NEW METHODS OF ONLINE CALIBRATION FOR ITEM BANK REPLENISHMENT

BY

YI ZHENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang, Chair
Professor Carolyn J. Anderson
Professor Steven A. Culpepper
Professor Jeffrey A. Douglas
Professor Katherine E. Ryan
Professor Jinming Zhang

Abstract

Item parameter calibration is important for tests based on item response theory, because the scoring, equating, bias analysis of the test, and item selection in adaptive tests are all based on the item parameters. As a test is continuously administered, item calibration needs to be conducted for new items at intervals to replace overexposed, obsolete, or flawed items in the item bank. Although it is possible to recruit examinees for the sole purpose of pretesting the new items, a more cost-effective and commonly employed approach is to embed the new items in operational tests. When this approach is employed in computerized adaptive tests (CAT), it is called "online calibration." Analogous to the tailored testing feature in CAT, where an optimal set of operational items are selected for each examinee to more efficiently estimate their ability levels, online calibration makes it possible to select an optimal sample of examinees for each pretest item to more efficiently calibrate their item parameters. During the operational tests, different pretest items can be selected for each examinee. The parameter values of the pretest items are constantly updated, based on which the sampling scheme is dynamically adjusted. A few pretest item selection methods have been proposed, but such development is still in its infant phase. This thesis proposes a new framework for pretest item selection in online calibration. A simulation study was conducted to compare the proposed methods with existing methods and also compare different estimation methods and pretest item seeding locations. Results show significant superiority of the proposed methods compared to existing methods in the 1PL and 2PL models. Middle and late seeding locations lead to more accurate calibration results. The Bayesian MEM estimation method is recommended among the six compared estimation methods.

To my parents, Que Zheng and Wei Peng, without whose nature or nurture I could never have accomplished this thesis.

Acknowledgments

Five years of graduate school is gone like the blink of an eye, but the growth I have experienced in the duration is tremendous. I owe my achievements to many people.

My first and foremost thanks, without doubt, belong to Dr. Hua-Hua Chang, my advisor. Ten thousand miles away from my own parents, I have been cared by Dr. Chang like one of his own children. Like a parent, he teaches us important survival skills rather than spoiling us with abundant wealth; he teaches us how to live a life, not only how to do research; he warns us when our own choices may not lead to an ideal end, even at the risk of looking too harsh; and with his rich experience and savvy insights, he teaches us how to navigate ourselves in the professional world. In addition to all these, thanks to him, I had the invaluable opportunities of working as the managing editor of Applied Psychological Measurement, completing two internships at major testing companies, writing two book chapters, and receiving numerous awards. These are all treasures in a graduate student's portfolio. His own endeavor and accomplishments have also set a great example of ever striving, staying positive, staying strong, and being the master of our own lives. Stepping out of his camp, I feel confident and prepared to enter the new chapter of my life and take new challenges.

I would like to also express my sincere gratitude to other professors at UIUC. Dr. Katherine Ryan has constantly offered generous support, spiritually, scholarly, and financially. Dr. Carolyn Anderson has always been patient and loving. Drs. Jeffrey Douglas, Jinming Zhang, Steven Culpepper, and Larry Hubert have also been very kind and supportive. They and many other professors here are always my cherished mentors and friends.

I had the pleasant and rewarding experiences working during two internships. I would like to thank my mentors at ACT, Inc., Drs. Yuki Nozawa, Xiaohong Gao, Rongchun Zhu, and my mentors at The College Board, Drs. David Chuah and Thomas Proctor. Their advice, kindness, and encouragement have greatly helped my professional development.

I learned a lot of knowledge and skills from my fellow students, especially Dr. Chun Wang, Dr. Ping Chen, and Justin Kern. I thank them and many others for their kindness and friendship, which is an important part of my life. Ms. Jialing He, Dr. Chang's wife, has given her warmest love to all Dr. Chang's students, which makes Champaign like home.

Dr. Mark Reckase has given me advice for my future career as well as valuable ideas for this thesis research, to which I am sincerely grateful.

Lastly, I owe everything to my loving family. Like most parents, my parents have given me selfless support all my life. More importantly, their unique and delicate education has equipped me with many good habits, versatile experiences, and all-rounded skills. Equally important is the example they have set to me as they live their own lives. All the past years I have witnessed how they work devotedly as professors, provide good care of the family, take daily life chores with big hearts, and enjoy all the beauty of life in their leisure time. They are always my role models. And lastly, my husband Zhenhua Li has been a wonderful companion since we met. I thank him for everything he has done in support of me. I feel really blessed to find him on my way and continue this journey with him together!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CTT	Classical test theory
IRT	Item response theory
SEM	Standard error of measurement
IRF	Item response function
1PL	The one-parameter logistic (model)
2PL	The two-parameter logistic (model)
3PL	The three-parameter logistic (model)
CAT	Computerized adaptive testing
MLE	Maximum likelihood estimation
EAP	Expected a-posteriori (estimation)
MAP	Maximum a-posteriori (estimation)
CMLE	Conditional maximum likelihood estimation
JMLE	Joint maximum likelihood estimation
MMLE	Marginal maximum likelihood estimation
EM	Expectation maximization (algorithm)
OIRPI	The ordered informative range priority index
OIRPI-O	The ordered informative range priority index with order statistics
OIRPI-S	The ordered informative range priority index with standardization
i	ID of the examinees
j	ID of pretest items
N_j	Number of calibration samples item j has received
$\boldsymbol{m}_i$	IDs of the operational items examinee i has received
$\boldsymbol{y}_i$	Examinee i's responses to the operational items
x_{ij}	Examinee i's response to pretest item j
$\boldsymbol{\beta}$	Item parameters

$\boldsymbol{\beta}_{op}$	Operational item parameters
$\hat{\boldsymbol{\beta}}_j$	Estimated item parameters of pretest item j
θ	Examinee ability parameter
$\hat{\theta}_i$	Final ability parameter estimate for examinee i
$P_j(\theta)$	Probability of answering item j correctly given examinee ability level θ

Chapter 1

Introduction

A strong trend towards the use of technology-enhanced formative and summative assessments has been seen in recent years. Many educational and psychological assessments are moving to computerized (adaptive) modes. The Race To The Top (RTTT) initiative, for example, requires that K-12 state assessments be computerized and make use of innovative items. As a result, the Partnership for Assessment of Readiness for College and Career (PARCC), a 23 state consortium, is rigorously preparing their online state assessments; and the Smarter Balanced Assessment Consortium (SBAC), consisting of 25 states, is collaborating to design a *computerized adaptive test* (CAT) for their state assessment. Other assessment programs are also undergoing technological and psychometric innovations. For example, the National Assessment of Academic Progress (NAEP) is moving toward computerization and diagnostics, and the Graduate Record Examinations (GRE) is now conducting computerized multistage testing. In the medical field, many patient-reported outcome measurements have adopted CAT to shorten the assessment time and relieve patient burden.

As test designs become more sophisticated and incorporate more modern technologies, more and more testing programs start to switch from the traditional *classical test theory* (CTT) to *item response theory* (IRT) as their underlying measurement theory. Compared to CTT, IRT offers more information about the items and great flexibility in scoring and scaling the assessed traits. More importantly, the item-independent scoring in IRT is necessary in the implementation of CATs. When IRT is used to model test data, *item calibration* is a crucial component of test operation. Item calibration refers to the procedure of fitting IRT models to response data collected from a sample of examinees and estimating the item parameters using the data. Modern assessments depend upon large item banks. Every item in the item bank needs to be calibrated before being used in operational tests, and the accuracy of the calibrated item parameters directly impacts the validity and reliability of the test, as reflected in examinee scoring, equating, differential item functioning analysis, etc.

As the test continues to be administered, item calibration is also needed at intervals to replenish the item bank by replacing overexposed, obsolete, or flawed items with new items. This is called “item bank replenishment”. When there is a frequent demand for item bank replenishment, finding methods to more efficiently and accurately calibrate items becomes imperative. This especially applies to the new state assessments, which are high stakes and therefore place high demands on frequent item bank replenishment.

Although it is possible to recruit examinees and conduct separate pretests for the sole purpose of pretesting the new items (this may be necessary when an item bank is built for the first time), a more cost-effective and commonly employed method is to embed new items in operational tests. These embedded new items are not used for scoring the examinees; instead, they are administered to collect response data for the calibration. After the responses are collected, the item parameters are estimated and linked to the existing scale. For example, in a large-scale state assessment, the entire examinee population is divided into six parts, and each part of the population takes 1/6 of all pretest items; The number of pretest items in a test is about 1/10 of the total test length. Besides cost-effectiveness, another important benefit of this strategy is that it generally ensures that examinees have the same motivation in completing the pretest items as completing the operational items. The consistency of test environment and examinee motivation strengthens the validity of item calibration and reduces item parameter drift between the pretest and operational tests. However, this is still not the most efficient way.

When the approach of embedding pretest items in operational tests is employed in CAT, it is then referred to as *online calibration* (Stocking, 1988). Analogous to the tailored testing feature in CAT, where an optimal set of operational items are selected for each examinee to more efficiently estimate their ability levels, online calibration also makes it possible to select an optimal sample of examinees for each pretest item to more efficiently calibrate their item parameters. During the operational tests, different pretest items can be selected for each examinee. The parameter values of the pretest items are constantly updated, based on which the sampling scheme is dynamically adjusted. The sampling can be terminated once a satisfactory accuracy of the parameter estimation is obtained. A few pretest item selection methods have been proposed for the CAT context, aiming to make the calibration more efficient. But such development is still in its infant phase.

One goal of my research is to develop and evaluate a new framework (with two algorithms) for selecting pretest items in online calibration. A simulation study has been conducted to evaluate the new methods and compare them with the existing methods. Two other goals are (1) to investigate the performance of different statistical estimation methods in the online calibration context, and (2) to compare the effects of different seeding locations (i.e., the embedding locations of pretest items in a test). These two dimensions are also included in the simulation study. The findings can be useful in not only the CAT context but also other computerized testing modes. With carefully developed online calibration designs, the item banks will hopefully be replenished more efficiently with more accurately calibrated new items.

To understand the online calibration study in this thesis, the basics of IRT and CAT that are most closely related to online calibration are first reviewed in Chapter 2. Subsequently, a general introduction to online calibration is presented in Chapter 3. A review of existing methods of online calibration is given in Chapter 4. The proposed pretest item selection methods are presented in Chapter 5, and the details of the estimation methods are described in Chapter 6. Chapter 7 presents the simulation study. Finally, this thesis concludes with a discussion and suggestions for future research directions.

Chapter 2

Backgrounds

2.1 Overview of Item Response Theory

Classical test theory (CTT) and item response theory (IRT) are two popular statistical frameworks for handling test design and analysis. CTT was developed earlier whereas IRT is more statistically sophisticated. Table 2.1 provides an incomplete summary the differences between CTT and IRT.

Allen and Yen (1979) offer a comprehensive introduction to CTT in their classic textbook. In the CTT framework, the most common examinee scoring strategy is to report the total score of the test. The difficulty index of an item is the proportion of examinees who answer the item correctly. As a result, neither the examinee score nor the item difficulty index is sample-invariant; in other words, the examinee scores depend on the difficulty of the items in the test, and the item difficulty index depends on the ability levels of the examinees who took the item. The *standard error of measurement* (SEM) in CTT is usually computed through the reliability index of the entire test, which leads to the counterintuitive fact that the SEM in CTT cannot vary for different examinees.

IRT uses a variety of probability models to model the probability of correct responses if the item is dichotomously scored, or of different levels of responses if the item has more than two possible score levels. These probabilities depend on the parameters of the specific item and examinee. When the items are dichotomously scored, such as multiple choice questions, the currently most common IRT models are the *one-parameter logistic* (1PL) model, *two-parameter logistic* (2PL) model, and the *three-parameter logistic* (3PL) model. The probability of a correct response to item j from examinee with ability level θ is modeled

Table 2.1: Incomplete summary of the differences between CTT and IRT

	CTT	IRT
Population dependency	Item statistics are population-dependent.	Item parameters are population-independent.
Item dependency	Examinee scores are test-dependent.	Examinee ability estimates are item-independent.
Reliability (standard error of measurement)	All scores have the same reliability.	Each ability value has its own reliability.
Assumptions	Unidimensionality	Local independence
Item difficulty	All items have the same contribution to the total score regardless of their difficulty, if not weighted.	The difficulty of items contribute to the estimation of examinee ability.
Item discrimination	Item discrimination is not considered in examinee scoring.	Item discrimination contributes to the estimation of examinee ability, unless the model holds it constant.
Scale comparability	Item difficulty and examinee score are on different scales.	Item difficulty and examinee scores are on the same scale.

by the following *item response functions* (IRFs):

1PL

$$P_j(\theta) = \frac{1}{1 + \exp\left[-(\theta - b_j)\right]}\,; \tag{2.1}$$

2PL

$$P_j(\theta) = \frac{1}{1 + \exp\left[-a_j(\theta - b_j)\right]}\,; \tag{2.2}$$

3PL

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp\left[-a_j(\theta - b_j)\right]}\,. \tag{2.3}$$

In these equations, a_j is the discrimination parameter of item j, b_j is the difficulty parameter of item j, and c_j is the pseudo-guessing parameter of item j. All these item parameters can vary by the individual items, which describe the characteristics of each item. The 1PL model is the simplest among the three, but it has the strongest assumption: all items are assumed to have equal discrimination power and no chance for guessing. The 2PL model assumes no chance for guessing but allows for varied discrimination power as modeled by the a-parameter. The 3PL model includes all three parameters, which can delineate a richer profile of an item. The mission of online calibration, or more generally item calibration, is to estimate these item parameters through certain statistical algorithms using a sample of response data.

When the responses to an item can be scored with more than two levels, such as short response with partial credits, there are a variety of polytomous IRT models to model the response data, such as the *graded response model* (Samejima, 1969), the *partial credit model* (Masters, 1982), the *generalized partial credit model* (Muraki, 1992), the *rating scale model* (Andrich, 1978), and the *nominal response model* (Bock, 1972). Embretson and Reise (2000), Nering and Ostini (2011), and van der Linden and Hambleton (1997) all give a comprehensive introduction to the most commonly-used polytomous IRT models.

In IRT, both examinee parameters and item parameters are sample-invariant. This means that if a different set of items are administered to the examinee, the estimation of ability parameter should generate the same value, excluding random perturbation. Likewise, if a different group of examinees took the item, the estimation of item parameters should also generate the same values, excluding random perturbation. This sample-invariant property sets the foundation for the adaptive item selection in CAT, and by the same means, it sets the foundation for the optimal sampling in online calibration, which is the main topic of this thesis.

In IRT, the SEM is no longer held constant across different levels of examinee ability. Instead, the *Fisher information*, a classic statistical index, was brought into IRT to provide the lower bound of the squared SEM at each θ level. In Statistics, the Fisher information is defined as:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\middle|\theta\right], \tag{2.4}$$

where θ denotes the unknown parameter, X denotes the data, and $\log f(X;\theta|\theta)$ denotes the log-likelihood function. Under dichotomous IRT models, the Fisher information of examinee ability θ can be reduced to:

$$I(\theta) = \sum_{j=1}^{J}\frac{[P_j'(\theta)]^2}{P_j(\theta)[1-P_j(\theta)]}, \tag{2.5}$$

where $j = 1, 2, \cdots, J$ denotes J items the examinee has responded to and $P_j(\theta)$ denotes the item response function formulated by, for instance, Equations 2.1–2.3. When the number of items gets large, the variance of the θ estimate approaches $1/I(\theta)$ from above. Therefore, $1/\sqrt{I(\theta)}$ provides a lower bound of the SEM at θ.

Fisher information plays a crucial role in the item selection in CAT. The most well-known item selection method in CAT, the maximum Fisher information method (F. M. Lord, 1980), selects the item that maximizes the Fisher information for estimating θ at the provisional estimated θ level. This method asymptotically minimizes the SEM of θ in a straightforward way. In item calibration, the Fisher information matrix for the item parameter vector is also essential for the optimal sampling of examinees. The details will be given in Section 5.2.3.

The estimation of the item or examinee parameters relies on statistical algorithms. When the item parameters are known (i.e., calibrated), common algorithms for estimating the examinee ability θ include *maximum likelihood estimation* (MLE), *expected A-posteriori estimation* (EAP), and *maximum A-posteriori estimation* (MAP). When both the item parameters and examinee parameters are to be estimated, the *joint maximum likelihood estimation* (JMLE) algorithm can be used. While JMLE has proven successful for the 1PL model, it meets difficulty with more complex models and larger numbers of examinees. A more commonly used estimation routine is the *marginal maximum likelihood estimation* (MMLE) with the *expectation-maximization* (EM) algorithm. MMLE-EM integrates out the examinee parameter θ using its posterior distribution obtained from the data and then finds the item parameter values that maximize this posterior expected likelihood. Baker and Kim (2004) give a thorough presentation of the variety of parameter estimation methods in IRT. The details of the item parameter estimation methods studied in this thesis will also be given in Chapter 6.

2.2 Overview of Computerized Adaptive Testing

CAT is a modern testing mode that capitalizes on the rapidly developing computer and Internet technology and revolutionizes the testing practices where paper-and-pencil tests have been the mainstream for a long time. The CAT testing mode contains two main components: the computer delivery system based on software engineering and the adaptive algorithms based on psychometric theory.

