0.2)	42 (0.2)	40 (0.1)	39 (0.1)	39 (0.1)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	0.58 (0.03)	0.71 (0.02)	0.89 (0.02)	0.76 (0.02)	0.90 (0.02)	0.93 (0.01)	0.93 (0.01)	0.87 (0.02)	0.97 (0.01)	0.99 (0.01)	0.98 (0.01)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.16 (0.00)	0.24 (0.01)	0.35 (0.00)	0.38 (0.01)	0.44 (0.01)	0.46 (0.01)	0 (0)	0.32 (0.01)	0 (0)	0.25 (0.01)
Mean Proportion of Off-grade Items from Two Grade levels only	0 (0)	0 (0)	0.13 (0.00)	0 (0)	0.21 (0.00)	0 (0)	0.26 (0.01)	0 (0)	0.18 (0.01)	0 (0)	0.13 (0.00)
Number of Tests that Include Off-Grade Items	0 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	0 (0)	100 (0)	0 (0)	100 (0)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 4. Means (SDs) of item pool performance under variable-length CAT: for high-performing simulees (SE threshold of 0.32)

High-Performing Simulees (V)	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9	**Pool 10**	Pool 11
Bias	-0.036 (0.019)	0.010 (0.020)	-0.008 (0.021)	0.003 (0.025)	-0.008 (0.021)	0.015 (0.024)	-0.043 (0.025)	-0.011 (0.024)	-0.004 (0.027)	0.001 (0.026)	-0.033 (0.029)
RMSE	0.341 (0.02)	0.343 (0.018)	0.364 (0.021)	0.337 (0.018)	0.329 (0.019)	0.342 (0.02)	0.333 (0.02)	0.357 (0.02)	0.320 (0.021)	0.342 (0.019)	0.345 (0.025)
Mean Standard Error	0.347 (0.001)	0.342 (0.001)	0.341 (0.001)	0.342 (0.000)	0.339 (0.001)	0.339 (0.001)	0.338 (0.000)	0.339 (0.000)	0.337 (0.001)	0.338 (0.001)	0.337 (0.001)
Mean Test Length	45 (0.2)	43 (0.2)	42 (0.2)	41 (0.2)	41 (0.1)	40 (0.2)	40 (0.2)	41 (0.2)	39 (0.1)	40 (0.1)	39 (0.1)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	0.66 (0.03)	0.85 (0.02)	0.87 (0.02)	0.89 (0.02)	0.92 (0.02)	0.93 (0.01)	0.94 (0.01)	0.91 (0.02)	0.97 (0.01)	0.97 (0.01)	0.98 (0.01)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.24 (0.00)	0.23 (0.01)	0.38 (0.01)	0.41 (0.01)	0.45 (0.01)	0.50 (0.01)	0 (0)	0.35 (0.01)	0 (0)	0.26 (0.01)
Mean Proportion of Off-grade Items from Two Grade levels only	0 (0)	0 (0)	0.11 (0.00)	0 (0)	0.18 (0.01)	0 (0)	0.23 (0.01)	0 (0)	0.15 (0.01)	0 (0)	0.12 (0.00)
Number of Tests that Include Off-Grade Items	0 (0)	100 (0.07)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	0 (0)	100 (0)	0 (0)	100 (0.1)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 5. Means (SDs) of item pool performance under fixed-length CAT: at the overall level

Overall Simulees (F)	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9
Bias	-0.007 (0.007)	0.009 (0.007)	-0.002 (0.007)	0.002 (0.008)	0.007 (0.008)	0.012 (0.008)	0.000 (0.008)	-0.005 (0.008)	-0.006 (0.009)
RMSE	0.339 (0.007)	0.337 (0.006)	0.33 (0.007)	0.337 (0.007)	0.328 (0.007)	0.326 (0.007)	0.330 (0.007)	0.332 (0.007)	0.327 (0.007)
Correlation	0.952 (0.002)	0.952 (0.002)	0.949 (0.002)	0.953 (0.002)	0.953 (0.002)	0.953 (0.002)	0.950 (0.002)	0.953 (0.002)	0.952 (0.002)
Mean Test Length	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)
Item Overlap Rate	0.31 (0.00)	0.24 (0.00)	0.24 (0.00)	0.19 (0.00)	0.19 (0.00)	0.16 (0.00)	0.16 (0.00)	0.16 (0.00)	0.11 (0.00)
Maximum Exposure Rate	0.54 (0.01)	0.46 (0.01)	0.48 (0.01)	0.41 (0.01)	0.43 (0.02)	0.36 (0.01)	0.38 (0.01)	0.34 (0.01)	0.33 (0.01)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.22 (0.00)	0.22 (0.00)	0.39 (0.00)	0.37 (0.00)	0.48 (0.00)	0.46 (0.00)	0 (0)	0.31 (0.00)
Mean Proportion of Off-grade Items from Two Grade Levels Only	0 (0)	0 (0)	0.04 (0.00)	0 (0)	0.07 (0.00)	0 (0)	0.08 (0.00)	0 (0)	0.06 (0.00)
Number of Tests that Include Off-Grade Items	0 (0)	1000 (0.12)	1000 (0)	1000 (0)	1000 (0)	1000 (0)	1000 (0)	0 (0)	1000 (0)

Table 4. 6. Means (SDs) of item pool performance under fixed-length CAT: for low-performing simulees

Low-Performing Simulees	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9
Bias	-0.011 (0.023)	-0.026 (0.022)	0.041 (0.023)	-0.049 (0.022)	0.036 (0.024)	0.020 (0.024)	-0.032 (0.023)	-0.053 (0.024)	-0.029 (0.028)
RMSE	0.407 (0.025)	0.409 (0.023)	0.329 (0.021)	0.364 (0.022)	0.340 (0.019)	0.342 (0.020)	0.347 (0.023)	0.34 (0.018)	0.345 (0.021)
Mean Standard Error	0.372 (0.002)	0.364 (0.002)	0.346 (0.001)	0.357 (0.002)	0.345 (0.001)	0.342 (0.001)	0.341 (0.001)	0.349 (0.001)	0.336 (0.001)
Mean Test Length	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.16 (0.00)	0.24 (0.01)	0.36 (0.01)	0.38 (0.01)	0.44 (0.01)	0.46 (0.01)	0 (0)	0.32 (0.01)
Mean Proportion of Off-grade Items from Two Grade levels only	0 (0)	0 (0)	0.13 (0.00)	0 (0)	0.21 (0.00)	0 (0)	0.26 (0.01)	0 (0)	0.18 (0.01)
Number of Tests that Include Off-Grade Items	0 (0)	100 (0.07)	100 (0.07)	100 (0)	100 (0)	100 (0)	100 (0)	0 (0)	100 (0)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 7. Means (SDs) of item pool performance under fixed-length CAT: for high-performing simulees

High-Performing Simulees	**Pool 1**	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7	**Pool 8**	Pool 9
Bias	-0.009 (0.021)	0.032 (0.025)	0.011 (0.020)	0.017 (0.025)	0.009 (0.026)	0.028 (0.024)	-0.037 (0.024)	0.003 (0.027)	0.004 (0.025)
RMSE	0.358 (0.021)	0.367 (0.022)	0.382 (0.021)	0.350 (0.021)	0.341 (0.020)	0.348 (0.019)	0.336 (0.021)	0.362 (0.021)	0.321 (0.020)
Mean Standard Error	0.367 (0.002)	0.354 (0.002)	0.349 (0.001)	0.347 (0.001)	0.344 (0.001)	0.340 (0.001)	0.338 (0.001)	0.342 (0.001)	0.336 (0.001)
Mean Test Length	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)	40 (0)
Content Blueprint Fulfillment(LB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Content Blueprint Fulfillment(UB)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Mean Proportion of Off-Grade Items for Each Test	0 (0)	0.24 (0.00)	0.23 (0.00)	0.38 (0.01)	0.41 (0.01)	0.45 (0.01)	0.50 (0.01)	0 (0)	0.35 (0.01)
Mean Proportion of Off-grade Items from Two Grade levels only	0 (0)	0 (0)	0.11 (0.00)	0 (0)	0.18 (0.01)	0 (0)	0.23 (0.01)	0 (0)	0.15 (0.01)
Number of Tests that Include Off-Grade Items	0 (0)	100 (0.1)	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)	0 (0)	100 (0)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 8. Means (SDs) of item pool performance under variable-length CAT: at the overall level (SE threshold of 0.32)

Item pool	Bias	RMSE	Correlation	Mean Test Length	Item Overlap Rate	Maximum Exposure Rate	Content Blueprint Fulfillment(LB)	Content Blueprint Fulfillment(UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.007 (0.008)	0.337 (0.007)	0.950 (0.002)	38 (0.0)	0.09 (0.00)	0.31 (0.01)	1 (0)	1 (0)	0 (0)	0 (0)	0 (0)
Pool 11	0.001 (0.010)	0.331 (0.008)	0.950 (0.002)	38 (0.0)	0.07 (0.00)	0.28 (0.01)	1 (0)	1 (0)	0.24 (0.00)	0.04 (0.00)	1000 (0.3)

Table 4. 9. Means (SDs) of item pool performance under variable-length CAT: for low-performing simulees (SE threshold of 0.32)

Item pool	Bias	RMSE	Mean Standard Error	Mean Test Length	Content Blueprint Fulfillment (LB)	Content Blueprint Fulfillment (UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.064 (0.024)	0.344 (0.019)	0.337 (0.000)	39 (0.1)	1 (0)	0.99 (0.01)	0 (0)	0 (0)	0 (0)
Pool 11	0.012 (0.032)	0.333 (0.024)	0.337 (0.000)	39 (0.1)	1 (0)	0.98 (0.01)	0.25 (0.01)	0.13 (0.00)	100 (0.07)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 10. Means (SDs) of item pool performance under variable-length CAT: for high-performing simulees (SE threshold of 0.32)

Item pool	Bias	RMSE	Mean Standard Error	Mean Test Length	Content Blueprint Fulfillment (LB)	Content Blueprint Fulfillment (UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.001 (0.026)	0.342 (0.019)	0.338 (0.001)	40 (0.1)	1 (0)	0.97 (0.01)	0 (0)	0 (0)	0 (0)
Pool 11	-0.033 (0.029)	0.345 (0.025)	0.337 (0.001)	39 (0.1)	1 (0)	0.98 (0.01)	0.26 (0.01)	0.12 (0.00)	100 (0.1)

Table 4. 11. Means (SDs) of item pool performance under variable-length CAT: at the overall level (SE threshold of 0.25)

Item pool	Bias	RMSE	Correlation	Mean Test Length	Item Overlap Rate	Maximum Exposure Rate	Content Blueprint Fulfillment(LB)	Content Blueprint Fulfillment(UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.007 (0.006)	0.263 (0.005)	0.969 (0.001)	64 (0.0)	0.15 (0.00)	0.36 (0.01)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pool 11	0.001 (0.007)	0.253 (0.005)	0.970 (0.001)	63 (0.0)	0.11 (0.00)	0.32 (0.01)	1 (0)	0 (0)	0.24 (0.00)	0.04 (0.00)	1000 (0)

Table 4. 12. Means (SDs) of item pool performance under variable-length CAT: for low-performing simulees (SE threshold of 0.25)