Let alone the psychometric components, the computer-based test delivery system already has many benefits. For example, computers can generate immediate score reports and save on test form printing, report printing, and human labor. Without the printed test forms stored at scattered locations and transported with innumerable vehicles, the computer delivery could potentially provide better control over the access of tests. Computers also make it possible to administer multimedia items, simulation-based items, and performance-based items. A good example of simulation/performance-based testing program is the medical licensing examination offered by the National Board of Medical Examiners (Clyman, Melnick, & Clauser, 1999). In this exam, examinees are presented with a virtual patient's symptoms and can interact with the computer to order different actions, which will then turn around with more information such as lab reports, based on which the examinees can take the next action until the case is completed.

Underlying the computer delivery system is an important and unique aspect of CAT: the adaptive algorithms based on psychometric theories. With the adaptive algorithms, CAT tailors the test to each individual. The algorithms automatically select the next item to match the examinee's ability level based on his/her responses to previous items. (Details will be given in the next section.) These algorithms can shorten the test by up to 50% without sacrificing the accuracy of the examinee scores (Wainer & Eignor, 2000).

Besides the above, the adaptive algorithms of CAT also make continuous administration possible. In continuous administration, test takers can choose to take the test at their preferred times and locations. If the traditional paper-and-pencil testing mode is adopted, due to test security and fairness concerns, the same test form cannot be used repeatedly during continuous administration, which leads to an unreasonably high demand for test forms. In contrast, CAT naturally fits in this setting, because in CAT, each individual receives a different test form tailored to him/her.

The CAT version of the US Armed Service Vocational Aptitude Battery (CAT-ASVAB) is one of the most successful large-scale applications of CAT (Sands, Waters, & McBride, 1997). It was initiated in the 1970's, developed in the 1980s, launched in the 1990's, and it continues to play a critical role in US military personnel selection. As the psychometric theories and methods advance rapidly, CAT-ASVAB has also been constantly revised and improved. Besides ASVAB, other famous large-scale CAT programs include the Graduate Management Admission Test (GMAT), the National Council of State Boards of Nursing, and the Graduate Record Examination (GRE).

A CAT system typically consists of the following main components:

1. The choice of the initial θ value;

2. Item selection methods;

3. Intermediate and final θ estimation methods; and

4. Stopping rules.

An initial θ value is needed for an examinee at the very beginning of the test when no information of the examinee is known. A simple option is to use the anticipated mean of the ability distribution as the initial value for all examinees. Some randomness can also be included to avoid giving similar initial items to examinees.

The item selection methods are the most important component in the CAT system. They not only need to serve the purpose of optimizing the statistical efficiency for estimating the examinee ability parameters, but also satisfying multiple nonstatistical constraints (e.g., content balancing, word count, answer key balancing) and controlling the exposure of each item. More details on item selection in CAT will be given in the next section.

The θ estimation can be done through a variety of methods developed for IRT, such as MLE. Mislevy and Chang (2000) and H.-H. Chang and Ying (2009) provide discussions that support the legitimacy of using traditional MLE to estimate θ in CAT.

The stopping rule component includes both fixed length CAT and variable length CAT with various stopping rules such as terminating the test once a satisfactory SEM is reached for the examinee.

2.3 Item Selection Methods in CAT

In plain language, the adaptive item selection in CAT mimics what a wise examiner would do: if the examinee answers an item correctly, he will not give an even easier item; instead, he gives a harder item. Likewise, if the examinee answers an item incorrectly, he will not give an even harder item; instead, he gives an easier item (Wainer, 2000). In this way, examinees can avoid taking too many redundant, non-informative items, and this leads to the original idea of adaptation that the items whose difficulty levels match the examinee's ability should be selected.

CTT is not flexible enough for supporting CAT, because the total score is item-dependent; in other words, if different examinees take different items like in CAT, their total scores are not directly comparable. However, the statistical estimation of the ability θ in IRT is item-independent, which naturally supports CAT. Moreover, because IRT puts the item difficulty parameter on the same scale with the examinee ability parameter, the idea of matching item difficulty with examinee ability then becomes simply matching the values of the item difficulty parameter and the examinee ability parameter.

Another angle to understand the item selection in CAT is through the SEM under the IRT framework. Recall that the lower bound of the SEM of θ is provided by $1/\sqrt{I(\theta)}$. Therefore, items with larger information values will produce smaller SEM of θ, namely, better measurement accuracy for the examinee proficiency level. In fact, too easy or too hard items will provide little information for estimating the examinee's proficiency level, while the items with matching difficulties are usually the most informative. This can be mathematically shown for the 1PL, 2PL, and 3PL models from their information functions (H.-H. Chang & Ying, 2009).

For the 1PL model,

$$I(\theta|b) = \frac{\exp(\theta - b)}{[1 + \exp(\theta - b)]^2} \,. \tag{2.6}$$

For a given examinee, $I(\theta)$ attains its maximum value $1/4$ at $b = \theta$.

For the 2PL model,

$$I(\theta|a, b) = \frac{a^2 \exp[a(\theta - b)]}{\{1 + \exp[a(\theta - b)]\}^2} \,. \tag{2.7}$$

For a given examinee and fixed a-parameter value, $I(\theta)$ attains its maximum value $a^2/4$ at $b = \theta$.

For the 3PL model,

$$I(\theta|a, b, c) = \frac{(1 - c)a^2 \exp[2a(\theta - b)]}{\{c + \exp[a(\theta - b)]\}\{1 + \exp[a(\theta - b)]\}^2} \,. \tag{2.8}$$

For a given examinee and fixed a-parameter and c-parameter values, $I(\theta)$ reaches its maximum when

$$b = \theta - \frac{1}{a} \log \frac{1 + \sqrt{1 + 8c}}{2} \,. \tag{2.9}$$

This is typically in the neighborhood of the θ value. For example, when $a = 2$ and $c = 0.25$, the optimal difficulty of the selected item $b \approx \theta - 0.15$.

A plethora of item selection methods for CAT have been developed since the first invention of CAT. The most popular one is the *maximum Fisher information method* (F. M. Lord, 1980). It selects the next item that maximizes the Fisher information at the current θ level. This method minimizes the standard error of measurement in a straightforward way as explained above. Many CAT programs used this method, but a well-known side-effect of this method is that it always prefers items with high a-parameter values (this is easily seen from Equations 2.7 and 2.8), causing a severely skewed distribution of item exposure rates. For example, Wainer and Eignor (2000) report an investigation of the empirical data obtained from an experimental CAT version of Scholastic Assessment Test (SAT) and they found that 17% of the verbal pool accounted for 50% of the tests administered, and about 33% of the pool accounted for 75% of tests administered. Two undesirable consequences of this side-effect are: (1) the high-a items are overly exposed, which poses a test security risk; (2) the low-a items are rarely or never used, which is an expensive waste of item development cost.

To level off the distribution of item exposure rates and prevent over-exposure of some items, some exposure control algorithms and alternative item selection methods were developed. The most famous and commonly used item exposure control algorithm is the *Sympson-Hetter method* (Sympson & Hetter, 1985). The general idea of the Sympson-Hetter method is to put a "filter" between the selection and administration of an item. After an item is selected, a random probability experiment is conducted to determine whether it will be administered. The conditional probability of administration given the item is selected, $P(A|S)$, is carefully calibrated for each item. The more "popular" an item is, the lower the conditional probability is. In this way, the maximum exposure rate can be kept below a certain desirable threshold, and the use of the over-exposed items will be spread over the less popular items.

One limitation of the Sympson-Hetter method is that it does not proactively raise the exposure rates of the under-exposed items. To address this, H.-H. Chang and Ying (1999) proposed the *a-stratified item selection* method, with an improved version *a-stratified with b-blocking* method (H.-H. Chang, Qian, & Ying, 2001), to intentionally select low-a items in the early stage of a test. This does not only proactively increase the use of the under-exposed item, but also, the low-a items can be more helpful than the high-a items in the early stage of a test when the θ estimate is not accurate enough (H.-H. Chang & Ying, 1996). Another alternative, the *maximum Kullback-Leibler information method* (H.-H. Chang & Ying, 1996), is also able to naturally level off the item exposure rates.

Besides the above-mentioned, many other item exposure control strategies have also been proposed. Georgiadou, Triantafillou, and Economides (2007) classifies them into five categories: (1) randomization strategies; (2) conditional selection strategies; (3) stratified strategies; (4) combined strategies; and (5) multiple stage adaptive test designs. They provide an outstanding review of each strategy.

An operational test blueprint usually contains some nonstatistical constraints, such as content balancing, word count, answer key balancing, enemy items (i.e., items that should not appear in the same test form together), etc. A few item selection methods have been developed to satisfy these nonstatistical constraints while simultaneously optimizing the statistical efficiency for estimating examinee ability levels. The *weighted deviation model* (Swanson & Stocking, 1993) and the *normalized weighted absolute deviation heuristic* (Luecht, 1998) treat all constraints as targets and form the criterion as the weighted sum of (normalized) deviations from the targets and then select the item that deviates the least from the target. The *maximum priority index* (Cheng & Chang, 2009) imposes a multiplier that accounts for the nonstatistical constraints before the traditional Fishier information term and selects the item that maximizes this product. The *shadow test* method (van der Linden & Veldkamp, 2004) utilizes the well-established integer programming framework to solve for the optimal "shadow test" that maximizes the information and satisfies all constraints. The integer programming method is not guaranteed to yield a solution, but any solution that is achieved strictly satisfies all constraints. The first three "heuristic" methods always produce a result and are less computationally intense, but they do not guarantee that all of the constraints will be met (Zheng, Wang, Culbertson, & Chang, 2014).

Chapter 3

Introduction to Online Calibration

3.1 General Procedures of Online Calibration

The general procedures of online calibration for item bank replenishment can be summarized by Figure 3.1. Steps 1–3 prepare the pretest item bank. Step 4 is the sampling step: during the operational test, when an examinee reaches a *seeding location* (i.e., the item position in an operational test allocated for pretest items), pretest items are selected from the pretest item bank based on certain item selection rules. Examinees' responses to these pretest items are not included in the scoring procedure. These responses are only used for calibrating the item parameters. When the examinee finishes his/her test, Step 5 is carried out where the item parameters of those administered pretest items are updated. Steps 4 and 5 are repeated for every new examinee, where the sampling algorithm is constantly adjusted based on the updated parameter values. The sampling process of each pretest item can be terminated and the item be exported from the pretest bank separately once a satisfactory accuracy of the parameter estimates is achieved or the maximum sample size is reached. As a final step (Step 6), the exported pretest items are reviewed and if approved, put into operational use.

3.2 Online Calibration as Applied Optimal Design

The general procedures of online calibration presented in the previous section fall under the umbrella of *optimal design*. Optimal design is an active subfield in statistics, which seeks statistical solutions to experiment design or sampling design in order to improve the efficiency and reduce the cost of the projects. Optimal design theories have been frequently applied to a broad range of fields, such as engineering, chemical engineering, education,

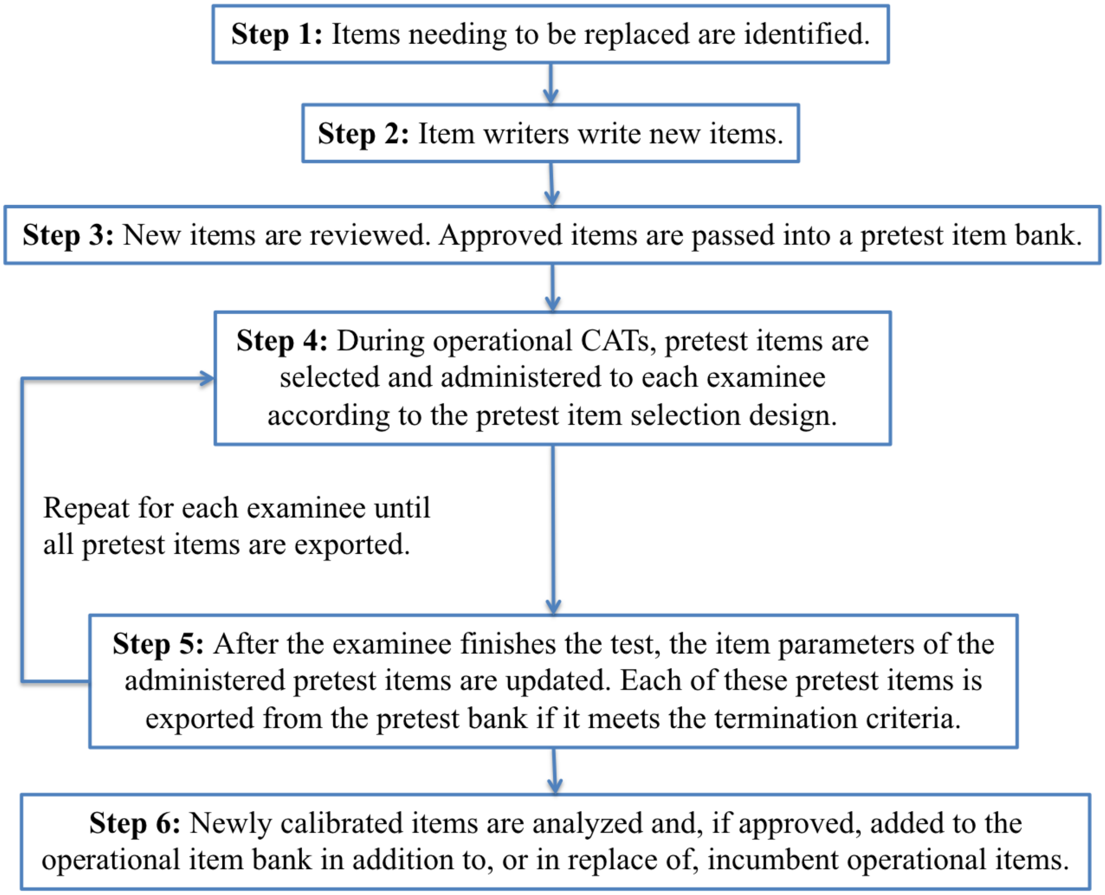

Step 1: Items needing to be replaced are identified.
Step 2: Item writers write new items.
Step 3: New items are reviewed. Approved items are passed into a pretest item bank.
Step 4: During operational CATs, pretest items are selected and administered to each examinee according to the pretest item selection design.
Repeat for each examinee until all pretest items are exported.
Step 5: After the examinee finishes the test, the item parameters of the administered pretest items are updated. Each of these pretest items is exported from the pretest bank if it meets the termination criteria.
Step 6: Newly calibrated items are analyzed and, if approved, added to the operational item bank in addition to, or in replace of, incumbent operational items.

Figure 3.1: The flowchart of item bank replenishment.

biomedical and pharmaceutical research, business marketing, epidemiology, medical research, environmental sciences, and manufacturing industry (M. P. Berger & Wong, 2005). In the setting of educational testing, the application of optimal design includes two main aspects. One aspect is administering the best items to an examinee to optimize the efficiency of estimating his/her ability level, and this is exactly what a CAT is purported for. The other aspect is finding the best sample of examinees to take the pretest items in order to optimize the efficiency of item calibration, which is the topic this thesis discusses.

When no information regarding the examinee ability is available, the best possible sampling method is perhaps random sampling, so that the sample is similar to the population of examinees. This is typical for the item calibration in the paper-and-pencil testing mode.

However, previous studies (e.g., M. P. F. Berger, 1991; F. Lord, 1962) showed that greater efficiency could be obtained if the calibration sample is chosen selectively based on both the item parameter values and examinee ability levels. This is then the motivation for online calibration in CAT.

However, in reality, the true values of neither the item parameters nor the examinee ability levels are ever known. A compromising solution is to use the θ values estimated from the operational items in place of the true examinee ability values, and use the provisional estimates of the item parameters in place of the true item parameter values. The provisional estimates of item parameters are updated sequentially as more data accumulate, and this design is called "optimal sequential design" (Jones & Jin, 1994), "sequential design" (Ying & Wu, 1997), or "sequential sampling design" (M. P. F. Berger, 1992). Luckily, Ying and Wu (1997) have shown that under certain regularity conditions, this sequential design converges to the optimal design; Y.-c. I. Chang (2011) has proven that under regularity conditions, the sequential design is asymptotically consistent and efficient when measurement errors of θ are present. These mathematical proofs provide the theoretical foundations for online calibration designs.

3.3 Fully Sequential Design versus Group Sequential Design in Online Calibration

In sequential sampling designs, there are "fully sequential design" and "group sequential design" (e.g., Armitage, 2002). In a fully sequential design, the sampling scheme is updated after every single new observation is obtained. In a group sequential design, the sampling scheme is only updated after a batch of samples are obtained and analyzed.

Note that many existing publications on online calibration use group sequential designs. In other words, Step 5 in Figure 3.1 is not carried out every time an examinee finishes his/her test; rather, the parameters of the pretest items are only updated after a batch of examinees finish their tests (e.g., Ban, Hanson, Wang, Yi, & Harris, 2001; Chen, Xin, Wang, & Chang, 2012; Kingsbury, 2009; van der Linden & Ren, 2014). Group sequential designs are useful when the interim analysis takes a considerably long time, which impedes the overall efficiency or practical feasibility. Otherwise, fully sequential designs are expected to be more efficient under ideal conditions. When a large-scale simulation is conducted, the choice of a fully sequential design could lead to unreasonable computation time. In this study, a group sequential design is used, where the pretest item parameters are updated after obtaining every 10 new samples. Nevertheless, when the situation allows, fully sequential designs should be considered in order to maximize the benefits of optimal sequential designs.

3.4 Intended Benefits of Online Calibration

Given the limited resources (e.g., examinees, time), optimal sequential sampling designs applied in online calibration could increase the accuracy of the calibrated item parameters (e.g., M. P. F. Berger, 1992; Buyske, 2005; Jones & Jin, 1994). In other words, to achieve the same calibration accuracy, an optimal sampling design requires fewer examinees than simple random sampling, which is typically implemented in traditional paper-and-pencil non-adaptive tests. Moreover, by assigning different pretest items to each individual, this adaptive online calibration design should pose less test security risk than assigning the same block of pretest items to a convenient sample (e.g., a school, a district) as often done in paper-and-pencil tests.

The techniques of online calibration may also be useful for the purposes of recalibrating items whose parameters may have drifted or establishing vertical scales across multiple grade levels. In the former setting, online calibration provides a fast turn-around of recalibrated item parameters during the normal test operation and an individual treatment for each focal item. In the latter setting, the CAT can be administered to students whose ability levels are beyond the normal range of their own grade and reach the ability ranges of the adjacent grades. This naturally provides response data for linking the scales of adjacent grades. With online calibration, these items from different grades can be automatically calibrated onto the universal vertical scale, which gradually builds up the vertical scale for the entire item bank across grades.

3.5 Main Design Factors in Online Calibration

There are perhaps four main design factors in an online calibration design:

1. Pretest item selection method: how to match the pretest items with the examinees based on the characteristics of the item and the examinee;
2. Seeding location: where in a test the pretest items are embedded;
3. Estimation method: which statistical algorithm is used to estimate the item parameters; and
4. Termination rule: when to terminate the sampling of a pretest item and export it from the pretest stage.

Previous literature on online calibration has two main foci: one line of literature focuses on developing pretest item selection methods, and the other line focuses on studying the estimation methods. The next Chapter will review the literature in both lines as well as some other aspects. The main purpose of this thesis is to present a new framework for pretest item selection. The estimation methods and seeding locations are also studied here.

Chapter 4

Review of Existing Methods in Online Calibration

This chapter reviews the currently existing methods in online calibration, including the existing methods for selecting pretest items, deciding seeding locations, estimating the item parameters, terminating the sampling, and others.

4.1 Pretest Item Selection Methods

The pretest item selection method determines how to match examinees with pretest items during the test. The existing pretest item selection methods in the literature can be summarized into three categories: random selection, examinee-centered selection, and item-centered selection. The following sections will review each category in detail.

4.1.1 Random Selection

In random selection, the computer randomly selects a pretest item from the pretest item pool when the examinee reaches the predetermined seeding locations in the test. When the examinee ability is normally distributed, this approach will result in a sample of examinees with roughly normally distributed ability for each pretest item. This method is easy to implement and will provide heterogeneous samples (e.g., Kingsbury, 2009; Chen et al., 2012). However, in an adaptive test where the difficulty of operational items generally follow a trend towards the examinees ability level, the difficulty of a randomly selected pretest item will likely stand out from the surrounding items, which may cause unnecessary confusion and anxiety to struggling examinees (Kingsbury, 2009).

4.1.2 Examinee-Centered Selection

In examinee-centered selection, pretest items are selected by the same item selection method used for selecting operational items (Chen et al., 2012; Kingsbury, 2009). As explained in Section 2.3, the operational item selection criteria in CAT are designed to optimize the estimation of examinee abilities (therefore this method is called "examinee-centered"), but they are not designed for the purpose of calibrating pretest items. Although the examinee-centered method may be a reasonable choice for the 1PL model, it may not be appropriate for other IRT models. A typical 1PL CAT optimizes the efficiency of examinee ability estimation by matching the item difficulty b with the examinee ability θ. It is easy to see from the 1PL item response function (equation 2.1) that using the same selection method for pretest items will also optimize the efficiency of item parameter estimation. However, in other IRT models, the optimal samples vary for different item parameters.

For example, a 3PL model (Equation 2.3) item has three parameters: the discriminating parameter (a), the difficulty parameter (b), and the pseudo guessing parameter (c). Figure 4.1 illustrates the information function for a 3PL model item with a set of reasonable item parameter values: $a = 1.5$, $b = 0$, $c = 0.2$. The higher the information, the more useful the corresponding ability θ is in estimating the parameter. The peaks of the information curves occur at different θ locations for the three parameters. Therefore, roughly matching b with θ, which is a typical operational item selection method in a 3PL model CAT, will provide a large amount of information for b but little information for either a or c. In other words, examinees whose ability levels match the item difficulty level usually provide less information than examinees at some other ability levels for estimating a and c. For a given item to be calibrated, the item should be assigned to a group of examinees so that they will provide a sufficient amount of information for estimating each item parameter. Selecting an inappropriate group of examinees could lead to inefficient or even seriously inaccurate item parameter estimation.

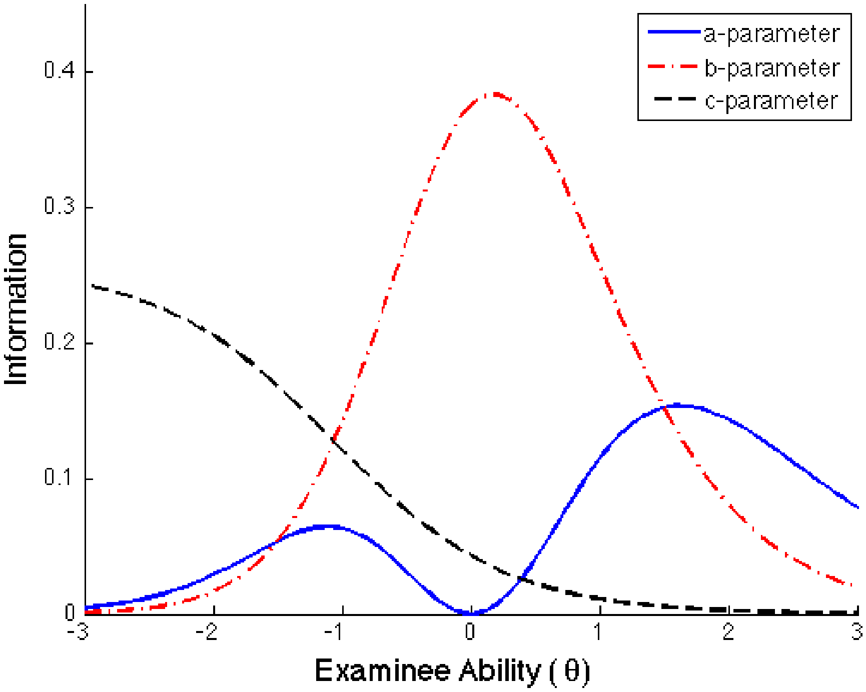


Figure 4.1: The information curves for a 3PL model item ($a = 1, b = 0, c = 0.2$). Note that the scale for the c-parameter curve has been reduced by 10 times for better presentation of all curves.

4.1.3 Item-Centered Selection

The item-centered selection methods match examinees with pretest items based on criteria directly designed to optimize the estimation of the pretest item parameters.

One of the most frequently adopted item-centered criterion in optimal calibration design literature (M. P. F. Berger, 1992; M. P. F. Berger, King, & Wong, 2000) as well as in online calibration literature (Y.-c. I. Chang & Lu, 2010; Jones & Jin, 1994; Zhu, 2006) is the *D-optimal* criterion. It minimizes the *generalized variance* (Anderson, 1984) of the item parameter estimates by maximizing the determinant of the Fisher information matrix of the item parameter vector. The D-optimal criterion is a traditional criterion in optimal design (Silvey, 1980). The detailed formulation of the D-optimal criterion in dichotomous IRT models will be given later in Section 5.2.3.

However, there is a common limitation of the above-cited online calibration papers: they all directly adopt the traditional optimal design paradigm, which is hardly feasible in the online calibration scenario. Using the traditional optimal design paradigm, they assume that there is an "examinee pool" filled with examinees at different ability levels (termed "design space" in optimal design literature); for each pretest item, **the examinees are compared** and the ones whose ability levels maximize the D-optimal criterion (i.e., provides the most information to the item) are selected.

For example, Chang and Lu's (2010) design can be summarized as two successive CATs: one for examinees and one for pretest items.

- Stage 1: Conduct CAT to examinees using operational items and estimate their abilities.
- Stage 2: Conduct CAT to pretest items. For each pretest item, select the examinees whose ability values estimated from Stage 1 satisfy the "2-point D-optimal" criterion. For 2PLM, the two target ability levels can be derived analytically as $\theta_1 = -1.5434/\hat{a} + \hat{b}$ and $\theta_2 = 1.5434/\hat{a} + \hat{b}$. For each pretest item, the CAT terminates when the length of the maximum axis of confidence ellipsoid for the item parameter estimates is no greater than a specified value.

Their design is hardly feasible beyond the simulation context. It requires that all examinees complete their operational tests in Stage 1 and form an "examinee pool" from which the optimal examinees are selected during Stage 2. In practice, operational CATs are administered at scattered times within a testing window. There can rarely be a static examinee pool to choose examinees from, and it is also unknown what ability level the next examinee is at.

Based on the same D-optimal criterion, van der Linden and Ren (2014) implemented a procedure that is practically feasible. In their design, when an examinee reaches the seeding location, **all pretest items are compared** and the item with the maximum Bayesian D-optimal statistic value is selected. However, this method also has its limitation. Among the pretest items, some items tend to produce consistently higher D-optimal statistic values than the others due to their own statistical superiority. Therefore, this design may tend to select those items even if other items need the current examinee more. Figure 4.2 illustrates this issue with a toy example of two 3PL model items. The true parameter values of item 1 are $a = 2, b = 1, c = 0.2$, and those of item 2 are $a = 1.5, b = 0, c = 0.25$. The curves in the figure are the D-optimal criterion values (i.e., the determinant of the Fisher information matrix) generated from a total of 51 examinees, including 50 examinees whose θ values are a random sample from the standard normal distribution, representing the samples the two pretest items have already accumulated, and the last examinee whose θ value is along the abscissa, representing the next examinee to be matched with the items. This figure illustrates that the D-optimal criterion values of item 1 are always greater than those of item 2 regardless of what value the next examinee's ability is. This means between the two items, an item selection method that directly compares the D-optimal criterion values will always select item 1 and neglect item 2. In other words, item 2 will receive no response data until item 1 is exported from the pretest item pool.

With this design, if a test developer terminates the calibration phase at a chosen time, he/she will be left with a pretest item pool where some items have very good parameter estimates but others have very unreliable parameter estimates or even no parameter estimates. In contrast to this situation, a more reasonable and desirable design is the test developer can terminate the calibration phase at a chosen time and obtain fairly equally accurate parameter estimates of all pretest items. If the he/she wishes to increase the accuracies of the item parameter estimates, he/she can continue with the pretest phase. Ren and Diao (2013) addressed this problem by imposing an exposure control constraint in the item selection procedure.

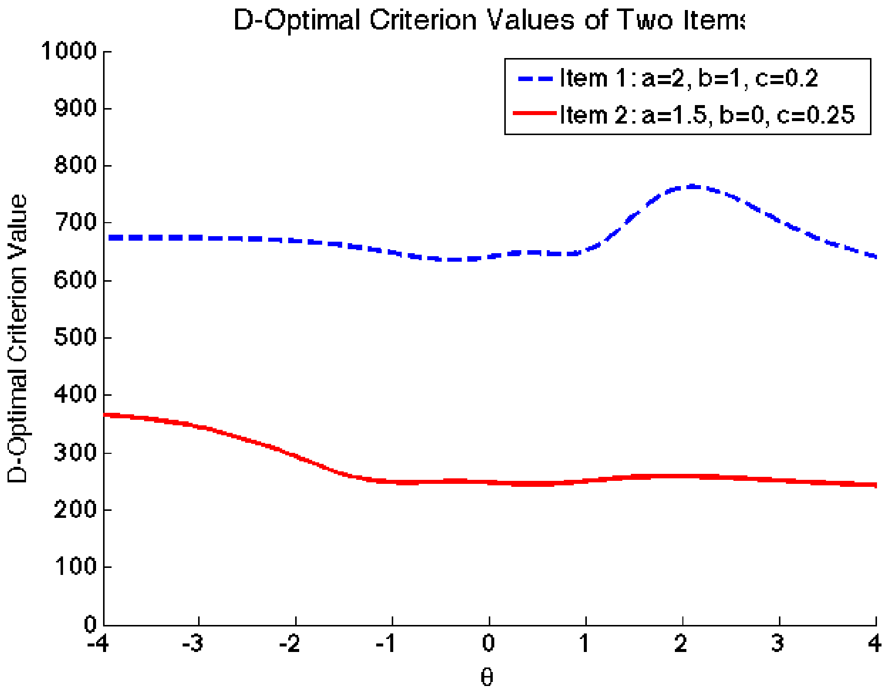


Figure 4.2: An illustration of the consistent superiority in the D-optimal criterion values of some items over the others.

Yet another existing item-centered selection method is the *suitability index* (SI) proposed by Ali and Chang (2011). This index is based on the target sample sizes and current sample sizes of each pretest item in several partitioned ability intervals. When an examinee reaches the seeding locations, the pretest item that maximizes the following SI index will be selected:

$$S_j = \frac{1}{|\hat{b}_j - \hat{\theta}|} \prod_{k=1}^{K} w_k f_{jk} \, , \tag{4.1}$$

where

$$f_{jk} = \frac{T_{jk} - t_{jk}}{T_{jk}} \, , \tag{4.2}$$

and j denotes pretest item, k denotes K ability intervals, T_{jk} denotes the target sample size of item j in ability interval k, and t_{jk} denotes the current sample size of item j in ability interval k. This design seeks to control the sample size from different ability intervals of each pretest item. Their simulation study shows promising results, but the specific partition of ability intervals and their target sample sizes are not clear from the currently available paper.

In summary, currently, pretest item selection methods that are both theoretically sound and practically feasible are still underdeveloped. My research seeks to develop a new item-centered pretest item selection framework for online calibration that is both theoretically sound and practically feasible. This framework can be combined with any criterion. Two algorithms are also developed under this framework. Chapter 5 will introduce the proposed new pretest item selection methods.

4.2 Seeding Locations

The factor of pretest item seeding location could also influence the calibration results. As the seeding location moves towards the end of the test, the θ estimate that is used in selecting the pretest items carries less measurement error, given satisfactory model fit. With a more accurate input parameter value, a pretest item selection method that has a correct mechanism should also generate more accurate output.

Ideally, to achieve the greatest efficiency, the pretest items should be seeded at the very end of the test, where the θ estimates contain the least measurement error. However, if this information is leaked to examinees, and they know that the pretest items will not be included in scoring, their motivation in completing these items may significant decrease, which can cause item parameter drift between the pretest and operational tests and impair the validity of item calibration. Therefore, a more practical way is to randomly determine the seeding locations.

Previous studies adopt different designs for seeding locations, such as randomly assigned throughout the entire test (Chen et al., 2012), fixed at items 22 and 28 in a test of 52 items

(Kingsbury, 2009), randomly assigned to three out of the last six positions in a test of 28 items (van der Linden & Ren, 2014). Kingsbury (2009) also suggests a rule that “no more than one pretest item should be administered consecutively”.

The factor of seeding location merits both theoretical and empirical study. A study of the effect of different seeding locations could reflect the soundness of a pretest item selection method or identify a potentially flawed pretest item selection method. For instance, if the calibration accuracy decreases with seeding locations later in the test, the pretest item selection method could be flawed, because later seeding locations come with more accurate θ estimates, which are taken as the input of the pretest item selection methods, and if the pretest item selection mechanism is correct, it should produce more accurate calibration results. The optimal choice of the seeding location may also interact with the pretest item selection method and may also vary in different stages of the accumulation of sample sizes. The simulation study in Chapter 7 will investigate the impact of different seeding locations.

4.3 Statistical Estimation Methods

A second line of literature addresses the statistical estimation methods in online calibration. Their research question is “how to utilize the known parameter values of the operational items to aid the estimation of the pretest items”. In a traditional calibration, oftentimes none of the items have known parameter values and all of them need to be calibrated. But there are also occasions where a part of the items have known parameter values and the remaining are calibrated, such as equating the scale of one test form to the known scale of another test form. These methods have been studied under the name “fixed-parameter calibration” (e.g., Kim, 2006). The estimation problem in online calibration is essentially the same with fixed-parameter calibration, in which the operational item parameters are fixed and the pretest item parameters are calibrated.

Ban et al. (2001) summarized five statistical methods that are applicable in online calibration to estimate pretest item parameters. They are Stocking-A, Stocking-B, the marginal maximum likelihood estimation (MMLE) with one expectation-maximization (EM) cycle method (OEM), the MMLE with multiple EM cycles method (MEM), and the BILOG with strong prior method.

- The *Stocking-A* method (Stocking, 1988) first estimates examinee ability θs using all the administered operational items and then it estimates pretest item parameters using *conditional maximum likelihood estimation* (CMLE) conditional on the estimated θ values. Stocking-A is the simplest method but may suffer from scale drift caused by using the estimated examinee ability rather than the true values (Stocking, 1988).

- The *Stocking-B* method (Stocking, 1988) is Stocking-A with a follow-up equating step using anchor items to correct for the aforementioned scale drift. It is theoretically more rigorous than Stocking-A but practically difficult because the need for anchor items increases the test length and calibration sample size (Ban et al., 2001).

- The *OEM* method (Wainer & Mislevy, 2000) is the MMLE-EM method with one EM cycle. It first obtains the posterior θ distribution using all the administered operational items. Then it uses this posterior θ distribution to marginalize the likelihood function for estimating the pretest item parameters. The estimated item parameters are those that maximize the posterior expected likelihood function.

- The *MEM* method (Ban et al., 2001) is the MMLE-EM method with multiple EM cycles. The first cycle is the same with the only cycle in OEM; starting with the second cycle, both operational and pretest items are used to update the posterior θ distribution for the MMLE-EM estimation. The EM iteration continues as the pretest item parameters are updated; the iteration is deemed converged if the estimates differ by no greater than a small threshold.