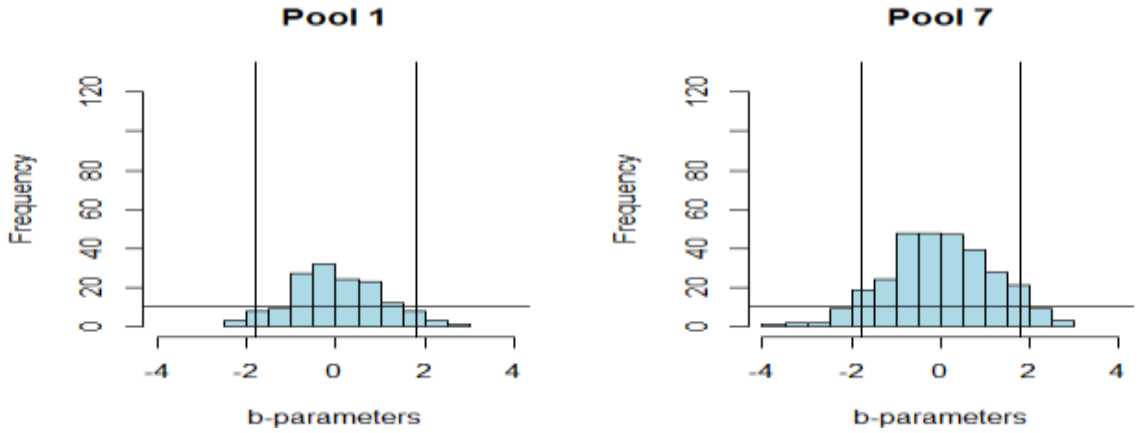
Item pool	Bias	RMSE	Mean Standard Error	Mean Test Length	Content Blueprint Fulfillment (LB)	Content Blueprint Fulfillment (UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.064 (0.015)	0.271 (0.013)	0.258 (0.000)	67 (0.1)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pool 11	0.008 (0.019)	0.263 (0.014)	0.258 (0.000)	66 (0.1)	1 (0)	0 (0)	0.26 (0.00)	0.14 (0.00)	100 (0)

Note. -1.8 and 1.8 on the θ scale are the approximate mean θ of low- and high-performing simulees

Table 4. 13. Means (SDs) of item pool performance under variable-length CAT: for high-performing simulees (SE threshold of 0.25)

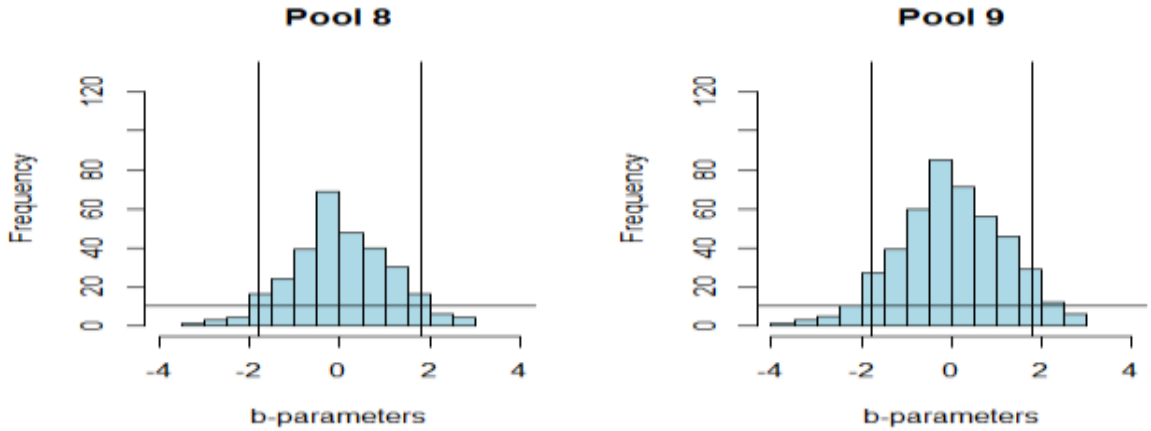
Item pool	Bias	RMSE	Mean Standard Error	Mean Test Length	Content Blueprint Fulfillment (LB)	Content Blueprint Fulfillment (UB)	Mean Proportion of Off-Grade Items for Each Test	Mean Proportion of Off-Grade Items from Two Grade Levels Only	Number of Tests that Include Off-Grade Items
Pool 10	0.010 (0.015)	0.265 (0.013)	0.259 (0.001)	67 (0.1)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pool 11	-0.026 (0.020)	0.265 (0.014)	0.258 (0.000)	65 (0.1)	1 (0)	0 (0)	0.29 (0.01)	0.13 (0.00)	100 (0)

Figure 4. 1. Histograms of the *b*-parameters across different baseline pools and item pools that incorporate off-grade items

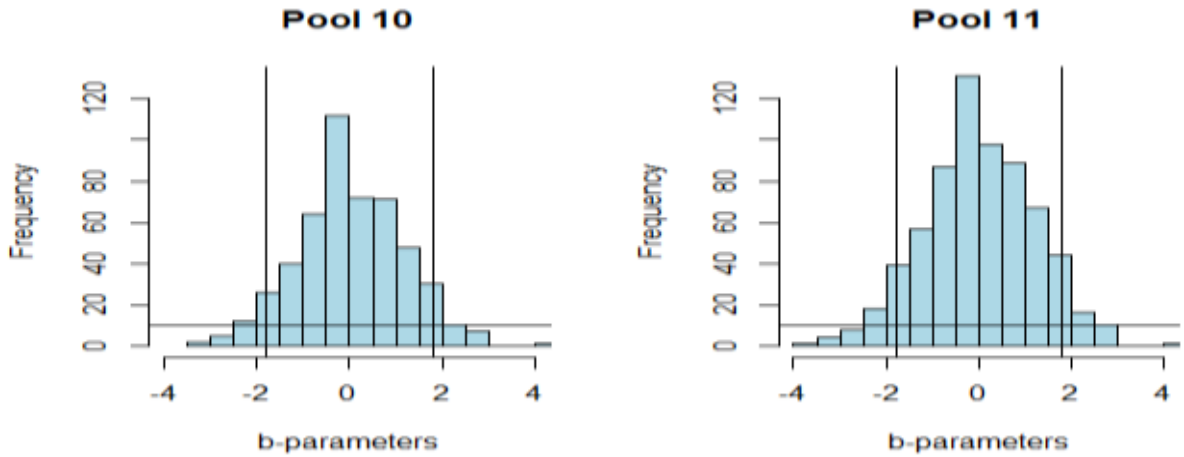


Notes: The small baseline item pool 1 included 150 on-grade items only, while item pool 7 incorporates 150 off-grade items from item pool 1.

Notes: The moderate baseline item pool 8 included 300 on-grade items only, while item pool 9

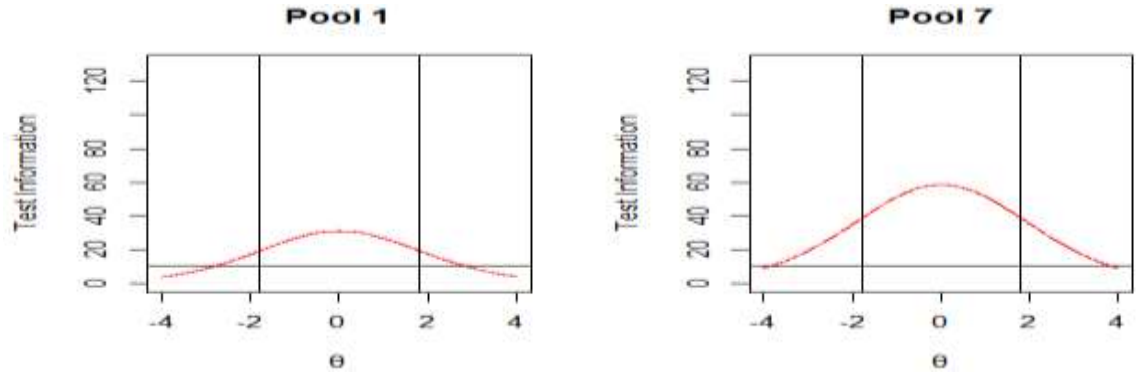


incorporates 150 off-grade items from item pool 8.

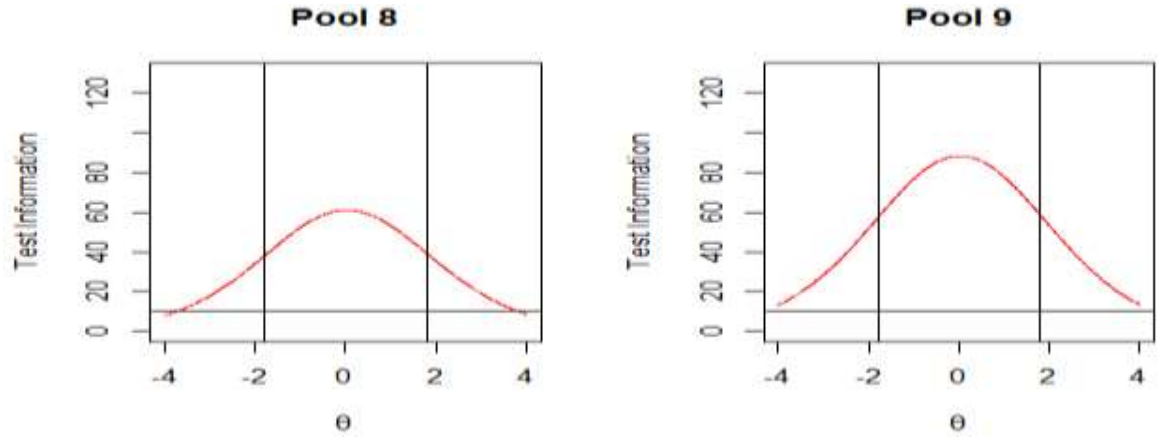


Notes: The large baseline item pool 10 included 500 on-grade items only, while item pool 11 incorporates 170 off-grade items from item pool 10.

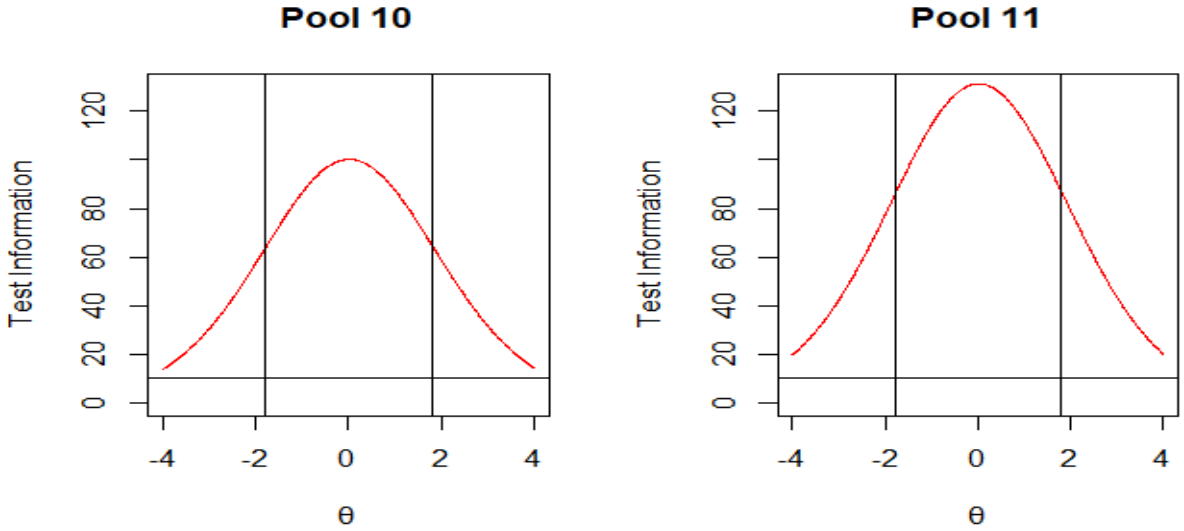
Figure 4. 2. Plots of test information across different baseline pools and item pools that incorporate off-grade items



Notes: The small baseline item pool 1 included 150 on-grade items only, while item pool 7 incorporates 150 off-grade items from item pool 1.

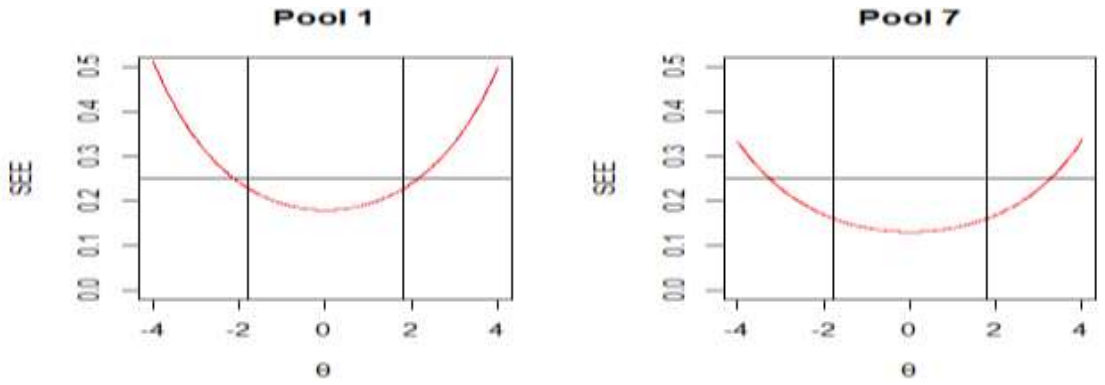


Notes: The moderate baseline item pool 8 included 300 on-grade items only, while item pool 9 incorporates 150 off-grade items from item pool 8.

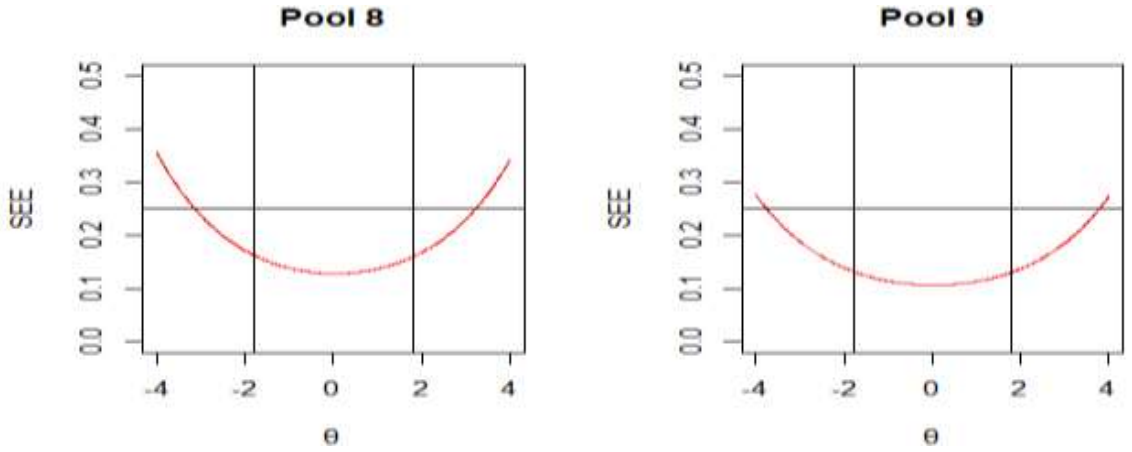


Notes: The large baseline item pool 10 included 500 on-grade items only, while item pool 11 incorporates 170 off-grade items from item pool 10.

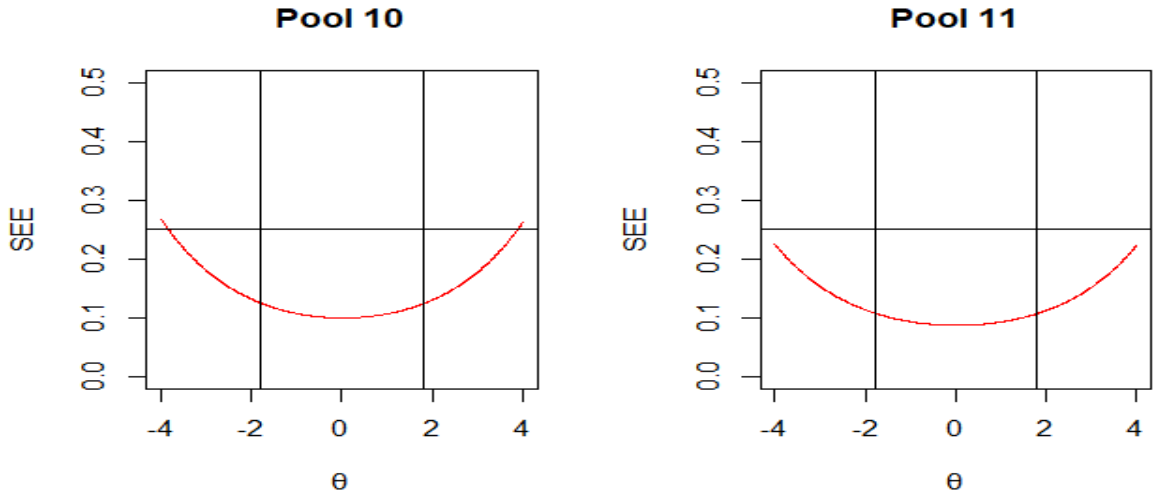
Figure 4. 3. Plots of the standard error of the estimate across different baseline pools and item pools that incorporate off-grade items



Notes: The small baseline item pool 1 included 150 on-grade items only, while item pool 7 incorporates 150 off-grade items from item pool 1.

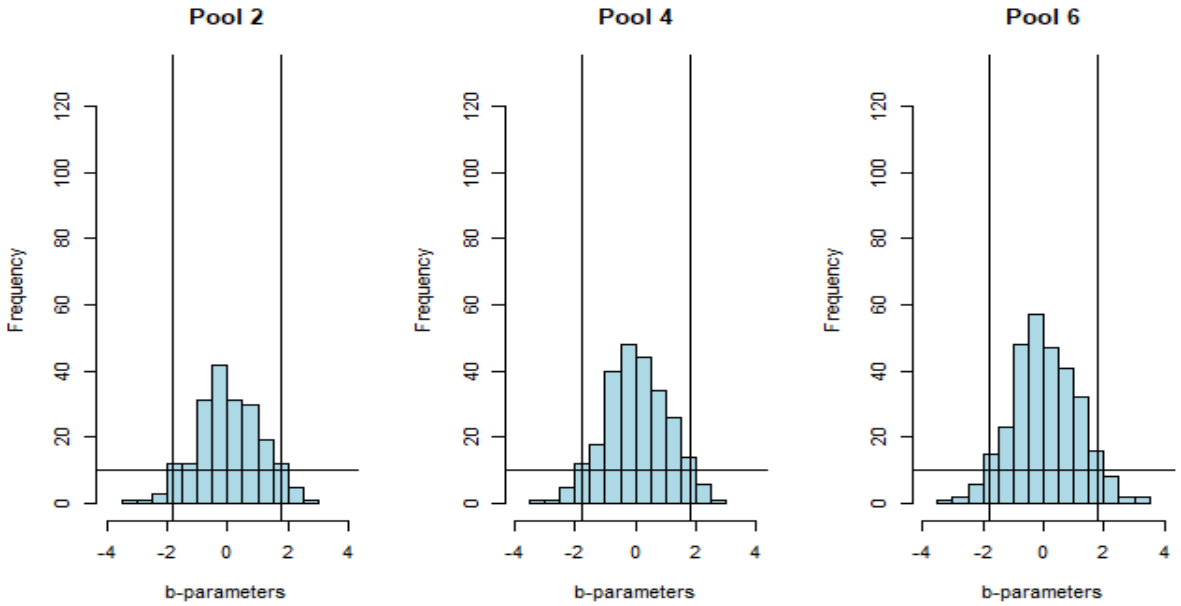


Notes: The moderate baseline item pool 8 included 300 on-grade items only, while item pool 9 incorporates 150 off-grade items from item pool 8.

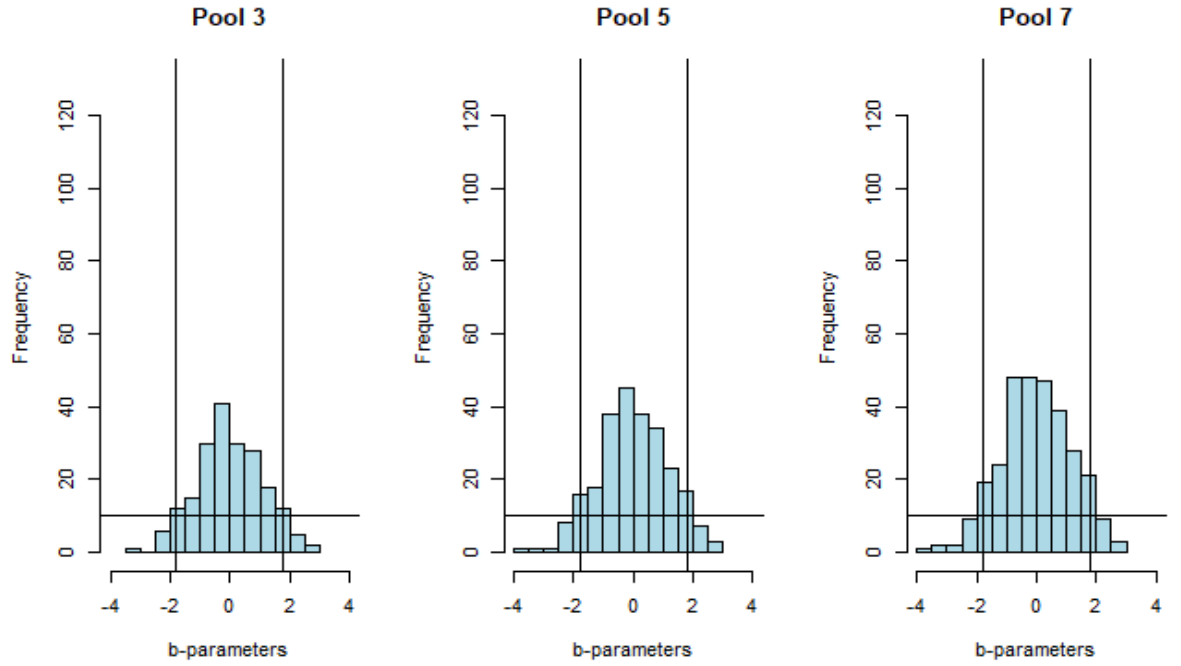


Notes: The large baseline item pool 10 included 500 on-grade items only, while item pool 11 incorporates 170 off-grade items from item pool 10.