- The *BILOG with Strong Prior* method (Ban et al., 2001) utilizes the BILOG (Mislevy & Bock, 1990) software to calibrate pretest items in one single run. It fixes the operational item parameters by setting fairly strong prior distributions on the operational items. It then calibrates the pretest and operational items simultaneously.

With a comprehensive simulation study, Ban et al. (2001) made the following conclusions: (1) MEM produced the smallest parameter estimation errors but was the most time-consuming; (2) OEM produced larger errors than MEM; (3) Stocking-B also produced satisfactory measurement accuracy but the need for anchor items lengthens the test and increases the sample size; (4) The BILOG with Strong Prior method may not be a reasonable choice for small sample sizes until more appropriate ways of handling sparse data are devised; and (5) Stocking-A had the largest weighted total error as well as theoretical weakness of scale drift. In summary, they recommend MEM among the five methods, with OEM as a less computationally intense alternative.

Based on their recommendations, this study includes Stocking-A, OEM, and MEM. When sample sizes are small, all these three methods can be combined with some Bayesian priors on the item parameters. These methods will be explained in greater detail in Chapter 6.

One major advantage of online calibration is that the calibrated item parameters are automatically on the existing scale. All of the aforementioned methods directly put the calibrated item parameters on the same scale with the operational item bank, and no linking is needed afterwards. This is also why the techniques of online calibration may aid the recalibration task for the purpose of detecting item parameter drift and also aid the construction of vertical scales.

4.4 Termination Rules

Termination rules of online calibration determine when to stop sampling and export pretest items from the pretest stage. The most straightforward termination rule is based on sample sizes (e.g., Ali & Chang, 2011; Kingsbury, 2009; Zhu, 2006): once the sample size of a pretest item reaches a predetermined value, the item is exported. However, with the same sample size, the parameters of different items can have various levels of standard error. If achieving a satisfactory accuracy of item parameter estimation is the main goal of calibration, termination rules may also be defined based upon standard errors: a pretest item is exported once the standard errors of its parameter estimates move below a predetermined threshold. This method is expected to be more efficient than the sample size rule. A maximal sample size can also be specified to prevent overlong pretest duration. Another possible termination rule, as proposed by Kingsbury (2009), is to stop sampling once the item parameter estimates stabilize. Many different specific rules can be designed along this line. Designing and evaluating termination rules can be an important research direction that follows this study.

4.5 Other Factors

Other factors in online calibration design may include the proportion of pretest items in a test, the minimum and maximum sample size, etc. The typical proportion of pretest items in a test found in literature is between 1/10 and 1/4. The decision may depend on the need for new items and other practical aspects. The minimum sample size also depends on various aspects such as requirement on estimation accuracy, urgency in replenishment, etc.

Chapter 5

New Pretest Item Selection Methods

This chapter introduces a new pretest item selection framework for online calibration. This framework is item-centered, meaning that pretest items are selected according to criteria based on the properties of the pretest items. This framework can be combined with any criterion. Two algorithms have also been developed under this framework.

Regardless of which specific criterion and algorithm are used, the following general steps are taken in the proposed online calibration design.

- Step 1: Initializing pretest item parameters.
  - Option 1: Random sampling can be used for each pretest item until a predetermined minimum calibration sample size is reached. Then the pretest items are calibrated using these data from random sampling, and the calibrated item parameters are used as the initial item parameters for the adaptive item selection in Step 2.
  - Option 2: Each pretest item can be classified into several difficulty levels by content experts. The difficulty parameters can then be initialized based on the expert classification and the discrimination or guessing parameters, if the model requires, can be initialized with the most common values.
- Step 2: During the operational CAT, when an examinee reaches the seeding locations, the program will select and administer the most desirable pretest item from the pretest item pool determined by the chosen item selection rule. The seeding locations can be predetermined and fixed or randomly chosen within a certain range.
- Step 3: When an examinee completes his/her CAT, the program will use one of the

statistical estimation methods to update the item parameters of each administered pretest item. Alternatively, when a pretest item has obtained an enough amount of new response data, its item parameters will be updated. For each item being estimated, all relevant response data, including those from the current examinee and those from previous examinees who have taken this item, are used for the estimation procedure. Note that all examinees who have taken the same pretest item did not take the same operational items, and this can be handled by the estimation algorithms as detailed in Chapter 6.

- Step 4: Steps 2 and 3 are iterated for each incoming examinee. The pretest items can be exported from the pretest phase individually once the termination rule is reached for the item. The termination rule can be based on sample sizes or measurement accuracy (e.g., standard error of measurement). Then, the iteration of Steps 2 and 3 continues with remaining pretest items (or the pretest item pool can be replenished with other pretest items) until the testing window is closed.

5.1 Two Approaches to Pretest Item Selection

In most IRT models, the optimal sampling design needs to consider the divergent needs of different parameters. The only exception is the 1PL model, which contains only one item parameter — the difficulty parameter. As shown earlier, Figure 4.1 illustrates the information curves for a 3PL item model item with item parameter values $a = 1, b = 0, c = 0.2$. The peaks of the information curves occur at different θ locations for the three item parameters. Therefore, the pretest item selection method must inclusively take care of the diverse needs of each parameter, and there can be two approaches to achieving this goal.

Approach I Treat the item parameters as a vector and use multivariate methods such as the D-optimal criterion.

As mentioned in the literature review chapter (Chapter 4), the existing pretest item selection methods based on the D-optimal criterion have some practical or theoretical limitations:

- Chang and Lu's (2010) two-stage design *compares all examinees after they finish their operational CATs and selects the most informative examinee for a pretest item.* This design is hardly feasible beyond the simulation context because in an operational CAT, examinees come to the test at varied times and leave afterwards, and there is hardly a static examinee pool to choose the optimal examinee from.
- van der Linden and Ren's (2014) design *compares all pretest items when an examinee reaches a seeding location during the operational CAT and selects the pretest item that generates the highest information criterion value.* This design may tend to select the statistically superior pretest items and ignore the others.

To address these limitations, this thesis proposes a new framework to match examinees with pretest items: *when an examinee reaches a seeding location during the operational CAT, compare all pretest items and select the item that* ***needs*** *the current examinee most.* Comparing this framework with van der Linden and Ren's (2014) design , the new framework is a "need-based" framework, while the latter can be thought of as "merit-based". A merit-based approach selects the item that carries the highest information to calibrate itself with the incumbent examinee; a need-based approach selects the item that needs the incumbent examinee most. For example, in Figure 4.2, if the ability of the current examinee is less than about -2, then item 2 needs this examinee in its sample more than item 1 does, because these θ values are among the most informative design points for item 2 across the entire θ scale, whereas more informative θ's can be found in other examinees for item 1. The central question is then how to quantify the needs of the items. The proposed *Ordered Informative Range Priority Index* (OIRPI) framework implements this "need-based" idea. The details will be given in Section 5.2.

Approach II Pool the different needs and balance them with a heuristic. Ali and Chang's (2011) Suitability Index method takes this approach. The construction of these indices is highly arbitrary. Another possible design is to let different parameters take turns to be the center of optimization, which can be regarded as an "element-wise priority index". Along this line, how to determine the turns, what are the targets, how to weight different elements, and when an element is determined satisfactory all need to be carefully designed. This thesis will not discuss the designs with this approach. It can be a potential area to investigate.

5.2 The Ordered Informative Range Priority Index

The Ordered Informative Range Priority Index (OIRPI) is a new framework for selecting pretest items proposed here. The goal of the OIRPI method is to assign the current examinee to the pretest item that *needs* him/her most, and the method determines "how badly an item needs the current examinee" by how informative this examinee is to this item **compared to examinees at other ability levels**. To describe the OIRPI method at the most generalized level, the term "information" will be used here to represent any measure of information. Two algorithms have been developed for the OIRPI framework. The steps of each algorithm are given below.

5.2.1 Algorithm 1: OIRPI with Order Statistics

Step 1: The first step of OIRPI is to divide the examinee ability scale θ into R contiguous ranges and determine the representative θ value in each range.

One way to divide the θ-scale is equal spacing by the θ value and the representative values are the middle point of each ranges. Another way is equal spacing by the percentiles, for example, if ten ranges are to be created, let P_t denote the t^{th} percentile, the ten ranges are $P_0 \sim P_{10}$, $P_{10} \sim P_{20}$, $\cdots$, $P_{90} \sim P_{100}$, and the representative values are $P_5, P_{15}, \cdots, P_{95}$. Because in practice the θs are roughly normally distributed, the latter is expected to make the OIRPI algorithm more effective.

Then, when an examinee reaches a seeding location, the following steps are carried out **for each item** to obtain their OIRPI values.

Step 2.1: Calculate the information for calibrating this item provided by each θ-range using their representative θ values.

Step 2.2: Order all R θ-ranges by their information values from the smallest to the largest.

Step 2.3: Identify which range the current examinee ability θ belongs to. Then assign the order statistic of that θ-range as the priority index of this item. In this way, if the current examinee ability θ belongs to the θ-range that provides the highest information to the item, this item will receive the highest priority index value.

Step 3: After all items receive their OIRPI values, the item with the highest OIRPI value will be selected for the current examinee. If their are ties, one of them is randomly chosen.

5.2.2 Algorithm 2: OIRPI with Standardization

The first algorithm is straightforward and intuitive, but a potential problem is the ties. When the θ-range the current examinee falls in generates the highest information for more than one pretest items, there are ties and one of them is randomly chosen. This becomes more common when R (i.e., the number of θ-ranges) is small. To eliminate the ties, a second algorithm that is based on standardization is proposed. This algorithm differs from the first algorithm in Steps 2.2 and 2.3.

Step 1: Divide the examinee ability scale θ into R contiguous ranges and determine the representative θ value in each range.

Then, when an examinee reaches a seeding location, the following steps are carried out **for each item** to obtain their OIRPI values.

Step 2.1: Calculate the information for calibrating this item provided by each θ-range using their representative θ values.

Step 2.2: Standardize the information for all R θ-ranges by $S_r = (D_r - \min_{r \in R}(D_r))/(max_{r \in R}(D_r) - min_{r \in R}(D_r))$, where D_r denotes the information value for the rth range.

Step 2.3: Identify which range the current examinee ability θ belongs to. Then assign S_r of that range as the priority index of this item.

Step 3: After all items receive their OIRPI values, the item with the highest OIRPI value will be selected for the current examinee.

This algorithm based on standardization essentially eliminates the ties because S_r is a continuous function.

5.2.3 The D-optimal Criterion

The D-optimal criterion is one of the most popular criteria in optimal design methodology for summarizing the information of multiple parameters (e.g., M. P. F. Berger, 1991, 1992; M. P. F. Berger et al., 2000; Jones & Jin, 1994). In the IRT online calibration context, the D-optimal criterion is the determinant of the Fisher information matrix of the item-parameter vector given the ability parameter θs of all currently sampled examinees.

Specifically, the information matrix of the 3PL model item parameter vector (a, b, c) of the jth item provided by θ_i of the ith examinee is given by the following equations (Hambleton, Swaminathan, & Rogers, 1991):

$$\boldsymbol{I}_j(\theta_i) = \begin{pmatrix} I_{aaij} & I_{abij} & I_{acij} \\ I_{abij} & I_{bbij} & I_{bcij} \\ I_{acij} & I_{bcij} & I_{ccij} \end{pmatrix} . \tag{5.1}$$

Let ℓ_{ij} denote the log-likelihood of the parameter values of item j given the response from examinee i and his/her ability level θ_i, and let $P_j(\theta_i)$ denote the item response function given θ_i defined in equation 2.3,

$$I_{aaij} = -E[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial a_j}] = (\theta_i - b_j)^2 \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2 ; \tag{5.2}$$

$$I_{bbij} = -E[\frac{\partial^2 \ell_{ij}}{\partial b_j \partial b_j}] = a_j^2 \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2 ; \tag{5.3}$$

$$I_{ccij} = -E[\frac{\partial^2 \ell_{ij}}{\partial c_j \partial c_j}] = \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \frac{1}{(1 - c_j)^2} ; \tag{5.4}$$

$$I_{abij} = -E[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial b_j}] = -a_j(\theta_i - b_j) \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \left[\frac{P_j(\theta_i) - c_j}{1 - c_j}\right]^2 ; \tag{5.5}$$

$$I_{acij} = -E[\frac{\partial^2 \ell_{ij}}{\partial a_j \partial c_j}] = (\theta_i - b_j) \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \frac{P_j(\theta_i) - c_j}{(1 - c_j)^2} ; \tag{5.6}$$

$$I_{bcij} = -E[\frac{\partial^2 \ell_{ij}}{\partial b_j \partial c_j}] = -a_j \frac{1 - P_j(\theta_i)}{P_j(\theta_i)} \frac{P_j(\theta_i) - c_j}{(1 - c_j)^2} . \tag{5.7}$$

The corresponding formulas for the 1PL model and the 2PL model can be reduced from the formulas above. For a 1PL model item, $\boldsymbol{I}_j(\theta_i)$ has a single entry I_{bbij}, where I_{bbij} is computed by plugging in $a_j = 1$ and $c_j = 0$ in Equation 5.3. For a 2PL model item $\boldsymbol{I}_j(\theta_i)$ contains the first two rows by the first two columns of the matrix in Equation 5.1 and the entries are computed by plugging in $c_j = 0$ in Equations 5.2, 5.5, and 5.3.

Note that based on the assumption that the responses to an item j from different examinees are independent, the likelihood of the item parameter values given a vector of responses from

N examinees are the direct product of the likelihood given each examinee's response. Thus, the log-likelihood and consequently the Fisher information have this following nice additive property:

$$\boldsymbol{I}_j(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{I}_j(\theta_i)\,, \tag{5.8}$$

which means the Fisher information matrix of the parameter values of an item j given a vector of $(\theta_1, \theta_2, \cdots, \theta_N)$ is the direct summation of the Fisher information matrices of item j given each individual θ_i.

The determinant of $\boldsymbol{I}_j(\boldsymbol{\theta})$ provides a scalar summary of the information for estimating the parameters of item j provided by the sample of examinees. A greater $|\boldsymbol{I}|$ value indicates higher information, which is associated with smaller generalized variance of the vector of item parameter estimates, defined as $|\mathrm{Cov}(\hat{\boldsymbol{\beta}})|$ (Anderson, 1984), where $\hat{\boldsymbol{\beta}}$ denotes the estimated item parameter vector:

$$\operatorname*{argmax}_{\theta} |\boldsymbol{I}| = \operatorname*{argmin}_{\theta} \frac{1}{|\boldsymbol{I}|} = \operatorname*{argmin}_{\theta} |\boldsymbol{I}^{-1}| \xrightarrow{asy} \operatorname*{argmin}_{\theta} |\mathrm{Cov}(\hat{\boldsymbol{\beta}})|\,. \tag{5.9}$$

In other words, the higher the D-optimal criterion value, the smaller the estimation error for item parameter calibration, that is, the more accurate the item parameter estimates are. Holding the sample size constant, a higher D-optimal criterion value leads to a more efficient calibration. Note that Equation 5.9 holds when $\boldsymbol{I}$ is computed using the true values of examinee ability θ's and item parameter vector $\boldsymbol{\beta}$. In online calibration, a sequential procedure, which is described in Chapter 3 is adopted to gradually approximate the real optimal design. The θ values estimated from the operational items are used in place of the true examinee ability values, and the provisional estimates of the item parameters are used in place of the true item parameter values. The provisional estimates of item parameters are updated sequentially as more data accumulate. Ying and Wu (1997) have shown that under regularity conditions, this sequential design converges to the optimal design; Y.-c. I. Chang

(2011) proved that under regularity conditions, the sequential design is asymptotically consistent and efficient when measurement errors of θ are present. For more details, readers can also refer to M. P. F. Berger (1991).

In the OIRPI method, when an examinee reaches a seeding location, the D-optimal criterion value is computed in Step 2.1 for the representative value of each θ range for every pretest item j. Note that previous to this examinee's test, other examinees have each taken several of these pretest items, so each pretest item has accumulated their own samples. Let $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_{N_j})$ denote the ability estimates of the N_j examinee samples item j has already accumulated; let θ_r denote the representative value of the rth θ-range. The D-optimal criterion value for adding the (N_j+1)th sample to pretest item j provided by the rth θ-range is computed through the following formula:

$$D_{rj} = \Big| \sum_{i=1}^{N_j} \boldsymbol{I}_j(\hat{\theta}_i) + \boldsymbol{I}_j(\theta_r) \Big| . \tag{5.10}$$

In Step 2.2 of OIRPI, for each pretest item j, if Algorithm 1 is used, $D_{1j}, D_{2j}, \cdots, D_{Rj}$ are compared and ordered from smallest to largest, and the order statistic of the range the current examinee θ belongs to is assigned as the priority index for item j. If Algorithm 2 is used, D_{rj}'s are standardized by

$$S_{rj} = \frac{D_{rj} - \min_{r \in R}(D_{rj})}{max_{r \in R}(D_{rj}) - min_{r \in R}(D_{rj})} . \tag{5.11}$$

Then S_{rj} of the range the current examinee θ belongs to is assigned as the priority index for item j.

Chapter 6

Statistical Estimation Methods

This chapter describes the details of the six statistical estimation methods investigated in this study. As discussed in Section 4.3, based on Ban et al.'s (2001) recommendations, the first three estimation methods included in this study are (1) Stocking-A (Stocking, 1988), (2) OEM (Wainer & Mislevy, 2000), and (3) MEM (Ban et al., 2001). Because the calibration sample sizes in online calibration are typically small — one goal of online calibration is to achieve satisfactory calibration results using small samples — to stabilize the estimation algorithms, Bayesian priors on item parameters can be added to the estimation algorithms. Therefore, this study also includes the Bayesian versions of these three methods, that is, (4) Bayesian Stocking-A, (5) Bayesian OEM, and (6) Bayesian MEM. The following sections in this chapter describe the unique computational logistics in online calibration, the algorithms and formulations of the three original estimation methods, and Bayesian modifications of the three estimation methods.