Figure 4. 4. Histograms of the *b*-parameters across item pools that incorporate different proportions of off-grade items

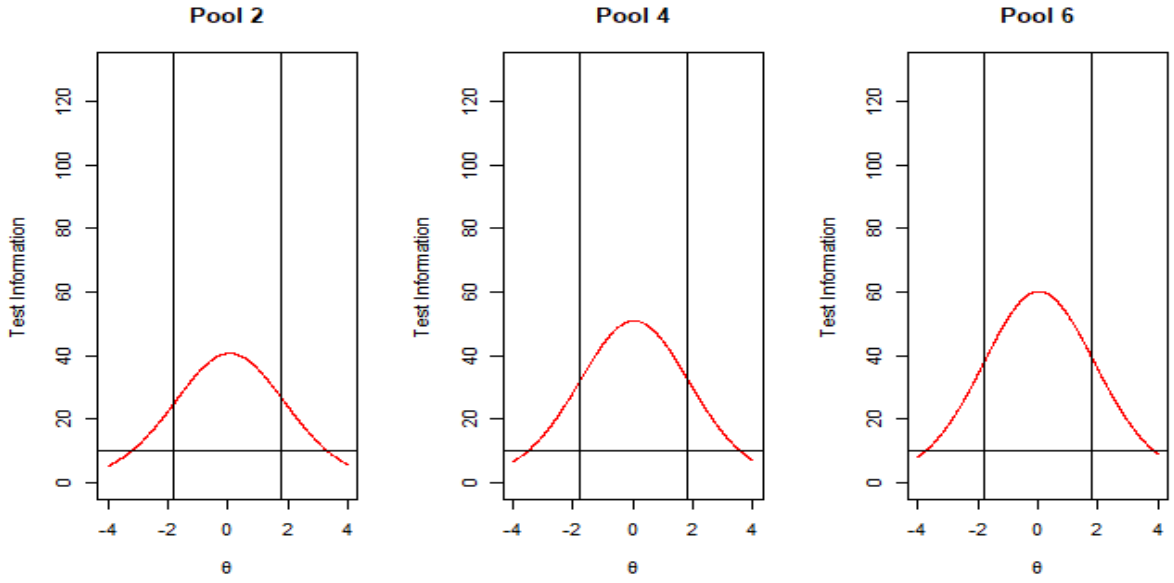


Notes: Item pools 2, 4 and 6 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

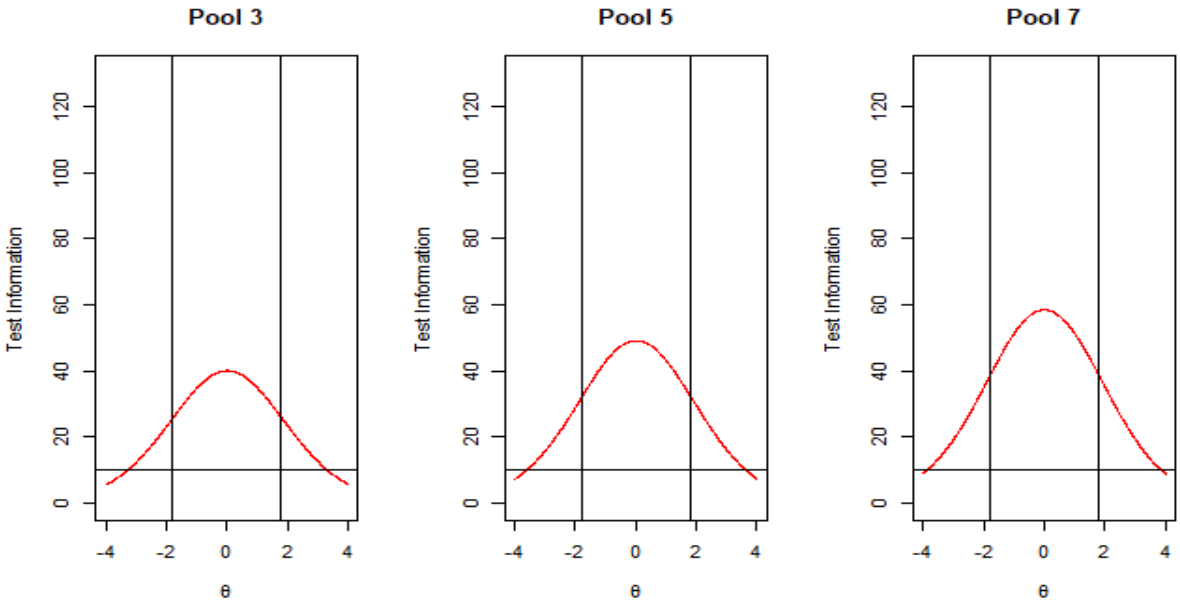


Notes: Item pools 3, 5 and 7 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

Figure 4. 5. Plots of test information across item pools that incorporate different proportions of off-grade items

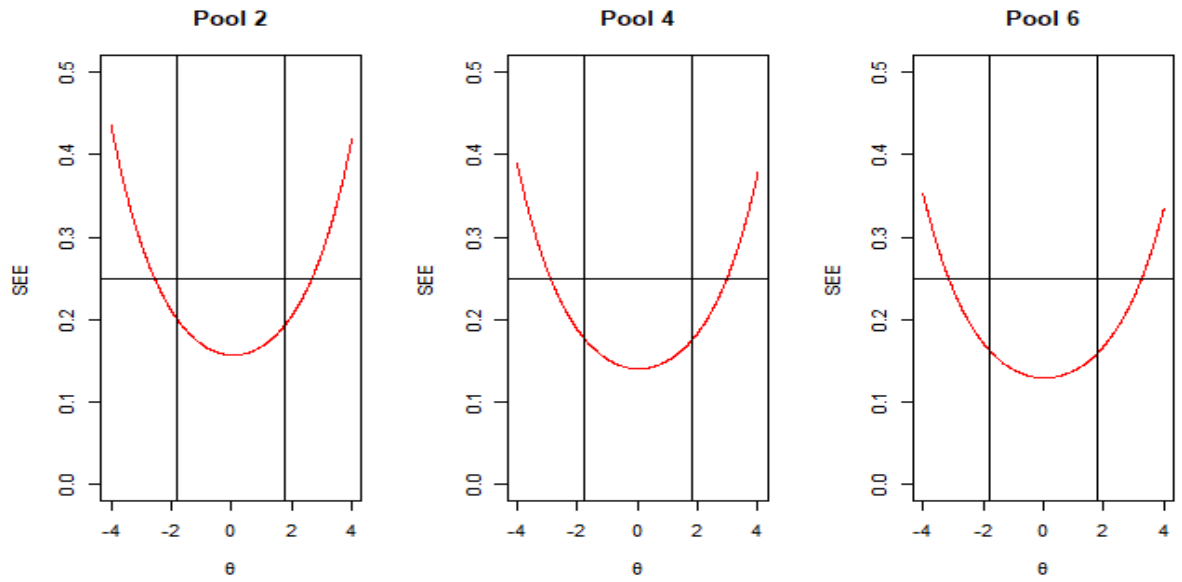


Notes: Item pools 2, 4 and 6 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

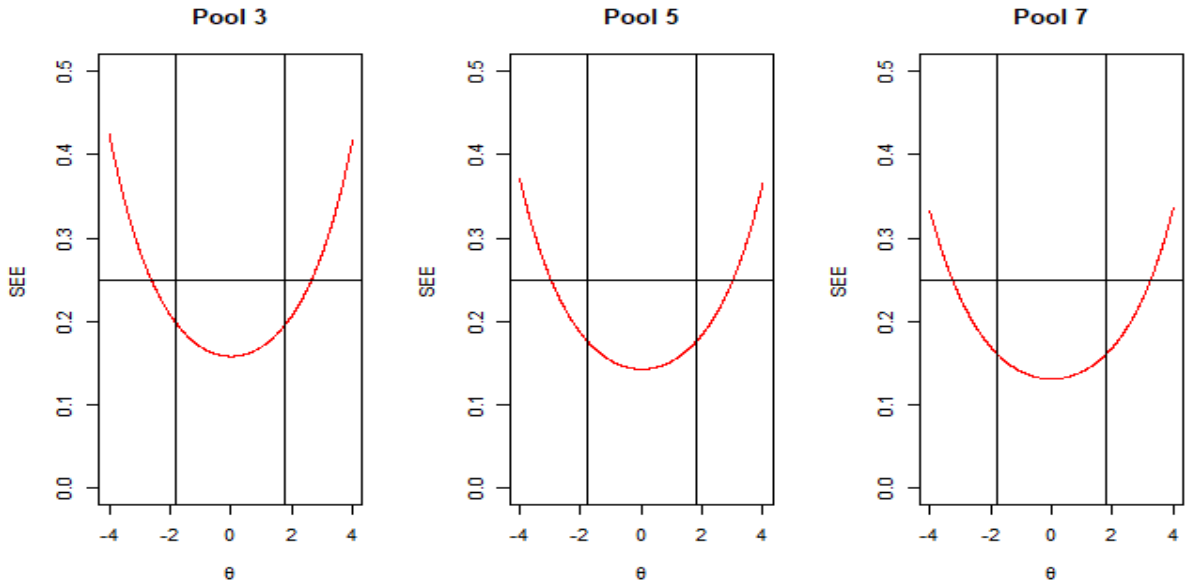


Notes: Item pools 3, 5 and 7 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

Figure 4. 6. Plots of the standard error of the estimate across item pools that incorporate different proportions of off-grade items

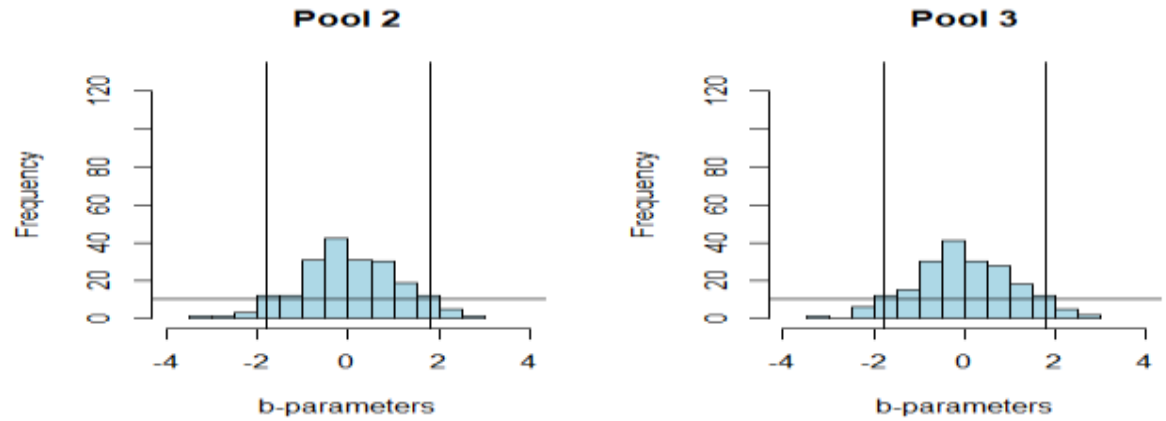


Notes: Item pools 2, 4 and 6 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

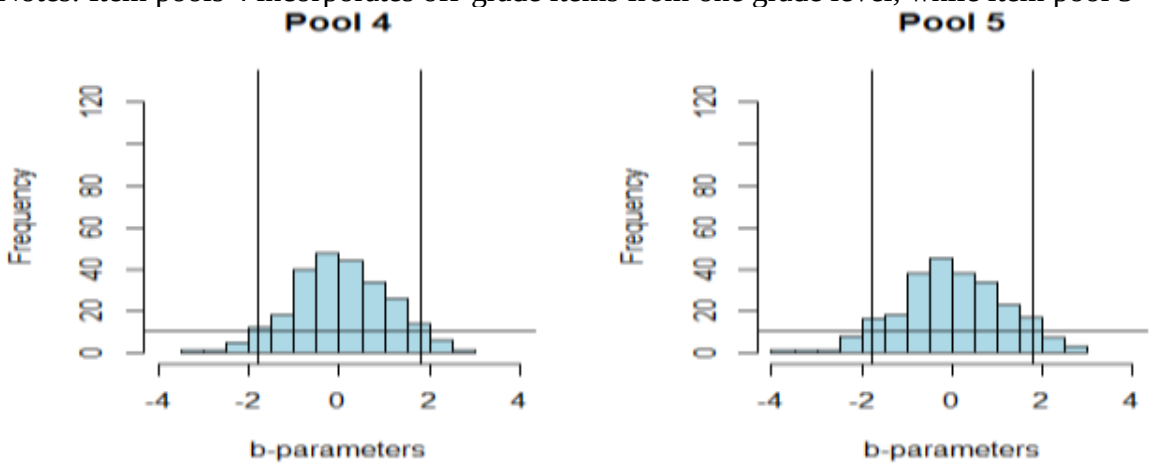


Notes: Item pools 3, 5 and 7 incorporate small, moderate and large proportions of off-grade items from the small baseline pool (the ratios of off-grade items to on-grade items are 1/3, 2/3 and 3/3 respectively).

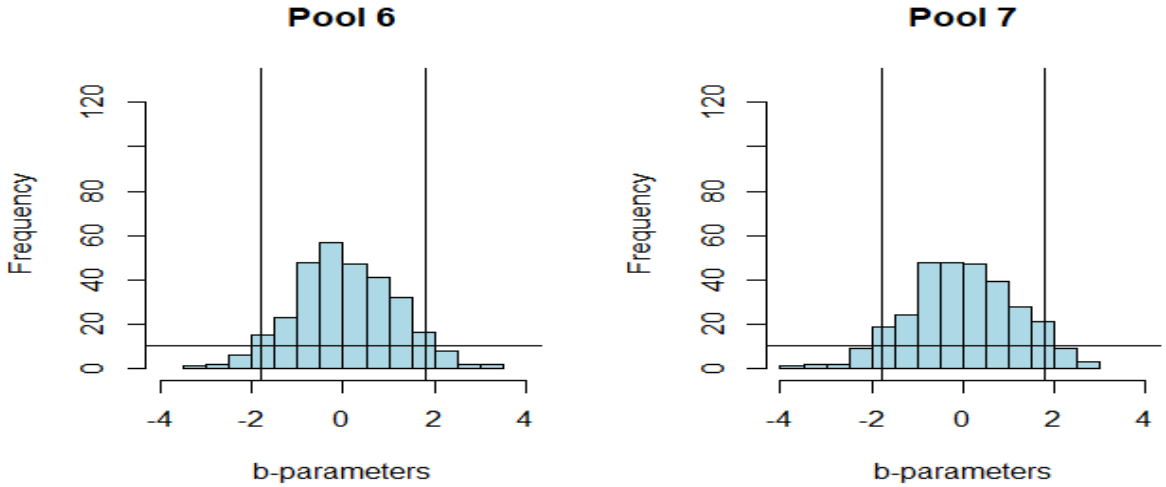
Figure 4. 7. Histograms of the *b*-parameters across item pools that incorporate different ranges of off-grade items



Notes: Item pools 2 incorporates off-grade items from one grade level, while item pool 3 incorporates off-grade items from two grade levels.

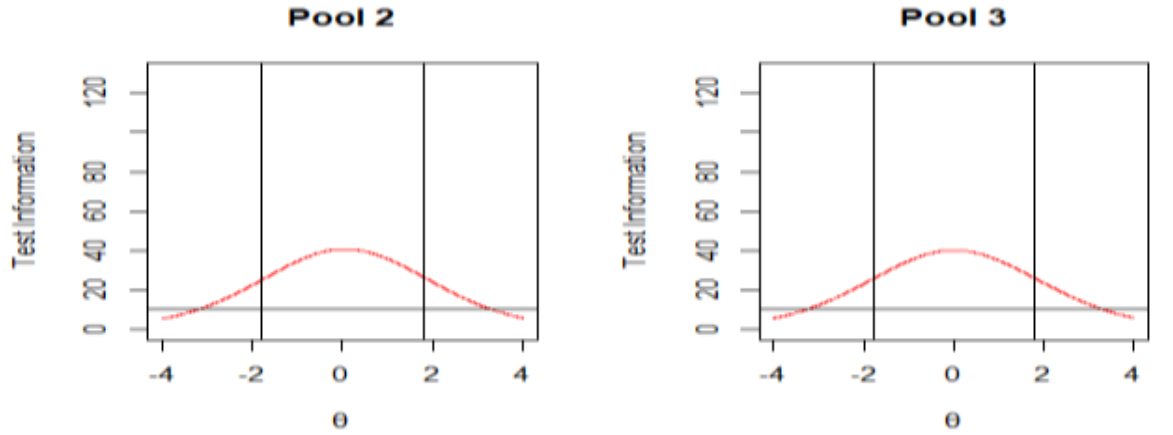


Notes: Item pools 4 incorporates off-grade items from one grade level, while item pool 5 incorporates off-grade items from two grade levels.



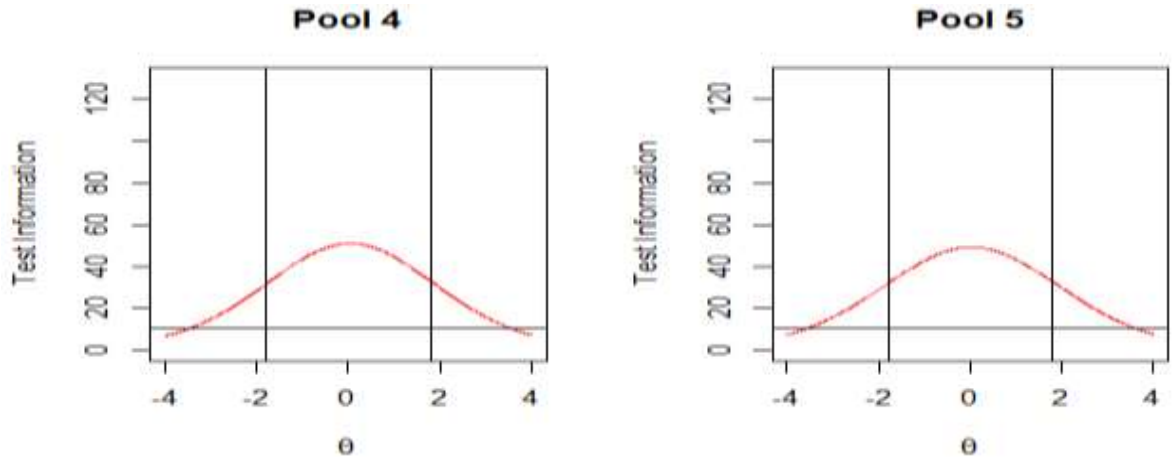
Notes: Item pools 6 incorporates off-grade items from one grade level, while item pool 7 incorporates off-grade items from two grade levels.

Figure 4. 8. Plots of test information across item pools that incorporate different ranges of off-grade items

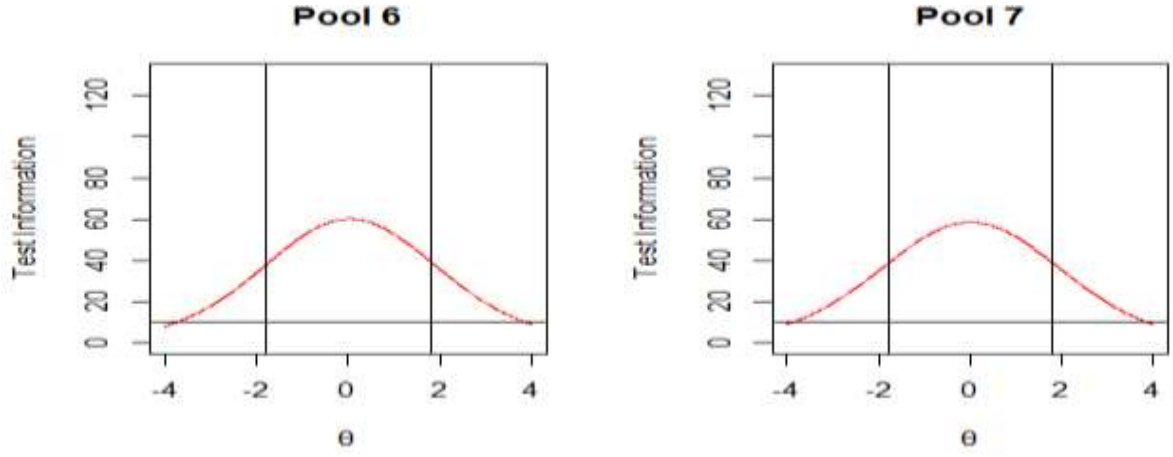


Notes: Item pools 2 incorporates off-grade items from one grade level, while item pool 3 incorporates off-grade items from two grade levels.

Notes: Item pools 4 incorporates off-grade items from one grade level, while item pool 5

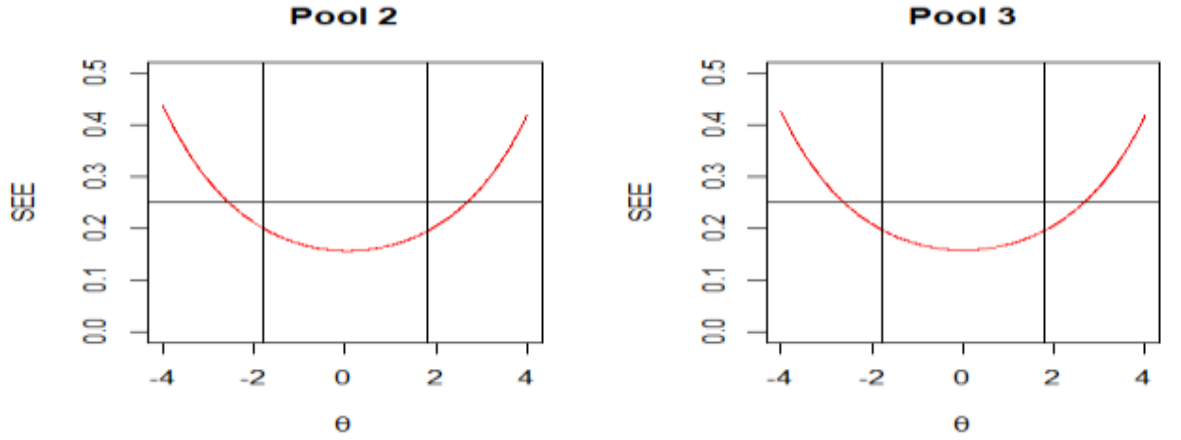


incorporates off-grade items from two grade levels.