6.1 Computational Logistics

Because the online calibration procedures are integrated in an operational CAT that is administered continuously, the computational logistics of online calibration is more complex than those of traditional item calibration. In traditional item calibration, all necessary response data are first collected in full and then a response matrix, which can be sparse sometimes, is prepared. This response matrix is then entered into a calibration program, which outputs the calibrated parameter values for all items of interest. In contrast, online calibration is an ongoing process. More response data accumulate as more examinees take the test sequentially; moreover, each pretest item is taken by a different sample of examinees, and these examinees also take different sets of operational items. To calibrate any pretest item j, data from all examinees who received this item in their tests are needed, which include both their responses to item j and the data for all the operational items they took.

Therefore, as each incoming examinee finishes his/her test, the following data must be stored in the database right away. First, for each pretest item j that the examinee received, assuming item j has already accumulated $N_j - 1$ samples, the following will be stored as the N_jth entry in each variable for item j:

- i: The ID of the examinee;
- x_{ij}: The examinee's response to pretest item j; and
- $\hat{\theta}_i$: The examinee's final ability parameter estimate, if Stocking-A is used for item calibration.

The IDs of all the N_j examinees who took item j will be used to pull necessary information from the operational item database. As each examinee i finishes his/her test, the operational item database will record the following data:

- $\boldsymbol{m}_i$: The IDs of the M_i operational items examinee i received; and
- $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \cdots, y_{iM_i})$: Examinee i's responses to these operational items.

The item parameters of all the operational items in the operational item bank, $\boldsymbol{\beta}_{op}$, are also stored in the database and will be used when a pretest item is calibrated.

If the individual sequential design is used, every time an examinee finishes his/her test, all pretest items administered to him/her will be estimated one by one. For a pretest item j being estimated, all necessary data from the previous $N_j - 1$ samples will be pulled from the database and used in combination with the new data from the current examinee to estimate its parameter values. If the group sequential design is used, every time a pretest item obtains an enough amount of new response data, its parameters are updated. The updated item parameter values will be used in the adaptive selection of pretest items for the incoming examinee.

6.2 Stocking-A: Conditional MLE of Item Parameters

The Stocking-A method (Stocking, 1988) is essentially a conditional MLE procedure. It first estimates examinee ability θ's using all the administered operational items and then it estimates pretest item parameters treating the estimated θ's as known and fixed. Stocking-A is the simplest among the three investigated methods. The following subsections will give the details of the two parts of the Stocking-A procedure — (1) estimating examinee ability parameters using operational items and (2) estimating item parameters conditional on the estimated examinee ability parameters.

6.2.1 Estimation of Examinee Ability Parameters

The examinee ability parameters are inherently estimated in the operational CAT. Once an examinee finishes his/her test, his/her final ability estimate is generated using all operational items he/she received. As described in Section 6.1, the final ability estimates are stored in the database and ready to be pulled at the time of calibration.

As mentioned in Section 2.1, there are three common methods for estimating examinee ability levels: MLE, EAP, and MAP. MLE is the most efficient estimator, but it is not available when the response vector contains all correct or all incorrect responses. EAP is computationally easy and provides reasonable estimates when the response vector contains all correct or all incorrect responses and when the sample size is small. Therefore, this study uses a hybrid method:

- If the number of items is no more than five or the response data are all correct or all incorrect, use EAP;

- Otherwise, use MLE.

EAP of examinee ability The EAP estimator for θ is the expectation of θ based on its posterior distribution given the response data. Let $\pi(\theta)$ denote the prior distribution of θ, which is most commonly standard normal distribution; let $L(\boldsymbol{y}_i, \theta)$ denote the likelihood of θ given examinee i's responses to all administered operational items,

$$\hat{\theta}_{i,EAP} = \int \theta \frac{L(\boldsymbol{y}_i, \theta)\pi(\theta)}{\int L(\boldsymbol{y}_i, \theta)\pi(\theta)\mathrm{d}\theta}\mathrm{d}\theta \,, \tag{6.1}$$

where

$$L(\boldsymbol{y}_i, \theta, \boldsymbol{\beta}_{op}) = \prod_{m=1}^{M_i} [P_m(\theta)]^{y_{im}} [1 - P_m(\theta)]^{1-y_{im}} \,. \tag{6.2}$$

and $P_m(\theta)$'s are computed using the operational item parameters $\boldsymbol{\beta}_{op}$.

MLE of examinee ability The MLE of θ can utilize the Fisher scoring procedure, which is a simplification of the Newton-Raphson algorithm. Let ℓ denote the log-likelihood of θ given the operational item responses: $\ell = \log L(\boldsymbol{y}, \theta)$; $\ell'(\theta)$ denote the first derivative of ℓ with respect to θ, $\ell''(\theta)$ denote the second derivative of the ℓ with respect to θ. A regular Newton-Raphson algorithm for MLE implements the following iterations:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{\ell'(\hat{\theta}_t)}{\ell''(\hat{\theta}_t)} \,. \tag{6.3}$$

The Fisher scoring procedure replaces $\ell''(\hat{\theta}_t)$ with its expectation, which is the negative Fisher information $-I(\hat{\theta}_t)$, so the iteration becomes:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{\ell'(\hat{\theta}_t)}{I(\hat{\theta}_t)} \,. \tag{6.4}$$

For the 3PL model,

$$\begin{aligned}\ell'(\theta) &= \frac{\partial \ell}{\partial \theta} \\ &= \sum_{m=1}^{M_i} (y_{im} - P_m(\theta)) \frac{P'_m(\theta)}{P_m(\theta)[1 - P_m(\theta)]} \\ &= \sum_{m=1}^{M_i} a_m \frac{y_{im} - P_j(\theta)}{P_m(\theta)} \frac{P_m(\theta) - c_m}{1 - c_m} \, ;\end{aligned} \tag{6.5}$$

$$\begin{aligned}I(\theta) &= -E[\frac{\partial^2 \ell}{\partial \theta^2}] \\ &= \sum_{m=1}^{M_i} \frac{P'^2_m(\theta)}{P_m(\theta)[1 - P_m(\theta)]} \\ &= \sum_{m=1}^{M_i} \frac{a_m^2}{(1 - c_m)^2} [P_m(\theta) - c_m]^2 \frac{1 - P_m(\theta)}{P_m(\theta)} \, .\end{aligned} \tag{6.6}$$

The formulas for the 1PL and 2PL models can be reduced from the above.

Because an initial estimate that is far away from the true parameter value may lead to non-convergence in Newton-Raphson iterations, a few iterations of the “Bisection” algorithm can be performed prior to the Fisher scoring procedure to provide a good initial estimate for the latter.

Let θ_{LB} and θ_{UB} denote the preset lower and upper bounds for a θ estimate. The steps of the MLE algorithms implemented in this study are the following.

Phase 1: Bisection Iterations

Step 1: $\theta_L \leftarrow \theta_{LB}$, $\theta_U \leftarrow \theta_{UB}$, $\hat{\theta} \leftarrow (\theta_L + \theta_U)/2$.

Step 2: Compute $\ell'(\theta_L)$, $\ell'(\theta_U)$, and $\ell'(\hat{\theta})$.

Step 3: If $\text{Sign}(\ell'(\theta_L)) = \text{Sign}(\ell'(\hat{\theta}))$, then $\theta_L \leftarrow \hat{\theta}$, $\hat{\theta} \leftarrow (\theta_L + \theta_U)/2$.
If $\text{Sign}(\ell'(\theta_U)) = \text{Sign}(\ell'(\hat{\theta}))$, then $\theta_U \leftarrow \hat{\theta}$, $\hat{\theta} \leftarrow (\theta_L + \theta_U)/2$.

Step 4: Iterate Steps 2 and 3 for a small number of times, for instance twice.

Phase 2: Fisher Scoring Iterations

Step 5: Take the final $\hat{\theta}$ from the Bisection iterations as the initial estimate in the Fisher scoring iterations.

Step 6: Perform the iteration in Equation 6.4 until the step size is less than a small threshold.

6.2.2 CMLE of Item Parameters given Examinee Ability Estimates

As mentioned in Section 6.1, for a pretest item j, the data from all examinees who took item j have been stored in the database and are ready to be pulled for calibrating item j. Let $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_{N_j})$ denote the ability estimates of the N_j examinees who took item j, and let $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \cdots, x_{N_j j})$ denote their responses to item j, respectively. Then the log-likelihood of the item parameter vector $\hat{\boldsymbol{\beta}}_j$ given $\boldsymbol{x}_j$ and $\hat{\boldsymbol{\theta}}_j$ is

$$\ell_j = \ell(\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\beta}}_j) = \sum_{i=1}^{N_j} \ell(x_{ij}, \hat{\theta}_i, \hat{\boldsymbol{\beta}}_j) = \sum_{i=1}^{N_j} x_{ij} \log[P_j(\hat{\theta}_i)](1 - x_{ij}) \log[1 - P_j(\hat{\theta}_i)] \,. \tag{6.7}$$

The CMLE method finds the item parameter estimates $(\hat{a}_j, \hat{b}_j, \hat{c}_j)$ that maximize ℓ_j. Similar to the MLE of θ, the Fisher scoring procedure can be used to implement the method. With certain initial values for $(\hat{a}_j, \hat{b}_j, \hat{c}_j)$, let $\hat{s}_j = \log \hat{a}_j$, the following step is iterated:

$$\begin{bmatrix} \hat{s}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{s}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_t + \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix}_t^{-1} \begin{bmatrix} \ell'_1 \\ \ell'_2 \\ \ell'_3 \end{bmatrix}_t . \tag{6.8}$$

For the 3PL model, the entries in the equation above are given as the follows. Again, the formulas for the 1PL and 2PL models can be reduced from them.

$$\ell_1' = \frac{\partial \log L_j}{\partial \hat{s}_j} = \sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \exp(\hat{s}_j) \frac{\hat{\theta}_i - \hat{b}_j}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \,; \tag{6.9}$$

$$\ell_2' = \frac{\partial \log L_j}{\partial \hat{b}_j} = -\sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \frac{\exp(\hat{s}_j)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \,; \tag{6.10}$$

$$\ell_3' = \frac{\partial \log L_j}{\partial \hat{c}_j} = \sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \frac{1}{P_j(\hat{\theta}_i)} \frac{1}{1 - \hat{c}_j} \,; \tag{6.11}$$

$$I_{11} = \sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 (\hat{\theta}_i - \hat{b}_j)^2 \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \,; \tag{6.12}$$

$$I_{22} = \sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \,; \tag{6.13}$$

$$I_{33} = \sum_{i=1}^{N_j} \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{1}{(1 - \hat{c}_j)^2} \,; \tag{6.14}$$

$$I_{12} = I_{21} = -\sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 (\hat{\theta}_i - \hat{b}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \,; \tag{6.15}$$

$$I_{13} = I_{31} = \sum_{i=1}^{N_j} \exp(\hat{s}_j) (\hat{\theta}_i - \hat{b}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{(1 - \hat{c}_j)^2} \,; \tag{6.16}$$

$$I_{23} = I_{32} = -\sum_{i=1}^{N_j} \exp(\hat{s}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{(1 - \hat{c}_j)^2} \,. \tag{6.17}$$

The iteration terminates when a certain convergence criterion is met, such as the maximum step size among all parameters between two consecutive iterations is less than a small threshold. The final output of $\hat{s}_j$ is then transformed back to $\hat{a}_j$ by $\hat{a}_j = \exp(\hat{s}_j)$.

6.3 OEM and MEM: Marginal MLE with EM Algorithm

The OEM and MEM methods are based on the marginal maximum likelihood estimation (MMLE) with EM algorithm, which integrates out the examinee ability parameter θ using the posterior distribution of θ obtained from the response data and finds the item parameter estimates that maximize the expectation of the log-likelihood with respect to the posterior θ distribution.

For a pretest item j, the calibration using OEM or MEM will use the following data from each examinee i who took item j:

- $\boldsymbol{m}_i$: The IDs of the M_i operational items examinee i received;
- $\boldsymbol{y}_i$: Examinee i's responses to all these operational items; and
- x_{ij}: Examinee i's response to item j.

The operational item parameters, $\boldsymbol{\beta}_{op}$, will also be used in the estimation procedure.

6.3.1 OEM

OEM has only one EM cycle.

- The *E-step* finds the posterior expectation of the log-likelihood of the item being estimated. This posterior expectation is based on the posterior ability distribution of every examinee $i = 1, 2, \cdots, N_j$ who took item j, which is obtained from their responses to the administered operational items, $\boldsymbol{y}_i$, and the known operational item parameters, $\boldsymbol{\beta}_{op}$.
- The *M-step* finds the item parameter vector $\boldsymbol{\beta} = (a, b, c)$ that maximizes the posterior expection of the log-likelihood.

These two steps are implemented through numerical procedures. Let θ_k denote q quadrature points on the θ-scale and $A(\theta_k)$ be their corresponding weights. The weights can be the standard normal probability density function, or Gaussian-Hermite quadrature points and weights can be used. First, the following two quantities are computed for each quadrature point $k = 1, 2, \cdots, q$.

$$\bar{f}_{jk} = \sum_{i=1}^{N_j} \left[\frac{L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}) A(\theta_k)}{\sum_{k=1}^{q} L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}) A(\theta_k)} \right] ; \tag{6.18}$$

$$\bar{r}_{jk} = \sum_{i=1}^{N_j} \left[\frac{x_{ij} L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}) A(\theta_k)}{\sum_{k=1}^{q} L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}) A(\theta_k)} \right] , \tag{6.19}$$

where

$$L(\boldsymbol{y}_i, \theta, \boldsymbol{\beta}_{op}) = \prod_{m=1}^{M_j} [P_m(\theta)]^{y_{ij}} [1 - P_m(\theta)]^{1-y_{ij}} . \tag{6.20}$$

Then, with certain initial values for $(\hat{a}_j, \hat{b}_j, \hat{c}_j)$, let $\hat{s}_j = \log \hat{a}_j$, the following step is iterated until a certain convergence criterion is met, such as the maximum step size among the three parameters between two consecutive iterations is less than a small threshold.

$$\begin{bmatrix} \hat{s}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{s}_j \\ \hat{b}_j \\ \hat{c}_j \end{bmatrix}_t + \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix}_t^{-1} \begin{bmatrix} \ell'_1 \\ \ell'_2 \\ \ell'_3 \end{bmatrix}_t . \tag{6.21}$$

For the 3PL model, the entries in the equation above are given as the follows. Again, the formulas for the 1PL and 2PL models can be reduced from them.

$$\ell'_1 = \sum_{k=1}^{q} [\bar{r}_k - \bar{f}_k P_j(\theta_k)] \exp(\hat{s}_j) \frac{\theta_k - \hat{b}_j}{P_j(\theta_k)} \frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} ; \tag{6.22}$$

$$\ell'_2 = -\sum_{k=1}^{q} [\bar{r}_{jk} - \bar{f}_{jk} P_j(\theta_k)] \frac{\exp(\hat{s}_j)}{P_j(\theta_k)} \frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} ; \tag{6.23}$$

$$\ell_3' = \sum_{k=1}^{q} [\bar{r}_{jk} - \bar{f}_{jk} P_j(\theta_k)] \frac{1}{P_j(\theta_k)} \frac{1}{1 - \hat{c}_j} \, ; \tag{6.24}$$

$$I_{11} = \sum_{k=1}^{q} \bar{f}_{jk} [\exp(\hat{s}_j)]^2 (\theta_k - \hat{b}_j)^2 \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \, ; \tag{6.25}$$

$$I_{22} = \sum_{k=1}^{q} \bar{f}_{jk} [\exp(\hat{s}_j)]^2 \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \, ; \tag{6.26}$$

$$I_{33} = \sum_{k=1}^{q} \bar{f}_{jk} \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \frac{1}{(1 - \hat{c}_j)^2} \, ; \tag{6.27}$$

$$I_{12} = I_{21} = -\sum_{k=1}^{q} \bar{f}_{jk} [\exp(\hat{s}_j)]^2 (\theta_k - \hat{b}_j) \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 \tag{6.28}$$

$$I_{13} = I_{31} = \sum_{k=1}^{q} \bar{f}_{jk} \exp(\hat{s}_j) (\theta_k - \hat{b}_j) \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \frac{P_j(\theta_k) - \hat{c}_j}{(1 - \hat{c}_j)^2} \, ; \tag{6.29}$$

$$I_{23} = I_{32} = -\sum_{k=1}^{q} \bar{f}_{jk} \exp(\hat{s}_j) \frac{1 - P_j(\theta_k)}{P_j(\theta_k)} \frac{P_j(\theta_k) - \hat{c}_j}{(1 - \hat{c}_j)^2} \, . \tag{6.30}$$

The final output of $\hat{s}_j$ is then transformed back to $\hat{a}_j$ by $\hat{a}_j = \exp(\hat{s}_j)$.

6.3.2 MEM

The MEM method allows multiple EM cycles. The first cycle is the same as OEM; starting with the second cycle, the data for both the operational items and the pretest item are used to update the posterior ability distribution in the E-step. The E-step and the M-step iterate until a certain convergence criterion is met.

The formulas for the first EM cycle in MEM are given in the previous section. Starting with the second cycle, the the posterior ability distribution of each examinee $i = 1, 2, \cdots, N_j$ who took item j is computed using his/her responses to the administered operational items, $\boldsymbol{y}_i$, his/her response to the pretest item, x_{ij}, the known operational item parameter values, $\boldsymbol{\beta}_{op}$, and the estimated parameter values, $\hat{\boldsymbol{\beta}}_j$.