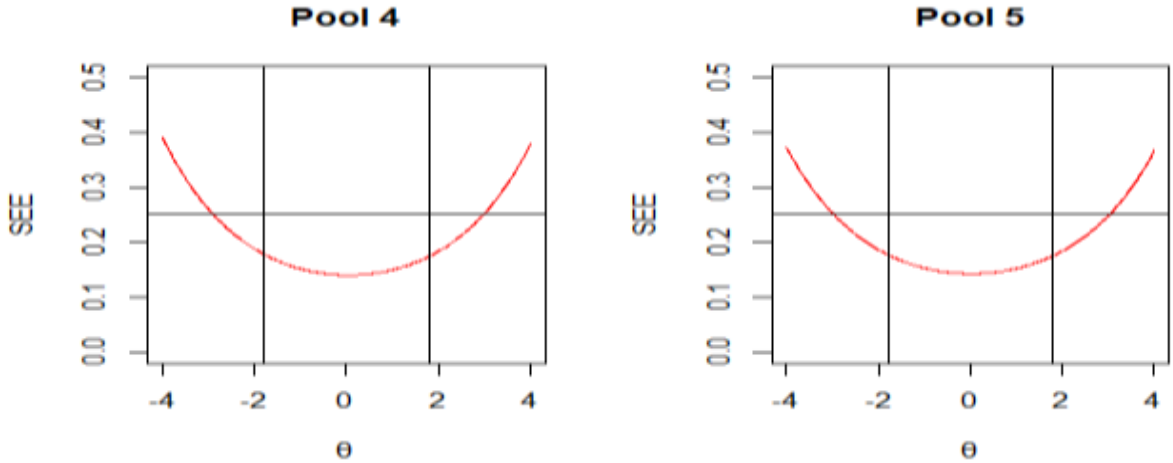


Notes: Item pools 6 incorporates off-grade items from one grade level, while item pool 7 incorporates off-grade items from two grade levels.

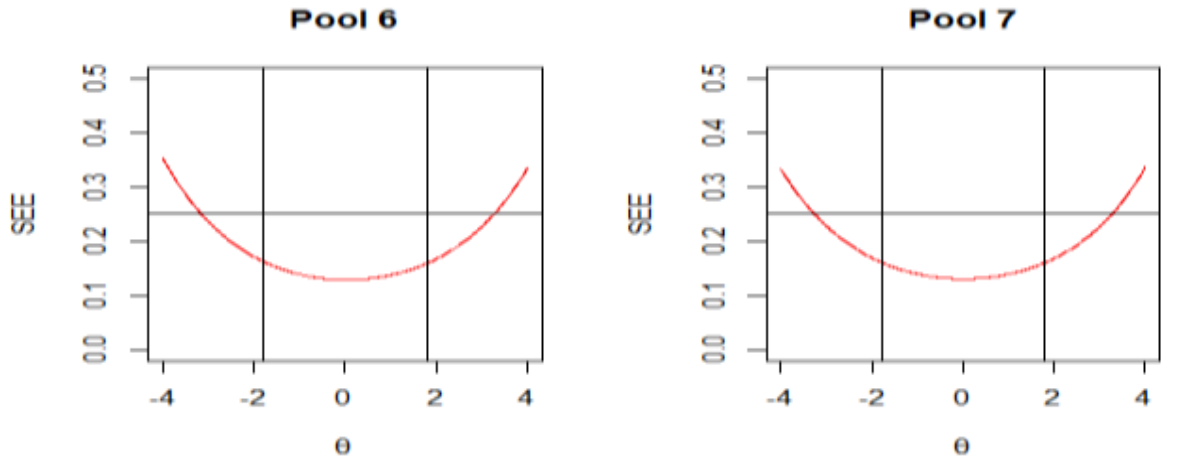
Figure 4. 9. Plots of the standard error of the estimate across item pools that incorporate different ranges of off-grade items



Notes: Item pools 2 incorporates off-grade items from one grade level, while item pool 3 incorporates off-grade items from two grade levels.



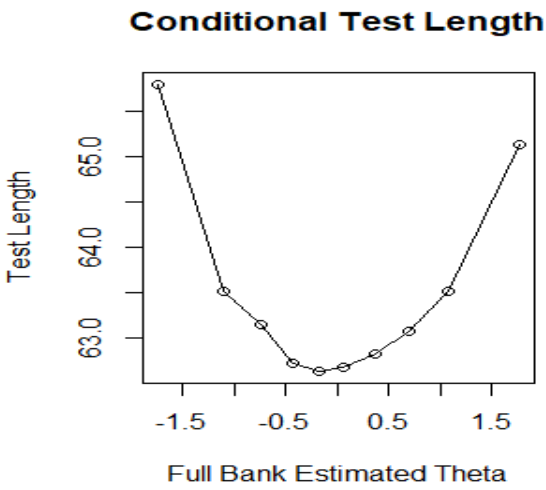
Notes: Item pools 4 incorporates off-grade items from one grade level, while item pool 5 incorporates off-grade items from two grade levels.



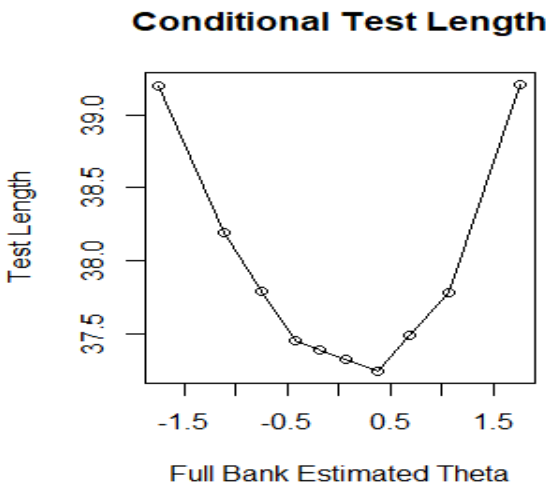
Notes: Item pools 6 incorporates off-grade items from one grade level, while item pool 7 incorporates off-grade items from two grade levels.

Figure 4. 10. The conditional test length and the conditional standard error of ability estimation of the deciles of the true ability levels with SE threshold of 0.25 and 0.32.

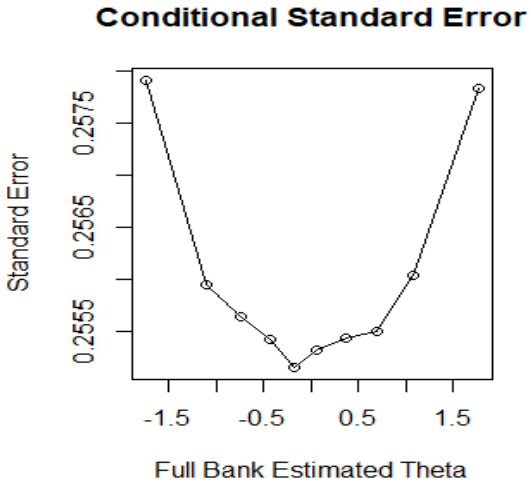
SE threshold of 0.25



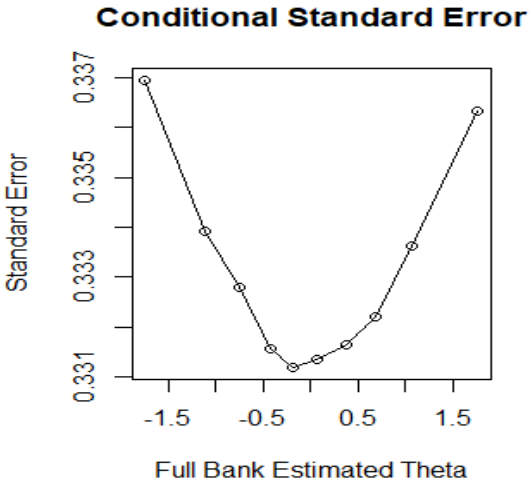
SE threshold of 0.32



SE threshold of 0.25



SE threshold of 0.32



Chapter 5 CONCLUSION AND DISCUSSION

This study investigated the impact of practical constraints of incorporating off-grade items in item pool characteristics and item pool performance under variable- and fixed-length constrained CAT. First, the research questions are addressed based on the results of the study. Then practical implications from the findings are presented. Finally, limitations and directions for future studies are discussed. The discussion focus on item pool performance with a SE threshold of 0.32 except for the discussion of the study factor of stopping rules.

5.1 Research Questions

How do different numbers of on-grade items in the item pools affect the impact of incorporating off-grade items in terms of item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?

Given the similar number and same range of off-grade items, the impact of incorporating off-grade items is highly dependent on the number of on-grade items in the item pool. The current study used three on-grade pool sizes (150, 300, and 500 items). This study employs a 500-item pool to target summative assessment or high-volume exams while employing 150- and 300-item pools to reflect formative or remedial assessment and advanced exams with a low-volume of examinees. In addition, the proportions of off-grade items for several item pools (7, 9 and 11) reflect reality in practice, where out of practical or content necessity a testing organization with a small item pool may incorporate a large proportion of off-grade items, and a testing organization with a large item pool may incorporate a small proportion of off-grade items.

The results suggest that the different numbers of on-grade items in the item pool have a positive impact on incorporating off-grade items. Incorporating off-grade items within a small

baseline pool (150 on-grade items) has more of an impact than within a large baseline pool (500 on-grade items). More specifically, compared with their baseline pools, item pool 7 (small baseline pool) leads to several more improvements than item pools 9 and 11 (moderate and large baseline pools). These improvements include, 1) consistent improvements in item pool characteristics (the smaller average value of the SEEs and the wider range of the *b*-parameters) at the overall level, which indicates increased measurement precision and more appropriate items for high- and low-performing simulees; 2) consistent improvements in measurement precision (the smaller RMSE and the smaller mean standard error) for high- and low-performing simulees; 3) consistent improvements in test efficiency (the shorter mean test length, the smaller item overlap rate, and the smaller maximum exposure rate) at the overall level, as well as improvements in test efficiency (the shorter mean test length) for high- and low-performing simulees; and 4) consistent improvements in content blueprint fulfillment (the higher upper-bound content blueprint fulfillment) for high- and low-performing simulees, as well as slight improvements in content blueprint fulfillment at the overall level. The improvements in measurement precision, test efficiency, and content blueprint fulfillment for item pools 9 and 11 are very small for low- and high-performing simulees, and at the overall level.

Previous to this, only Wei and Lin's (2015) study investigated incorporating off-grade items under variable- and fixed-length CAT. Their study supported the conclusion that incorporating off-grade items can improve measurement precision and test efficiency for high- and low-performing examinees and at the overall level. The current study confirms Wei and Lin's finding that incorporating off-grade items yields improvements in item pool performance for high- and low-performing examinees than on-grade item only item pool, but the improvements in item pool performance at the overall level are not obvious. This maybe due to

the fact that the proportions of off-grade items in the item pool are more appropriately reflect reality in this study (in Wei and Lin's study, the ratio of the off-grade item to on-grade items is 2).

In summary, if the goal is to identify high- and low-performing examinees in order to provide information to support precisions for advanced or remedial coursework, a test organization that has a small item pool and could tolerate the content of off-grade items could consider incorporating additional items and increase measurement precision for low- and high-performing students. Otherwise, off-grade items should be used with caution. However, for moderate or larger baseline pools, the improvements in item pool performance are very small and the increase in the proportions of off-grade items may not be justified.

How does incorporating different proportions of off-grade items affect the item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?

The impact of incorporating off-grade items is dependent on the number of off-grade items. Incorporating off-grade items is a good way to increase item pool size without new item development. However, too many off-grade items will lead to a challenge for tests that are designed to measure grade-level standards. This is especially problematic for high-performing students taking below-grade items and low-performing students taking above-grade items. There are fixed guidelines on the acceptable proportion of off-grade items in the item pool. The current study employed three proportions of off-grade items in the item pool: small, moderate, and large (the ratios of the off-grade item to on-grade items in the item pool are 1/3, 2/3 and 3/3 accordingly). The choice of these three proportions was made based on the results of the preliminary simulation.

The results suggest that increasing different proportions of off-grade items has an impact in item pool characteristics and item pool performance. Increasing proportions of off-grade items leads to several improvements: 1) consistent improvements in item pool characteristics (increased the maximum value of the TIFs, the smaller average value of the SEEs, and the wider range of the *b*-parameters) at the overall level, which indicate more information and increased measurement precision at the overall level and additional appropriate items for high- and low-performing simulees; 2) consistent improvements in measurement precision (the smaller mean standard error) for high- and low-performing simulees; 3) consistent improvements in item overlap and maximum exposure rate (the smaller item overlap and the smaller maximum exposure rate) at the overall level, as well as slight improvements in average test length (the shorter mean test length) for high- and low-performing simulees; and 4) consistent improvements in content blueprint fulfillment (the higher upper-bound content blueprint fulfillment) for high- and low-performing simulees. However, there remains the issue that increasing proportions of off-grade items will increase the proportion of inappropriate items.

In summary, compared with the impact of the different numbers of on-grade items in the item pool, different proportions of off-grade items only yield a positive impact in item pool characteristics and item pool performance. Even the smallest proportions (1/3) of off-grade items maybe questionable if test users have strict consideration for measuring on-grade standards.

How does incorporating different ranges of off-grade items affect the item pool characteristics and item pool performance under variable- and fixed-length constrained CAT?

Off-grade items can be incorporated in a variety of ways. The range of off-grade items is one important factor. One reason to expand the range of off-grade items from one grade level to two grade levels maybe that administering off-grade items will prevent a low-performing

examinee from being overwhelmed by overly difficult questions and a high-performing examinee from being unmotivated by overly easy questions (Bong, 2016). However, broadening the range of off-grade items to a two grades range maybe a more serious challenge for tests that are designed to measure grade-level standards. There are frequently no guidelines on the appropriate range of off-grade items in the item pool. The current study provides two conditions for the range of incorporating off-grade items: one grade above and one grade below the target grade, and two grade levels above and two grade levels below the target grade. The choice of these two ranges was made based on previous studies.

The results suggest that different ranges of off-grade items have an impact in item pool characteristics and item pool performance. Broadening the range of off-grade items leads to several improvements: 1) consistent improvements in item pool characteristics (the wider range of *b*-parameters) at the overall level, which indicates more appropriate items for high- and low-performing simulees; 2) consistent improvements in measurement precision (the smaller mean standard errors) for high- and low-performing simulees; and 3) consistent improvements in content blueprint fulfillment (the higher upper-bound content blueprint fulfillment) for high- and low-performing simulees. However, there is an issue that broadening the range of off-grade items will raise the proportion of off-grade items from two grade level for each test.

In summary, compared with the impact of different proportions of on-grade items in the item pool, different ranges of off-grade items also yield improvements in measurement precision and content blueprint fulfillment, but not on test efficiency. However, the improvements found with broadening the range of off-grade items (i.e. measurement precision and content blueprint fulfillment) are larger than the improvements of increasing the proportion of off-grade items.

Broadening the range of off-grade items maybe a better option than increasing the proportion of off-grade items for low- and high-performing simulees in some situations.

How does incorporating off-grade items affect item pool performance differently under variable- and fixed-length CAT?

Stopping rules and appropriate SE threshold levels are other important factors that need to be considered regarding incorporating off-grade items. Fixed-length CAT provides better content blueprint fulfillment for each examinee. Variable-length CAT provides better measurement precision and shorter test length. Wei and Lin (2015) found that under variable-length CAT, a smaller SE threshold level improved measurement precision (the smaller RMSE and the larger correlations between true and estimated ability) but decreased content blueprint fulfillment. This study implemented CAT under variable- and fixed-length CAT, and two SE thresholds levels, 0.32 and 0.25, under variable-length CAT.

The results suggest that different stopping rules have an impact on incorporating off-grade items. Compared with variable-length CAT, fixed-length CAT produces a shorter mean test length for low- and high-performing simulees and produces consistent improvements in content blueprint fulfillment (100% upper-bound content blueprint fulfillment) at the overall level and for low- and high-performing simulees. However, the variable CAT shows consistent improvements in measurement precision (smaller RMSE and smaller mean standard errors) for low- and high-performing simulees.

The results suggest that different SE threshold levels have an impact on incorporating off-grade items. Specifically, changing SE threshold level from 0.32 to 0.25 leads to improvements in measurement precision (the smaller RMSE and the higher correlations between estimated and true ability at the overall level. The smaller Bias, the smaller RMSE, and the

smaller mean standard error for low- and high-performing simulees). However, it also yields some undesirable item pool performance, including 1) a decrease in test efficiency (increased the mean test length from 38 to 63 items, higher item exposure rate and maximum exposure rate at the overall level, and longer test length for low- and high-performing simulees); 2) a decrease in content blueprint fulfillment (0% upper-bound content blueprint fulfillment at the overall level and for low- and high-performing simulees); and 3) an increase the mean proportion of off-grade items for each test for low- and high-performing simulees.

Wei and Lin (2015) found that a smaller SE threshold level would yield improvements in measurement precision. The results of this study are consistent with these findings. However, considering the increase in test length (about 65%), content blueprint fulfillment (0% of upper-bound content blueprint fulfillment) and the mean proportion of off-grade items, it is difficult to justify a SE threshold of 0.25 is appropriate even for the largest item pool (670 items).

In summary, different SE threshold levels have more impact than different stopping rules. The choice of SE threshold levels will affect whether CAT will yield appropriate item pool performance at the overall level and for low- and high-performing simulees. However, the choice of fixed- or variable-length CAT will only affect certain test efficiency and content blueprint fulfillment for low- and high-performing simulees.

Are some indices of evaluation criteria more sensitive than other indices for evaluating the impact of incorporating off-grade items?

The current study applied several evaluation criteria to examine item pool characteristics and performance. It is worthwhile examining which index is more sensitive and consistent to evaluate the impact of incorporating off-grade items.

First, the impact of incorporating off-grade items are obvious for low- and high-performing simulees than at the overall level. For item pool characteristics, the descriptive statistics and plots of the TIFs and SEEs are more sensitive and consistent than the descriptive statistics and histograms of the *b*-parameters. For measurement precision, the index of the mean standard error is more sensitive and consistent than other indices (bias, RMSE, and the correlations between estimated and true ability). For example, the differences in bias and the correlations between estimated and true ability are very small. For test efficiency, the index of item overlap rate is more sensitive and consistent than other indices (mean test length and maximum item exposure). For content blueprint fulfillment, upper-bound content blueprint fulfillment is more sensitive than lower-bound content blueprint fulfillment. The index of mean proportion of off-grade items for each test is more sensitive than the other two indices (mean proportion of off-grade items from two grade levels only and a number of tests that include off-grade items).

5.2 Conclusions and Practical Applications

Incorporating off-grade items within an on-grade item pool is an ongoing issue for K-12 assessment. For example, SBAC incorporated off-grade items within an on-grade item pool (AIR, 2015). Several studies (Wei & Lin, 2015; AIR, 2015; Way et al., 2010) have concluded that incorporating off-grade items improved measurement precision, test efficiency, and content blueprint fulfillment for high- and low-performing examinees. However, increasing off-grade items may raise concerns for tests that are designed to measure only grade-level standards. This study shows that there are many practical constraints of incorporating off-grade items that need to be considered. In general, the optimal design of incorporating off-grade items is characterized by two criteria: 1) it leads to improvements in item pool characteristics and item pool

performance for low- and high-performing simulees; and 2) it leads to a minimum proportion of off-grade items for each test.

Based on the criteria above, item pool 7 (incorporate 150 off-grade items into a small baseline pool) maybe a good design to incorporate off-grade items, with improvements in item pool characteristics and item pool performance for high- and low-performing simulees. However, as shown in Tables 4.3 and 4.4, the mean proportions of off-grade items are high (46% and 50%) for low- and high-performing simulees in item pool 7. The results of this study indicate that there are some situations in which incorporating off-grade items would be beneficial. For example, some formative assessments could tolerance moderate to a large proportion of off-grade items on each test. The design present in this study can be applicable to any test organization that aims to improve item pool performance for high- and low-performing examinees. Future studies could limit off-grade items only to high- and low-performing simulees.

The contribution of this study has four points. First, previous studies (Wei & Lin 2015, Way et al., 2010) focused on incorporating off-grade items or not, while the current study investigates the practical constraints of incorporating off-grade items. The results also show that practical constraints of incorporating off-grade items, organized here from most impact to least impact in item pool characteristics and item pool performance, are: 1) incorporating off-grade items into small baseline pool or large baseline pool; 2) broadening the range of off-grade items from one grade level to two grade levels; 3) increasing the proportion of off-grade items in the item pool; and 4) applying variable- or fixed-length CAT. For example, the results indicated that broadening the range of off-grade items yields improvements in measurement precision and content blueprint fulfillment when compared to increasing the proportion of off-grade items.

Second, this study shows that the optimal design of incorporating off-grade items should take into account the complete picture of the effects of incorporating off-grade items. For example, Wei and Lin's (2015) finding that incorporating off-grade items yields improvements for bothr high- and low-performing examinees and at the overall level. However, this study found the improvement of incorporating off-grade items at the overall level is slight. Thus, if the purpose of a test is to not to measure the performance of high- and low-performing examinees, for example, when a test is more concerned about test takers in the middle of the distribution. Then incorporating off-grade items may not be necessary.

Third, this study could serve as guidance for test organizations that are considering incorporating off-grade items within an on-grade assessment. For example, this study shows that the impact of incorporating off-grade items are more evident for low- and high-performing simulees than at the overall level. This study also shows that some indices of item pool characteristics and item pool performance are more sensitive and consistent for evaluating the impact of incorporating off-grade items. This study indicates that a SE threshold of 0.25 under variable-length CAT would be a difficult target, given the pool characteristics, resulting in longer tests.

Fourth, the current study indicates that incorporating off-grade items may not improve measurement precision (bias, RMSE, and correlations), test length, or test security (item overlap rate and maximum exposure rate) for overall simulees. However, incorporating off-grade items will reduce the chances that a low-performing examinee will be overwhelmed by overly difficult questions and a high-performing examinee will be bored by overly easy questions. The goal of the CAT is to provide a unique test that fits a student's ability level. The current study provides a foundation for future studies to study the inclusion of items that are closer to high- and low-

performing student's ability level. This method will improve the efficiency of the CAT without the cost of developing new items.