In the actual implementation, the first EM cycle of MEM is carried out using Equations 6.18 – 6.30. In each following EM cycle, first the following two quantities are computed for each quadrature point $k = 1, 2, \cdots, q$.

$$\bar{f}_{jk} = \sum_{i=1}^{N_j} \left[\frac{L(\boldsymbol{y}_i, x_{ij}, \theta_k, \boldsymbol{\beta}_{op}, \hat{\boldsymbol{\beta}}_j) A(\theta_k)}{\sum_{k=1}^{q} L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}, \hat{\boldsymbol{\beta}}_j) A(\theta_k)} \right] ; \tag{6.31}$$

$$\bar{r}_{jk} = \sum_{i=1}^{N_j} \left[\frac{x_{ij} L(\boldsymbol{y}_i, x_{ij}, \theta_k, \boldsymbol{\beta}_{op}, \hat{\boldsymbol{\beta}}_j) A(\theta_k)}{\sum_{k=1}^{q} L(\boldsymbol{y}_i, \theta_k, \boldsymbol{\beta}_{op}, \hat{\boldsymbol{\beta}}_j) A(\theta_k)} \right] , \tag{6.32}$$

where

$$L(\boldsymbol{y}_i, x_{ij}, \theta, \boldsymbol{\beta}_{op}, \hat{\boldsymbol{\beta}}_j) = [P_j(\theta)]^{x_{ij}} [1 - P_j(\theta)]^{1-x_{ij}} \prod_{m=1}^{M_j} [P_m(\theta)]^{y_{ij}} [1 - P_m(\theta)]^{1-y_{ij}} . \tag{6.33}$$

Then the iterations described in Equations 6.8 – 6.30 are carried out. The output item parameter values from these iterations are then plugged into Equations 6.31 and 6.32 in the next EM cycle. The EM cycle terminates until a certain convergence criterion is met, such as the maximum absolute change in the item parameters between two consecutive EM cycles are less than a small threshold.

6.4 The Bayesian priors on item parameters

Estimation procedures that are based on the Newton-Raphson algorithm are prone to non-convergence problems. Possible causes of non-convergence include starting in a “bad” neighborhood, existence of local maxima, the sample size is relatively small, etc. Having small sample sizes is typical in online calibration, especially at the beginning stage of calibration. To alleviate this problem, Bayesian priors on item parameters can be added into the estimation procedure. A common choice of the Bayesian priors for the 3PL model item parameters is

a-parameter:

$$s = \log(a) \sim N(\mu_s, \sigma_s)\,; \tag{6.34}$$

b-parameter:

$$b \sim N(\mu_b, \sigma_b)\,; \tag{6.35}$$

c-parameter:

$$c \sim \mathrm{Beta}(\alpha_c, \beta_c)\,. \tag{6.36}$$

While previous studies (e.g.. Ban et al., 2001) used certain fixed Bayesian priors, in the online calibration setting, a new solution for finding the parameter values for the Bayesian priors is to fit the corresponding distributions to the operational item parameter values. This solution may be more informative than an arbitrary choice, and less subject to the self-validating problem in finding "empirical priors" from the items being calibrated.

Essentially, the change in the estimation algorithms brought by the Bayesian component is that a Bayesian prior term $\log g(\boldsymbol{\beta})$ is added to the log-likelihood equations in each estimation method. The details of the Bayesian modifications of the three estimation methods are given next (Baker & Kim, 2004).

6.5 Bayesian Stocking-A

When Bayesian priors are incorporated in Stocking-A, Equation 6.7 becomes

$$\ell_j = \sum_{i=1}^{N_j} x_{ij} \log[P_j(\hat{\theta}_i)](1 - x_{ij}) \log[1 - P_j(\hat{\theta}_i)] + \log g(\boldsymbol{\beta})\,. \tag{6.37}$$

Equations 6.9 – 6.17 become

$$\ell_1' = \sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \exp(\hat{s}_j) \frac{\hat{\theta}_i - \hat{b}_j}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} - \frac{\hat{s}_j - \mu_s}{\sigma_s^2}\,; \tag{6.38}$$

$$\ell_2' = -\sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \frac{\exp(\hat{s}_j)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} - \frac{\hat{b}_j - \mu_b}{\sigma_b^2}\,; \tag{6.39}$$

$$\ell_3' = \sum_{i=1}^{N_j} [x_{ij} - P_j(\hat{\theta}_i)] \frac{1}{P_j(\hat{\theta}_i)} \frac{1}{1 - \hat{c}_j} + \frac{\alpha_c - 1}{\hat{c}_j} - \frac{\beta_c - 1}{1 - \hat{c}_j}\,; \tag{6.40}$$

$$I_{11} = \sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 (\hat{\theta}_i - \hat{b}_j)^2 \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 + \frac{1}{\sigma_s^2}\,; \tag{6.41}$$

$$I_{22} = \sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2 + \frac{1}{\sigma_b^2}\,; \tag{6.42}$$

$$I_{33} = \sum_{i=1}^{N_j} \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{1}{(1 - \hat{c}_j)^2} + \frac{\alpha_c - 1}{\hat{c}_j^2} + \frac{\beta_c - 1}{1 - \hat{c}_j^2}\,; \tag{6.43}$$

$$I_{12} = I_{21} = -\sum_{i=1}^{N_j} [\exp(\hat{s}_j)]^2 (\hat{\theta}_i - \hat{b}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \left[\frac{P_j(\hat{\theta}_i) - \hat{c}_j}{1 - \hat{c}_j} \right]^2\,; \tag{6.44}$$

$$I_{13} = I_{31} = \sum_{i=1}^{N_j} \exp(\hat{s}_j)(\hat{\theta}_i - \hat{b}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{(1 - \hat{c}_j)^2}\,; \tag{6.45}$$

$$I_{23} = I_{32} = -\sum_{i=1}^{N_j} \exp(\hat{s}_j) \frac{1 - P_j(\hat{\theta}_i)}{P_j(\hat{\theta}_i)} \frac{P_j(\hat{\theta}_i) - \hat{c}_j}{(1 - \hat{c}_j)^2}\,. \tag{6.46}$$

6.6 Bayesian OEM and Bayesian MEM

For both OEM and MEM, when Bayesian priors are incorporated, Equations 6.22 – 6.30 become

$$\ell_1' = \sum_{k=1}^{q} [\bar{r}_k - \bar{f}_k P_j(\theta_k)] \exp(\hat{s}_j) \frac{\theta_k - \hat{b}_j}{P_j(\theta_k)} \frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} - \frac{\hat{s}_j - \mu_s}{\sigma_s^2}\,; \tag{6.47}$$

$$\ell_2' = -\sum_{k=1}^{q}[\bar{r}_{jk} - \bar{f}_{jk}P_j(\theta_k)]\frac{\exp(\hat{s}_j)}{P_j(\theta_k)}\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j} - \frac{\hat{b}_j - \mu_b}{\sigma_b^2}\,; \tag{6.48}$$

$$\ell_3' = \sum_{k=1}^{q}[\bar{r}_{jk} - \bar{f}_{jk}P_j(\theta_k)]\frac{1}{P_j(\theta_k)}\frac{1}{1 - \hat{c}_j} + \frac{\alpha_c - 1}{\hat{c}_j} - \frac{\beta_c - 1}{1 - \hat{c}_j}\,; \tag{6.49}$$

$$I_{11} = \sum_{k=1}^{q}\bar{f}_{jk}[\exp(\hat{s}_j)]^2(\theta_k - \hat{b}_j)^2\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j}\right]^2 + \frac{1}{\sigma_s^2}\,; \tag{6.50}$$

$$I_{22} = \sum_{k=1}^{q}\bar{f}_{jk}[\exp(\hat{s}_j)]^2\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j}\right]^2 + \frac{1}{\sigma_b^2}\,; \tag{6.51}$$

$$I_{33} = \sum_{k=1}^{q}\bar{f}_{jk}\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\frac{1}{(1 - \hat{c}_j)^2} + \frac{\alpha_c - 1}{\hat{c}_j^2} + \frac{\beta_c - 1}{1 - \hat{c}_j^2}\,; \tag{6.52}$$

$$I_{12} = I_{21} = -\sum_{k=1}^{q}\bar{f}_{jk}[\exp(\hat{s}_j)]^2(\theta_k - \hat{b}_j)\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\left[\frac{P_j(\theta_k) - \hat{c}_j}{1 - \hat{c}_j}\right]^2\,; \tag{6.53}$$

$$I_{13} = I_{31} = \sum_{k=1}^{q}\bar{f}_{jk}\exp(\hat{s}_j)(\theta_k - \hat{b}_j)\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\frac{P_j(\theta_k) - \hat{c}_j}{(1 - \hat{c}_j)^2}\,; \tag{6.54}$$

$$I_{23} = I_{32} = -\sum_{k=1}^{q}\bar{f}_{jk}\exp(\hat{s}_j)\frac{1 - P_j(\theta_k)}{P_j(\theta_k)}\frac{P_j(\theta_k) - \hat{c}_j}{(1 - \hat{c}_j)^2}\,. \tag{6.55}$$

Chapter 7

Simulation Study

7.1 Design

A simulation study was conducted using a Fortran program written by the author to investigate the effects of different pretest item selection methods, statistical estimation methods, and seeding locations. The comparison was conducted under the 1PL, 2PL, and 3PL models.

7.1.1 Compared Pretest Item Selection Methods

Five pretest item selection methods were compared: (1) OIRPI with order statistics (OIRPI-O), (2) OIRPI with standardization (OIRPI-S), (3) direct comparison of D-optimal values, (4) examinee-centered selection, and (5) random selection. The random selection condition, which is the easiest method and also the best possible solution in conventional paper-and-pencil tests, provides a baseline for the comparison.

OIRPI In both OIRPI-O and OIRPI-S, the θ scale was divided into 13 ranges. The division is based on percentiles. Let P_t denote the t^{th} percentile of the standard normal distribution, the 13 ranges are $P_0 \sim P_{1/13\times100}$, $P_{1/13\times100} \sim P_{2/13\times100}$, $\cdots$, $P_{12/13\times100} \sim P_{100}$, and the representative values are $P_{1/26\times100}, P_{3/26\times100}, \cdots, P_{25/26\times100}$. A simulation study was conducted to compare this with finer divided ranges (i.e., 61 ranges). No significant difference was detected. So 13 ranges was chosen for better computational efficiency.

Direct comparison of D-optimal values The direct comparison of D-optimal value method selects the pretest item that produces the maximum D-optimal value. The D-optimal value of a pretest item is computed using both the θ estimate of the current examinee and the θ estimates of all past examinees who took this item.

Examinee-centered selection For the examinee-centered method, because the classical maximum Fisher information method (i.e., selecting the item that provides the maximum Fisher information for estimating examinee ability θ) was used as the operational item selection method in this study, the examinee-centered method used the same method to select pretest items.

7.1.2 Compared Statistical Estimation Methods

The six statistical estimation methods compared in this study are Stocking-A, OEM, MEM, Bayesian Stocking-A, Bayesian OEM, and Bayesian MEM. The Bayesian prior distributions of the item parameters used in the last three methods are given in Equations 6.34 – 6.36.

To mimic a real situation where the Bayesian priors of the item parameters may not be known, and to better utilize the known information from the operational items, the Bayesian prior parameters in this study were obtained by fitting the lognormal, normal, and beta distributions to the operational *a*-, *b*-, and *c*-parameters, respectively. This solution may be more informative than an arbitrary choice, and less subject to the self-validating problem in finding "empirical priors" from the items being calibrated.

Convergence is defined as the largest absolute change in all parameters being no greater than a small critical value. For (Bayesian) Stocking-A and (Bayesian) OEM, the convergence thresholds for the Fisher scoring iterations is 0.001, and the maximum number of iterations is 100. For (Bayesian) MEM, a "pseudo-EM" procedure (Cai, personal communication) is used: each EM cycle always executes five Fisher scoring iterations; the convergence thresholds for the EM cycles is 0.001, and the maximum number of EM cycles is 100. This pseudo-EM procedure is recommended when the model is complex and no simple analytical solution is available.

7.1.3 Compared Seeding Locations

The test length is 40, and five pretest items are randomly seeded in the following three conditions for the seeding range:

1. Early in the test: randomly seeded among items 1 through 13;
2. In the middle of the test: randomly seeded among items 14 through 26; and
3. Late in the test: items 27 through 40.

These three conditions for seeding locations are expected to generate different results, because the θ estimates, which are used in selecting pretest items, are of different levels of accuracy at different stages of the test. Thus, in this study, the effect of the seeding location as revealed from the entire dynamic system of online calibration was empirically investigated. Besides, randomness was incorporated in the seeding design because fixed seeding may lead to differentiated motivation and consequently item parameter drift from pretest to operational tests, if the seeding locations are known by examinees.

7.1.4 Test Specifications

The simulation was replicated for 100 times. In each replication, 30 pretest items and 300 operational items were randomly generated from the same distributions. The distributions were chosen to mimic realistic situations. Specifically,

$$\begin{bmatrix} \log(a) \\ b \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0.4 \\ 0.0 \end{bmatrix}, \begin{bmatrix} 0.10 & 0.25 \\ 0.25 & 1.00 \end{bmatrix}\right), \tag{7.1}$$

and

$$c \sim \mathrm{Beta}(4, 16)\,, \tag{7.2}$$

where the c-parameters are independent from the a- and b-parameters.

Examinee ability θ's were generated from the standard normal distribution. In the simulation, examinees take the CAT test sequentially. The operational items are selected from the operational item bank and the examinee ability parameter is updated after each operational item is administered. As described in Section 6.2.1, examinee abilities were estimated by the EAP method when the number of administered operational items is no more than five or the responses are all correct or all incorrect; otherwise, MLE was used.

When an examinee reaches the seeding locations, pretest items are selected from the pretest item bank. Because the parameter estimation is highly unstable when the sample size is too small, within an average of the first 100 responses that each pretest item receives, which means for the first about $100 \times 30/5 = 600$ examinees, pretest items are selected randomly and the item parameters are not updated. After that, different adaptive pretest item selection methods are used, and the item parameters of each pretest items are updated after it obtains every 10 new samples. The pretest item selection algorithms are adjusted once the item parameters are updated.

The termination of the sampling for each pretest item was determined by sample size. Once a pretest item receives a total of 500 examinee responses, it is exported from the pretest item bank. By holding the sample sizes constant, the performance of different pretest item selection methods and estimation methods can be compared by comparing the accuracy of the estimated parameters. The more accurate the parameter estimates are, the more efficient the methods are.

7.1.5 Evaluation Criteria

The performance of the compared methods are evaluated through two criteria. The first criterion focuses on the accuracy of the individual item parameter estimates. Specifically, the RMSEs of the estimates of each item parameter, formulated by Equation 7.3 are evaluated.

$$RMSE_p = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\beta}_p - \beta_p)^2}\,, \tag{7.3}$$

where p denotes the specific element in the item parameter vector, such as the a-parameter, b-parameter, and c-parameter, and $j = 1, 2, \cdots, J$ denotes the J pretest items in one replication ($J = 30$ in this study) .

The second criterion focuses on the overall recovery of item parameter vectors. The chosen criterion is the average weighted area difference between the true *item characteristic curve* (ICC) and the estimated ICC. The area difference is computed by numeric integration using the following formula:

$$AD = \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{q}\left|P_j(\theta_k) - \hat{P}_j(\theta_k)\right|g(\theta_k)\,, \tag{7.4}$$

where θ_k's are quadrature points, and the weighting function $g(\theta_k)$ is the density of the standard normal distribution. This weighting strategy reflects the overall effect of the item parameter recovery on the entire normally distributed population.

7.2 Results

7.2.1 Comparing Estimation Methods

First, the six statistical estimation methods are compared. In each crossed condition, the RMSE values of an item parameter are averaged across the 100 replications. These averaged RMSEs are presented by Figures 7.1 – 7.6. Each figure is for one of the item parameters in one model, and the subfigures are for different seeding locations. In each figure, the horizontal axis represents the five pretest item selection methods, where "Examinee" stands for the examinee-centered selection and "D-optimal" stands for the direct comparison of the D-optimal values. The six estimation methods are grouped together for each pretest item selection method, ordered by Stocking-A, OEM, MEM, Bayesian Stocking-A, Bayesian OEM, and Bayesian MEM. The estimation methods are also distinguished by different colors.