5.3 Limitations and Directions for Future Research

The current study has several limitations. First, SBAC incorporated off-grade items within an on-grade item pool only after an examinee has responded to two-thirds of the operational items in order to help ensure off-grade items are administered only to low- and high-performing examinees (AIR, 2016). This would greatly decrease the mean proportion of off-grade items for each test and the number of tests with off-grade items. The current study incorporates off-grade items within on-grade item pools before an examinee responds to any item (catR does not have a CAT algorithm similar to SB's).

Second, the current study limited content blueprint fulfillment to just satisfying content specifications (i. e., the lower and upper bounds for each content area). However, the depth of knowledge or cognitive level constraints should also be considered. The current study employs the content balancing method CCAT, which was implemented successfully with mutually exclusive content constraints. However, to satisfy multiple constraints, other content balancing methods need to be employed.

The current study provides direction that future studies could address. First, the current study focuses on investigating how incorporating off-grade items affects overall simulees as well as high- and low-performing simulees. The results of the current simulation study indicate that incorporating off-grade item only improves item pool performance for high- and low-performing simulees. It is worthwhile to investigate the impact of incorporating off-grade items for only high- and low-performing simulees in future studies.

Second, the results of the current simulation study indicated that broadening the range of off-grade items yields more improvements in measurement precision and content blueprint fulfillment than increasing the proportion of off-grade items. However, educational policy may prevent using items from two grade levels. Therefore, it may be worthwhile limiting off-grade items to one grade level only. The current simulation study chose off-grade items randomly from the whole normal distribution of grade 3 N (-0.7, 1.05) and grade 5 N(0.6, 1.15) (if the range of off-grade items is one grade level). An alternative way is choosing off-grade items where the item difficulty parameters are close to low- and high-performing simulee's ability scale. Future studies could consider offering off-grade items only for high- and low-performing simulees' ability scale to improve the efficiency of the CAT.

Third, this study applied the 1PL IRT model in order to prevent overuse of highly discriminating items and equalized item exposure. Future studies should investigate the impact of incorporating off-grade items under the 2PL or 3PL IRT model, which allows investigation of the effects of variation of discrimination parameters and guessing parameters in CAT simulation.

Fourth, this study generated each item pool and simulees with a normal distribution consistent with previous studies (Gonulates, 2015; Tay, 2015; Lim, 2010; Xiong, 2010; Xing & Hambleton, 2004). However, this generation might not have reflected reality for item pool and examinees' ability distribution. Future studies should investigate the impact of a uniform distribution or other distributions for item pools and simulees. Gorin, Dodd, Fitzpatrick, and Shieh (2005) found that a mismatch between ability distributions and item pool characteristics will result in a larger bias for ability estimation.

Fifth, this study only incorporates three item pool sizes, three proportions, and two ranges of off-grade items. Future studies should consider other combinations.

Lastly, this study is limited to the R package catR. The catR package is useful and was recently updated by the authors (Magis & Barrada, 2017). However, only the constrained content balancing method (Kingsbury & Zara, 1989) and randomesque exposure control method are available in the catR for now. Future research should investigate the impact of incorporating off-grade items with varying content balancing methods and exposure control methods. Future studies should try to limit the administration of off-grade items only to low- and high-performing simulees by improving the CAT simulation tools.

REFERENCES

American Institutes for Research (AIR). (2015). *Smarter Balanced Summative Assessments Testing Procedures for Adaptive Item-Selection Algorithm 2014–2015 Test Administrations*. Retrieved from https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

American Institutes for Research. (2016). *Smarter Balanced Summative Assessments Simulation Results in 2016–2017 Test Administrations*. Retrieved from https://portal.smarterbalanced.org/library/en/2016-17-summative-assessments-simulation-results.pdf

Babcock, B., & Weiss, D. J. (2013). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*, 1–18.

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*(2), 137-149.

Bong, D. (2016). *Fixed Form Vs. Adaptive Test Design in Language Proficiency Testing*. Retrieved from avant assessment: https://avantassessment.com/blog/2016/fixed-form-vs-adaptive-test-design-in-language-proficiency-testing

Chajewski, M. (2011). *MLE vs. Bayesian item exposure in non-cognitive type adaptive assessments with restricted item pools: Trait estimation, item selection and reliability* (Ph.D.). Fordham University, the Bronx

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of educational Measurement, 40*(1), 71-103.

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 241-255.

Cohen, J., & Albright, L. (2014). *Smarter Balanced adaptive item selection algorithm design report*, Washington, D.C, Retrieved from http://www.smarterapp.org/specs/AdaptiveAlgorithm.html

National Center for Research on Evaluation, Standards & Student Testing (CRESST) Psychometrics Team. (2016). *Simulation-Based Evaluation of the 2014-2015 Smarter Balanced Summative Assessments: Accommodated Item Pools.* Retrieved from http://www.smarterbalanced.org/assessments/development/additional-technical-documentation/

CRESST Psychometrics Team. (2017). *Simulation-Based Evaluation of the 2016-2017 Smarter Balanced Summative Assessments: General and Accommodated Item Pools.* Retrieved from http://www.smarterbalanced.org/assessments/development/additional-technical-documentation/

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. *Journal of Applied Testing Technology, 13*(1).

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems. *Council of Chief State School Officers*.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and psychological measurement, 53*(1), 61-77.

Eggen, T., & Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*(5), 713–734.

Evans, J. J. (2010). Comparability of examinee proficiency scores on computer adaptive tests using real and simulated data (Doctoral dissertation, Rutgers University-Graduate School-New Brunswick).

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247-261.

Gonulates, E. (2015). *A novel approach to evaluate item pools: The Item Pool Utilization Index* (Ph.D.). Michigan State University, Ann Arbor.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433-456.

Grossman, J. (2010). *Evaluating the Efficiency and Accuracy of a Computer Adaptive Curriculum-Based Measurement* (Ph.D.). Columbia University, New York.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, *74*(4), 677-696.

He, W., & Reckase, M. D. (2014). Item Pool Design for an Operational Variable-Length Computerized Adaptive Test. *Educational & Psychological Measurement*, *74*(3), 473-494.

Hembry, I. F. (2014). *Operational Characteristics of Mixed-format Multistage Tests Using the 3PL Testlet Response Theory Model* (Ph.D.). The University of Texas at Austin.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: administration mode effects and a comparison with short forms. *Applied Psychological Measurement, 31*(5), 412–429.

Jodoin, M. G. (2003). Psychometric properties of several computer-based test designs with ideal and constrained item pools. Doctoral dissertation, University of Massachusetts Amherst.

Kang, H.-A., & Chang, H.-H. (2016). Parameter Drift Detection in Multidimensional Computerized Adaptive Testing Based on Informational Distance/Divergence Measures. *Applied Psychological Measurement*, *40*(7), 534–550.

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer adaptive testing (CAT): A faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology*, *26*(2), 227.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Ph.D.). The University of Texas at Austin, Retrieved from ProQuest Dissertations & Theses Global. (230674954)

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8-14.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159-175.

Lei, P. W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied psychological measurement*, *36*(1), 21-39.

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*(5), 857-874.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment*, *2*(5).

Lim, E. Y. (2010). *The effectiveness of using multiple item pools to increase test security in computerized adaptive testing* (Ph.D.). The University of Illinois at Urbana-Champaign.

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education, 3*(1), 29-41.

Luecht, R. (2014). Design and Implementation of a Large-scale Multistage Testing System. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized Multistage Testing: Theory and Applications* (pp. 69-83). Chapman and Hall/CRC.

Luecht, R. (2015). Applications of Item Response Theory: Item and Test Information Functions for Designing and Building Mastery Tests. In S. Lane, M.R. Raymond, & T.M. Haladyna, (Eds.). *Handbook of test development*. Routledge.

Luo, X. (2015). *Incorporating mixed item formats in CAT: A comparison of shadow test and bin-structured approaches* (Ph.D.). Michigan State University, Ann Arbor. Retrieved from ProQuest Dissertations & Theses Global. (1750071954)

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315.

Magis, D., Raiche, G., Barrada, J.R., (2017) Generation of IRT Response Patterns under Computerized Adaptive.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(1), 1-19.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). Past and Present Understandings of Out-of-Level Testing: A Research Synthesis. Out-of-Level Testing Report 1.

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing Flexible Constraints and Measurement Precision in Computerized Adaptive Testing. *Educational & Psychological Measurement, 72*(4), 629–648.

Northwest Evaluation Association. (2013). Measures of academic progress: A comprehensive guide to the MAP K-12 computer adaptive interim assessment.

Piromsombat, C. (2014). *Differential Item Functioning in Computerized Adaptive Testing: Can CAT Self-Adjust Enough?* (Doctoral dissertation, University of Minnesota).

Reckase, M. D. (2003). Item pool design for computerized adaptive tests. In the *annual meeting of the National Council on Measurement in Education, Chicago, IL*.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, 52(2), 127-141.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). Incorporating Content Constraints into a Multi-Stage Adaptive Testlet Design. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Risk, N. M. (2015). *The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT)* (Ph.D.). The University of Illinois at Chicago, Retrieved from ProQuest Dissertations & Theses Global. (1729174743)

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). *Comparison of Multistage Tests with computerized Adaptive and Paper-and-Pencil Tests.* ETS Research Report Series, 2007(1).

Smarter Balanced Assessment Consortium (SBAC) (2016). *Smarter Balanced Assessment Consortium: 2014-15 Technical Report*. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Scantron Corporation. (2004). *Technical Manual of Performance Series Computer Adaptive Internet Assessment for Schools.* www.bcvic.net/bcps/bcps-edperformance/PerformanceTechManual.pdf

Shin, C., Chien, Y., & Way, D. (2012). *A comparison of two content balancing methods for fixed and variable length computerized adaptive test*. Paper presented at the 2012 NCME annual conference, Vancouver, Canada.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*(3), 277-292.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools. *ETS Research Report Series*, *1994*(1).

Stone, E., & Davey, T. (2011). Computer Adaptive Testing for Students with Disabilities: A Review of the Literature. *ETS Research Report Series*, *2011*(2).

Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* (Ph.D.). The University of Illinois at Urbana-Champaign, Champaign.

Texas Education Agency (2013). *State of Texas Assessments of Academic Readiness Vertical Scale Technical Report*.

Thompson, N. A. (2009). Item Selection in Computerized Classification Testing. *Educational & Psychological Measurement*, *69*(5), 778–793.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1), 1-9.

Thurlow, M., Elliott, J., & Ysseldyke, J. (1999). Out-of-level testing: Pros and cons (Policy Directions No. 9). Minneapolis, MN: the University of Minnesota, National Center on Educational Outcomes.

van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3-30). Springer New York.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. *Computerized adaptive testing: A primer*, 233-272.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Wang, S., & Zhang, L. (2017). *Effects of Ignoring Discrimination Parameter in CAT Item Selection on Student Scores*. Presented at the National Council on Measurement in Education (NCME) conferences, San Antonio, TX.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*(4), 17-27.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). Some Considerations Related to The Use of Adaptive Testing for The Common Core Assessments. https://www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf

Wei, H., & Lin, J. (2015). Using Out-of-Level Items in Computerized Adaptive Testing. *International Journal of Testing*, *15*(1), 50–70.

Weiss, D. J. & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*(1), 5-21.

Xiong, X. (2010). *An Optimization Technique for Item Pool Evaluation with Guidance for Modification of Item Pools and Test Specifications* (Ph.D.). Fordham University, The Bronx.

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* (Ph.D.). Michigan State University, Ann Arbor.

Yi, Q., Wang, T., & Ban, J. C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 267-292.