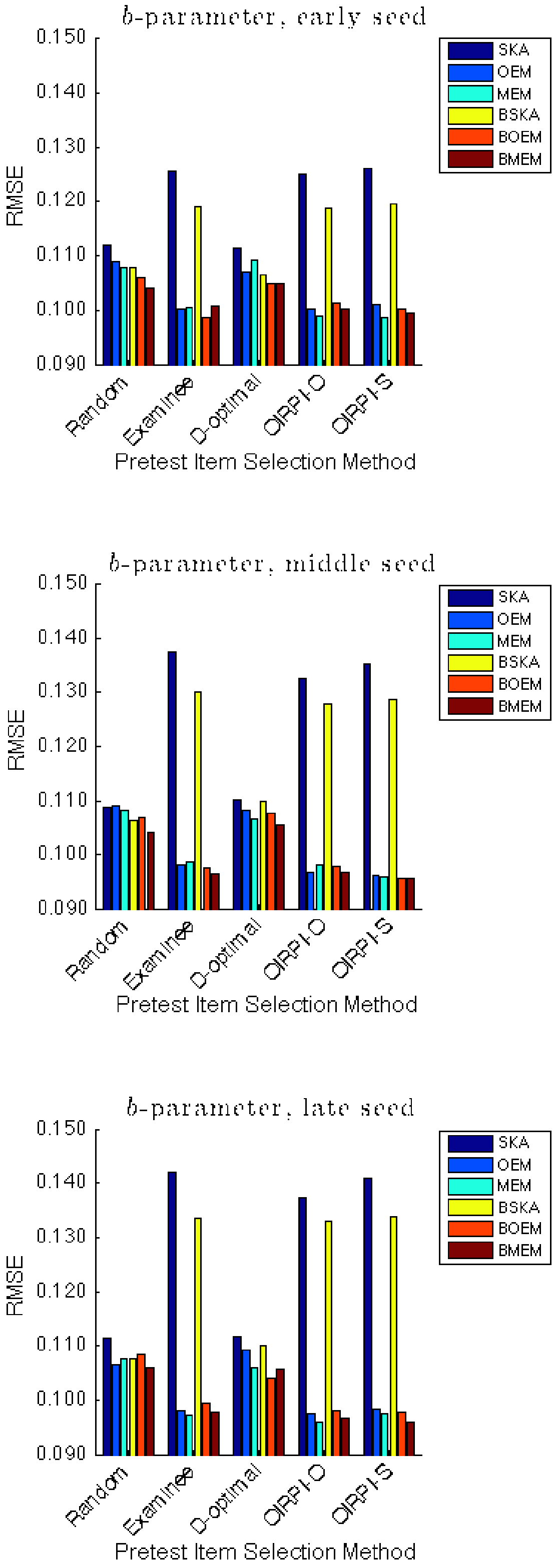
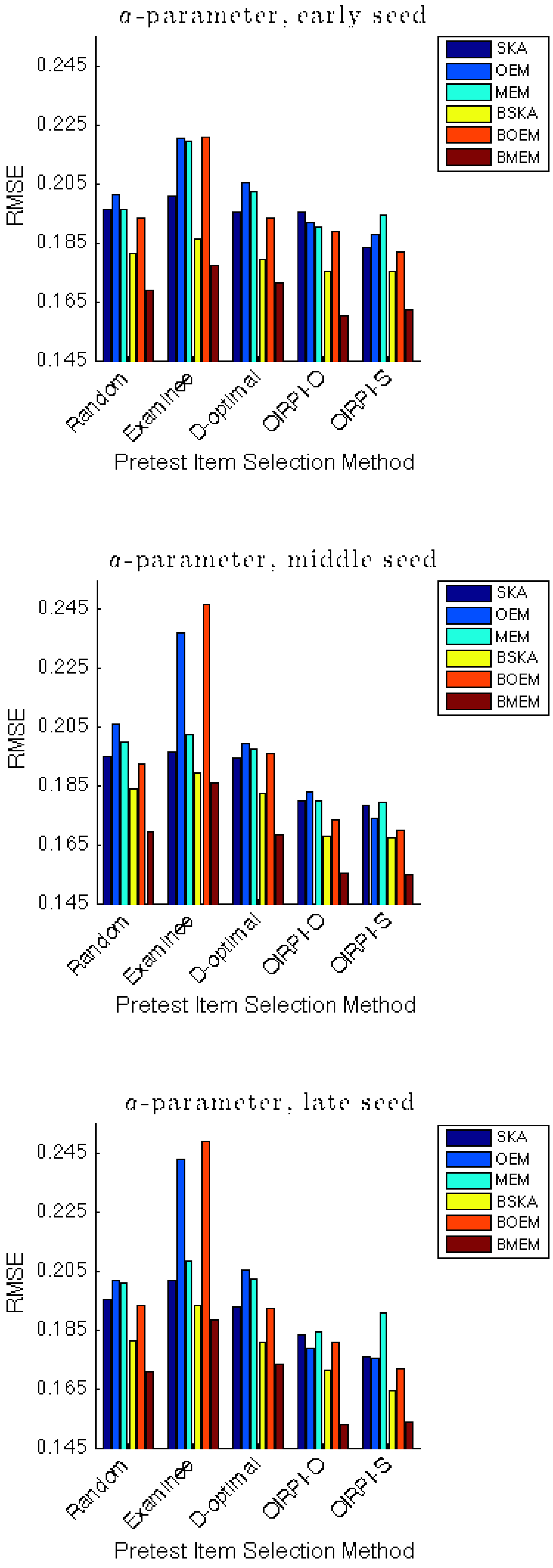


Figure 7.1: The RMSE of the b-parameter estimates under the 1PL model: comparing estimation methods.



a-parameter, early seed

a-parameter, middle seed

a-parameter, late seed

Figure 7.2: The RMSE of the a-parameter estimates under the 2PL model: comparing estimation methods.

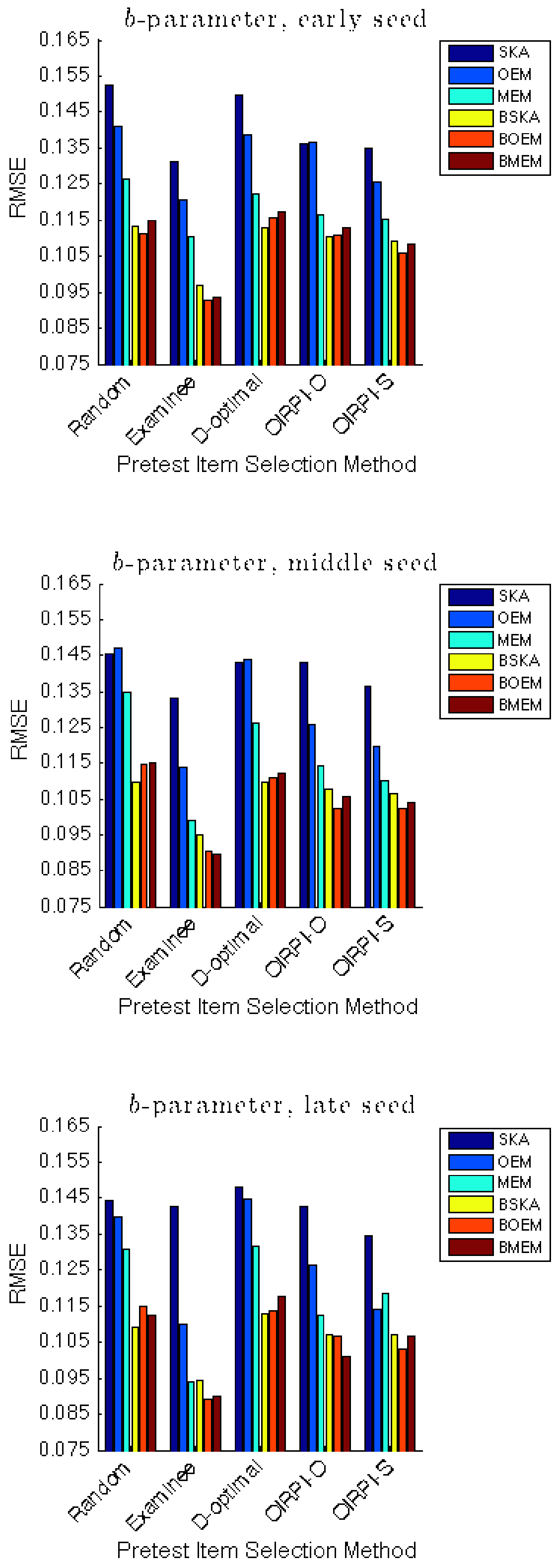
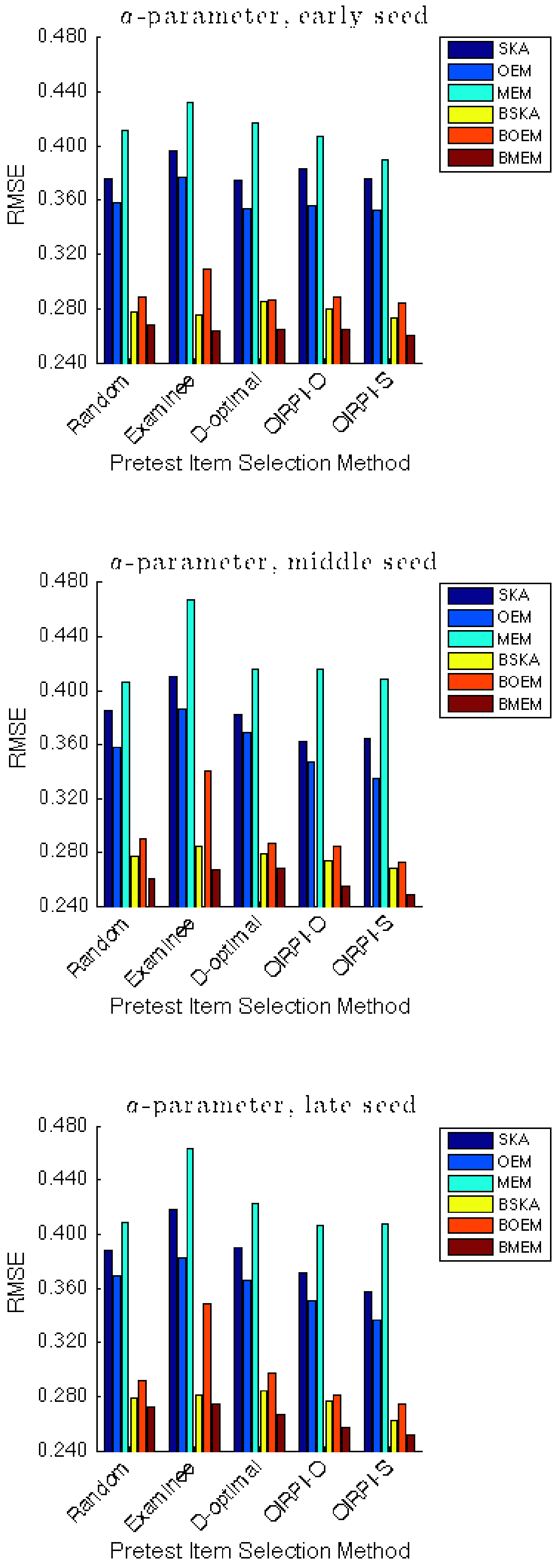


Figure 7.3: The RMSE of the b-parameter estimates under the 2PL model: comparing estimation methods.



a-parameter, early seed

a-parameter, middle seed

a-parameter, late seed

Figure 7.4: The RMSE of the a-parameter estimates under the 3PL model: comparing estimation methods.

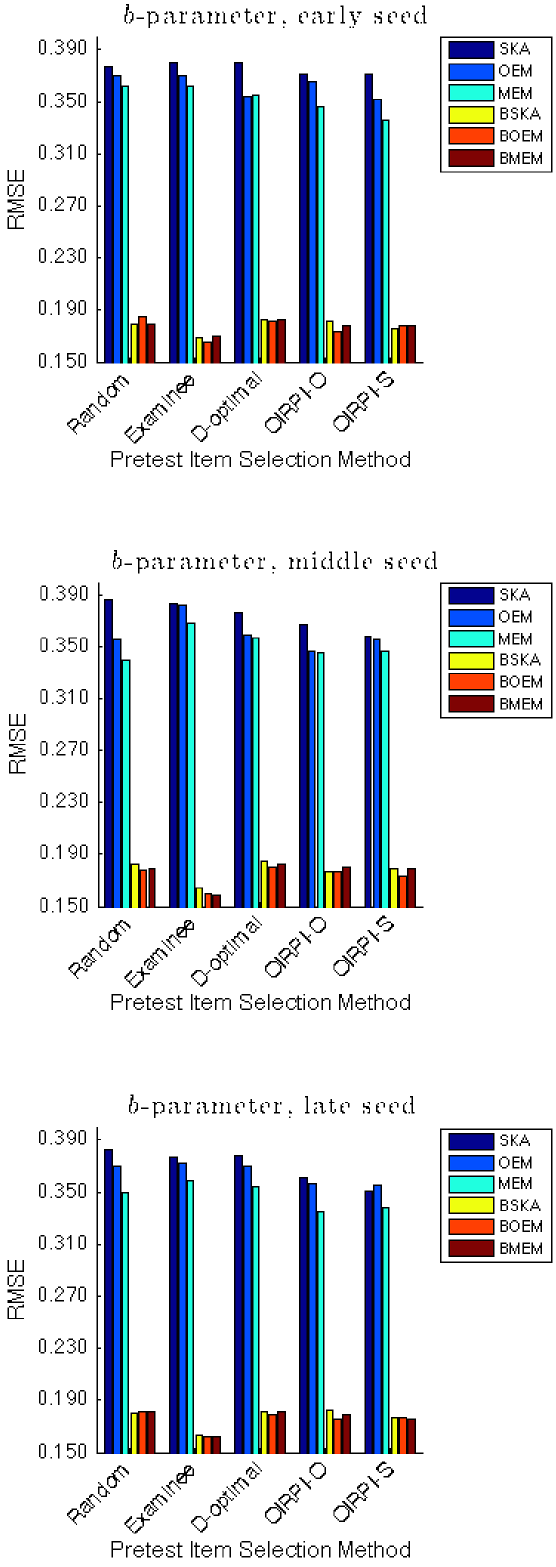
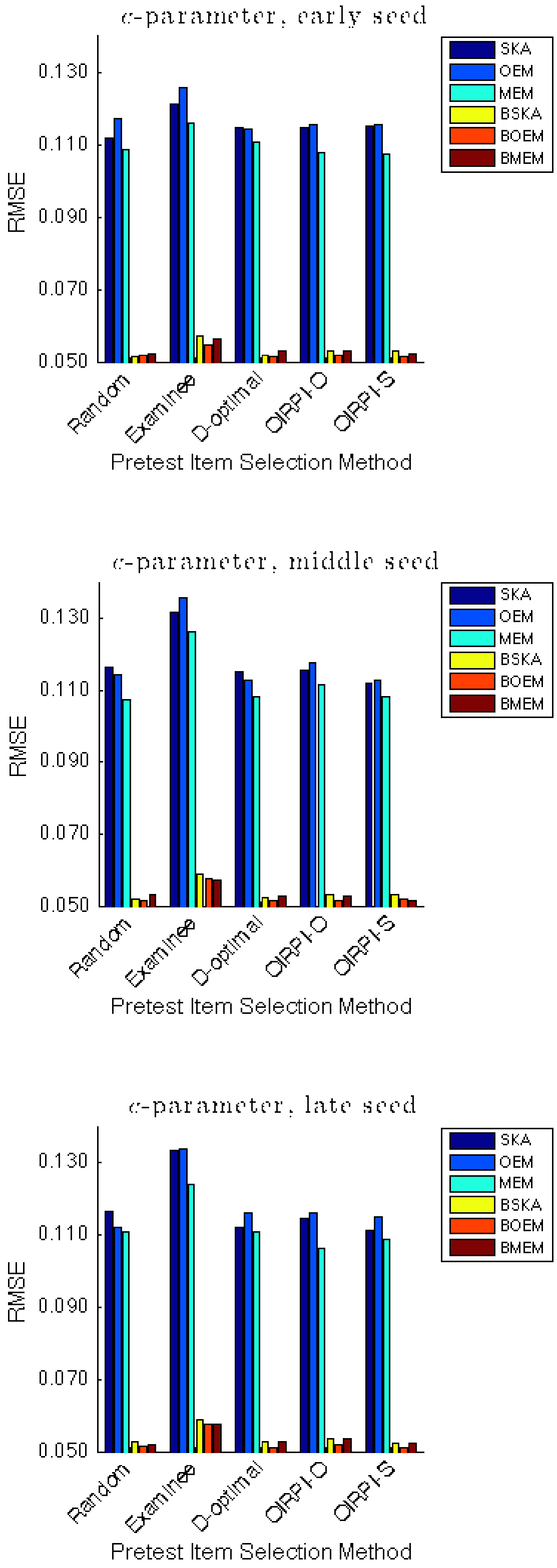


Figure 7.5: The RMSE of the b-parameter estimates under the 3PL model: comparing estimation methods.



c-parameter, early seed

c-parameter, middle seed

c-parameter, late seed

Figure 7.6: The RMSE of the c-parameter estimates under the 3PL model: comparing estimation methods.

Results show that in different models, parameters, seeding locations, and pretest item selection methods, the patterns of the RMSEs produced by the six estimation methods are not the same. For the b-parameter under the 1PL model, in the "D-optimal" method and the two OIRPI methods, the Stocking-A and the Bayesian Stocking-A methods produced significantly larger RMSEs than the other four estimation methods. For the 2PL model, in the examinee-centered method, the a-parameter RMSEs from the OEM and Bayesian OEM are significantly larger than the others. In the b-parameter of the 2PL model and all three parameters of the 3PL model, the three non-Bayesian methods are significantly less accurate than the three Bayesian methods.

Nevertheless, one thing consistent across all conditions is that the Bayesian MEM method appears to be the most stable method among the six and in most cases generated the smallest RMSEs. Therefore, the Bayesian MEM method is recommended in general, and all following analyses are conducted using the data generated from the Bayesian MEM method only.

Note that Ban et al. (2001) comment that OEM is faster than MEM because OEM has only one EM cycle and MEM iterates through many EM cycles until the convergence criterion is met. However, it was found in this simulation study that the MEM using the pseudo EM algorithm, as explained in Section 7.1.2, consumes a similar amount of computation time as the OEM method. This result demonstrates that the pseudo EM algorithm is effective. When the pseudo EM algorithm is used, MEM is not only more accurate than OEM but also comparable in terms of speed.

7.2.2 Pretest Item Selection Methods and Seeding Locations

Figures 7.7 – 7.10 present the comparison of the three seeding locations and five pretest item selection methods. Again, in all figures, the label "D-optimal" denotes the direct comparison of D-optimal values, "Examinee" denotes the examinee-centered item selection method, "OIRPI-O" denotes the OIRPI method with order statistics, and "OIRPI-S" denotes the OIRPI method with standardization. For both the RMSE and the weighted ICC area difference, a smaller value means the estimation is more accurate.

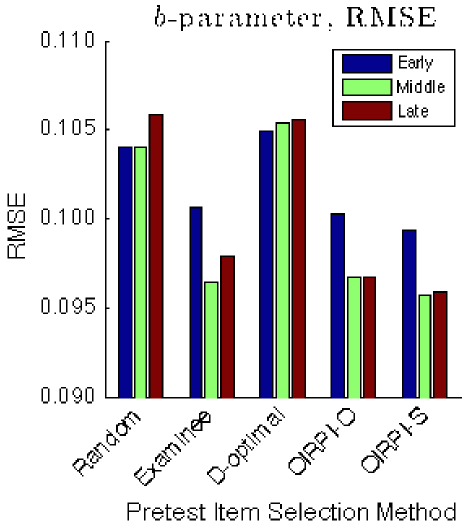
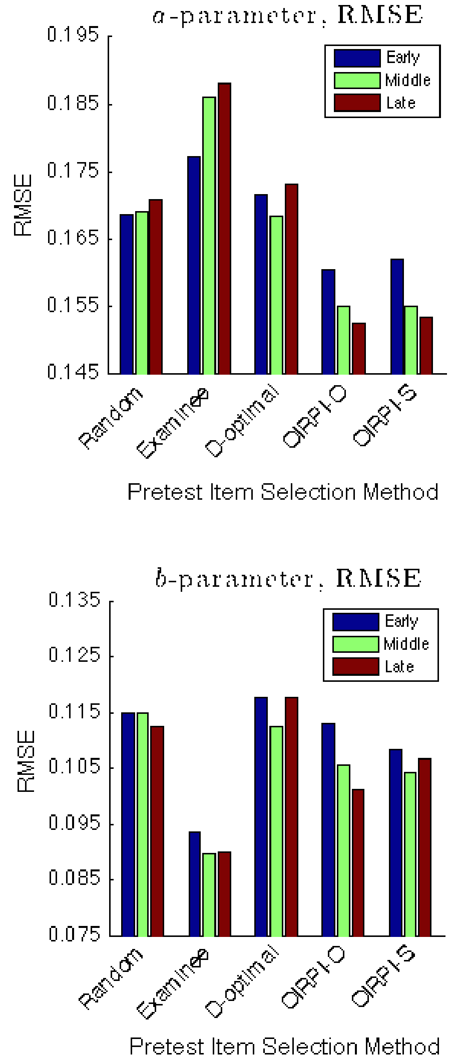


Figure 7.7: The RMSE of item parameter estimates under the 1PL model: comparing pretest item selection methods and seeding locations.

For the 1PL model, the b-parameter is the only item parameter in the model. Figure 7.7 shows that the examinee-centered selection and the two OIRPI methods generated the smallest RMSE values, which means the most efficient calibration. These three methods are more efficient than both random selection and the direct comparison of the D-optimal criterion values. Note that theoretically, the examinee-centered selection is expected to generate superior results, as explained earlier in this paper. The equivalent performance of the two OIRPI methods as the examinee-centered selection from the simulation results demonstrates the effectiveness of the OIRPI methods.

Comparing the three conditions for seeding locations, a slight decreasing trend is observed in the examinee-centered selection and the two OIRPI methods. The "middle of the test" and "late in the test" conditions generated mostly equivalent results. This could indicate the correctness of the mechanism of these three methods. As the seeding location moves towards the end of the test, the θ estimate that is used to select the pretest items becomes more accurate; with a more accurate input parameter value, a correct mechanism should also generate more accurate output, which is reflected by a decreasing trend. In contrast, an increasing trend may indicate that the mechanism is flawed in this setting.



a-parameter, RMSE

b-parameter, RMSE

Figure 7.8: The RMSE of item parameter estimates under the 2PL model: comparing pretest item selection methods and seeding locations.

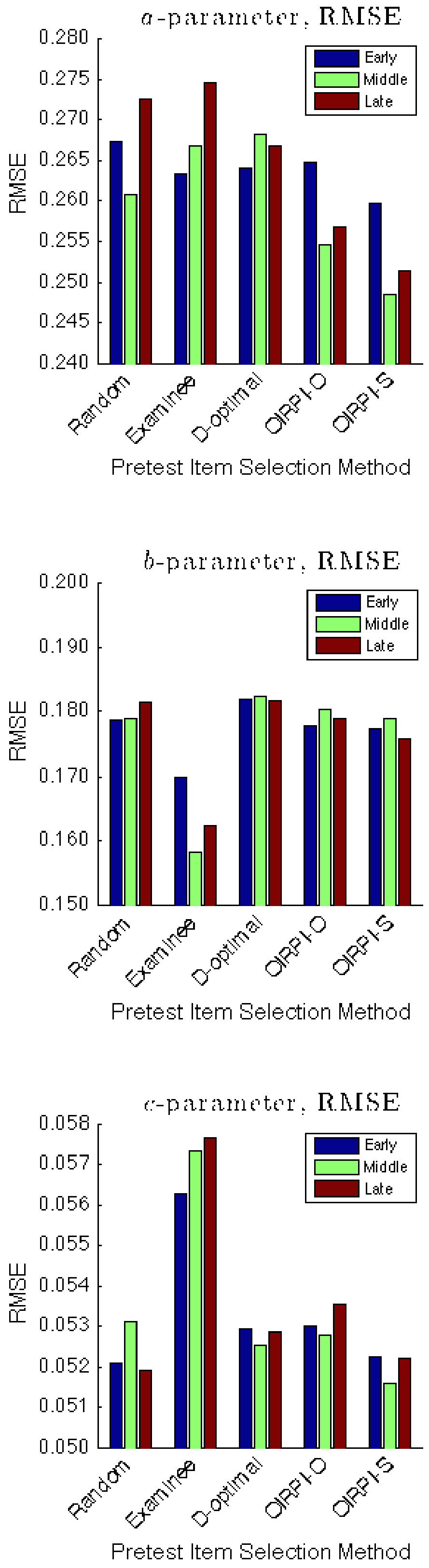


Figure 7.9: The RMSE of item parameter estimates under the 3PL model: comparing pretest item selection methods and seeding locations.

As Figures 7.8 and 7.9 show, for both the 2PL and 3PL models, the examinee-centered selection generated the most accurate b-parameter estimates, but least accurate a-parameter estimates (and c-parameter estimates in the 3PL model). This confirms the earlier theoretical reasoning about the issue in the examinee-centered selection. Besides the examinee-centered selection, the two OIRPI methods generated slightly more accurate a- and b-parameter estimates than random selection and the direct comparison of the D-optimal criterion values. All four methods except the examinee-centered selection have similar levels of recovery of the c-parameters.

Again, when comparing the seeding locations, a generally decreasing trend is seen in the two OIRPI methods in both parameters of the 2PL model and the a-parameter in the 3PL model, but not in random selection and the direct comparison of the D-optimal criterion values. Moreover, a clear decreasing trend in the b-parameter and an increasing trend is seen in the a- and $c-$parameters for the examinee-centered selection, which verifies the flawed mechanism of the examinee-centered selection in the 2PL and 3PL models.

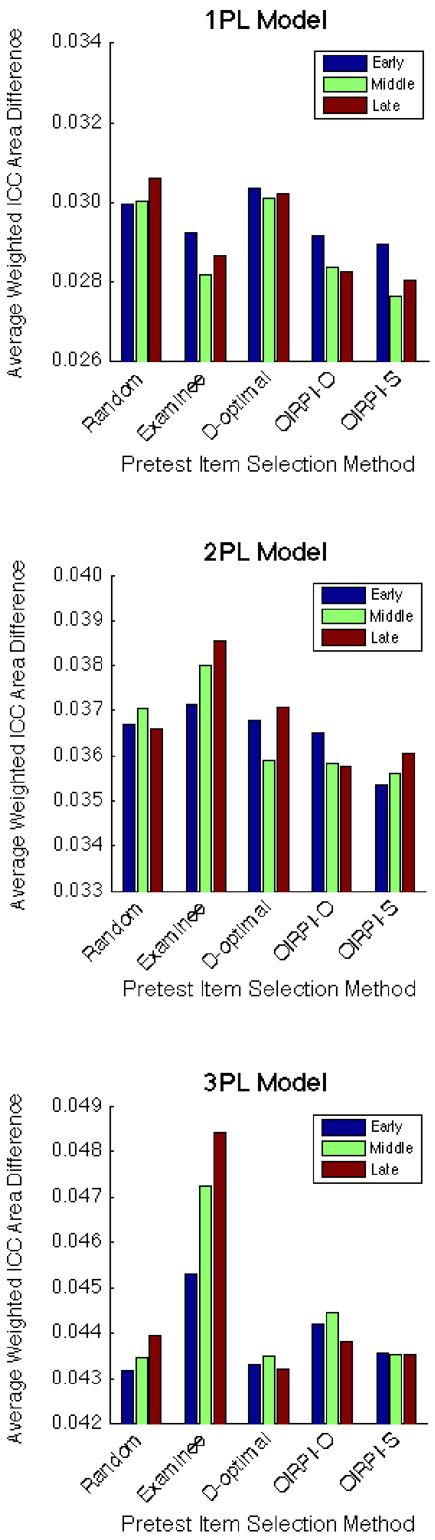


Figure 7.10: The average weighted ICC area differences: comparing pretest item selection methods and seeding locations.

While the RMSE values reflect the accuracy of recovering individual item parameters, the weighted ICC area differences (Figure 7.10) reflect the overall effect on examinee scoring. The weighted ICC area difference is more practically important, because for an operational test, while the accuracy of individual item parameters is still an intermediate result, what eventually matters is the accuracy of examinee ability estimation, and that is reflected in the accuracy of the ICCs. Figure 7.10 shows that (1) under the 1PL model, the two OIRPI methods are equivalently efficient with the examinee-centered selection and more efficient than random selection and the direct comparison of the D-optimal criterion values; (2) under the 2PL model, the two OIRPI methods are slightly more efficient than random selection and the direct comparison of the D-optimal criterion values, and the examinee-centered selection is the least efficient; (3) under the 3PL model, all four methods except the examinee-centered selection are equivalently efficient whereas the examinee-centered selection suffers significant problems.

7.2.3 Statistical Significance and Interactions

Eyeballing the figures could raise a question of statistical significance. Although there appear to be differences among the methods, are these differences statistically significant? To answer this question, linear regression models were applied to the 100 replications of outcome data. These 100 replications are deemed independent random samples because the item parameters, examinee ability parameters, and examinee responses are all generated randomly in each replication.

Using only data from the Bayesian MEM estimation method, the remaining factors include (1) five pretest item selection methods and (2) three seeding locations. A separate linear model was fitted for each of the following outcome variables: (1) RMSE of b in the 1PL model, (2) RMSE of a in the 2PL model, (3) RMSE of b in the 2PL model, (4) RMSE of a in the 3PL model, (5) RMSE of b in the 3PL model, (6) RMSE of c in the 3PL model,

(7) weighted ICC area difference in the 1PL model, (8) weighted ICC area difference in the 2PL model, and (9) weighted ICC area difference in the 3PL model. For each dependent variable, first a linear model with two main factors and their interactions were fitted. Then, if no significant interactions were detected, a second model (the final model) was fitted with the main factors only.

Tables 7.1 and 7.2 present the regression coefficients and their significance levels. In each regression model, the baseline for the pretest item selection method is the OIRPI-S method, and the baseline for the seeding location is "early in the test". A positive coefficient in the table indicates the corresponding method is less efficient than the baseline method, and the star signs indicate that the corresponding terms are significantly different from the baselines.

The results in the tables are consistent with the figures in the previous section; moreover, statistical significance is found in the comparisons. Conditioning on pretest item selection methods, both the "middle of the test" and "late in the test" seeding locations generated smaller RMSE values for all parameters in all models except for c in the 3PL model, and four of them are significant.

Holding the seeding location constant, OIRPI-O is not significantly different from the baseline OIRPI-S method. Most other methods generated larger RMSE values than the baseline OIRPI-S method, except for b-parameters from the examinee-centered selection in the 2PL and 3PL models. The superiority of the proposed OIRPI methods in both individual parameter estimation and overall ICC differences are significant compared to the direct comparison of D-optimal values method and random method in the 1PL and 2PL models. The superiority of the proposed OIRPI methods is significant compared to the examinee-centered selection in the a-parameter of the 2PL model. In the 3PL model, no significant superiority was found in the proposed OIRPI methods as compared with the direct comparison of D-optimal values method and random method. Some interactions are observed between seeding locations and item selection methods in the a-parameters of the

2PL and 3PL models.

7.3 Conclusions

In summary, the following conclusions can be made within the settings of the simulation study:

1. The two OIRPI methods are the most efficient or one of the most efficient methods in online calibration, and this is consistent across the three IRT models.
2. The two OIRPI methods both improved the calibration efficiency over traditional random sampling in the 1PL and 2PL models.
3. Consistent with the theoretical analysis, the examinee-centered method generated poor results for 2PL and 3PL models.
4. For the two OIRPI methods, seeding locations in the middle of the test or late in the test generated more accurate calibration results than early in the test, and the results are similar between themselves.

	1PL, b	2PL, a	2PL, b	3PL, a	3PL, b	3PL, c
(Intercept)	0.0982^{***}	0.1622^{***}	0.1090^{***}	0.2597^{***}	0.1781^{***}	0.0518^{***}
Seed(Middle)	-0.0022^{*}	-0.0072	-0.0041^{*}	-0.0113	-0.0013	0.0002
Seed(Late)	-0.0015	-0.0086^{*}	-0.0039^{*}	-0.0083	-0.0011	0.0003
Sel(OIRPI-O)	0.0009	-0.0018	0.0001	0.0051	0.0018	0.0011^{*}
Sel(D-Optimal)	0.0083^{***}	0.0093^{*}	0.0095^{***}	0.0045	0.0046	0.0008
Sel(Examinee)	0.0013	0.0150^{***}	-0.0153^{***}	0.0036	-0.0138^{***}	0.0051^{***}
Sel(Random)	0.0076^{***}	0.0064	0.0077^{***}	0.0077	0.0024	0.0004
Seed(M):Sel(OIRPI-O)		0.0019		0.0011		
Seed(L):Sel(OIRPI-O)		0.0008		0.0003		
Seed(M):Sel(D-Optimal)		0.0041		0.0153		
Seed(L):Sel(D-Optimal)		0.0103^{*}		0.0110		
Seed(M):Sel(Examinee)		0.0162^{**}		0.0149		
Seed(L):Sel(Examinee)		0.0196^{***}		0.0195^{*}		
Seed(M):Sel(Random)		0.0078		0.0047		
Seed(L):Sel(Random)		0.0109^{*}		0.0135		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7.1: Linear regression of the RMSEs of the item parameter estimates in the 1PL, 2PL, and 3PL models

	1PL	2PL	3PL
(Intercept)	0.0285^{***}	0.0356^{***}	0.0436^{***}
Seed(Middle)	-0.0007^{*}	0.0000	0.0000
Seed(Late)	-0.0004	0.0003	0.0000
Sel(OIRPI-O)	0.0004	0.0004	0.0006
Sel(D-Optimal)	0.0020^{***}	0.0009^{**}	-0.0003
Sel(Examinee)	0.0005	0.0022^{***}	0.0018^{**}
Sel(Random)	0.0020^{***}	0.0011^{***}	-0.0004
Seed(M):Sel(OIRPI-O)			0.0003
Seed(L):Sel(OIRPI-O)			-0.0004
Seed(M):Sel(D-Optimal)			0.0002
Seed(L):Sel(D-Optimal)			-0.0001
Seed(M):Sel(Examinee)			0.0019^{*}
Seed(L):Sel(Examinee)			0.0031^{***}
Seed(M):Sel(Random)			0.0003
Seed(L):Sel(Random)			0.0008

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7.2: Linear regression of the weighted ICC area differences in the 1PL, 2PL, and 3PL models

Chapter 8

Discussion

Recently, many educational and psychological assessments are moving towards computerized (adaptive) modes, such as the new K-12 state assessments created under the Race To The Top (RTTT) program, the National Assessment of Academic Progress (NAEP), and the patient-reported outcome (PRO) measurements in medical practices. Many of these CAT programs are high stakes and administered over multiple years, and therefore, this places high demands on item calibration and it is crucial to calibrate the items efficiently and economically.

Although CAT brings in challenges, it also provides a great solution to large-scale calibration due to the nature of “online” and “sequential”. In fact, CAT offers the unique opportunity to assign different pretest items to each examinee adaptively. One major advantage of CAT is that it provides more efficient latent trait estimates with fewer items, and the current study shows that it should also be true that CAT provides more accurate calibration of item parameters with fewer examinees than that is required in conventional paper-and-pencil tests.

Online calibration has been studied for decades to dynamically sample examinees for calibrating new items more efficiently than the traditional pretesting methods. However, the pretest item selection methods that are both practically feasible and theoretically well-founded appear to be rather limited. In contrast to the existing item selection methods in online calibration, this thesis made the argument that what is practically feasible and possibly effective is not comparing all examinees from an unavailable examinee pool, not comparing the pretest items and select the one that simply generates the highest information value for either estimating examinee abilities or estimating item parameters, but to compare all pretest items when an examinee reaches a seeding location and select the item that *needs* the current examinee most.

In this spirit, this thesis proposed an new pretest item selection framework, the Ordered Informative Range Priority Index (OIRPI), including two algorithms to implement it — OIRPI with order statistics and OIRPI with standardization. Both proposed methods showed superior efficiency than the existing methods with varied margins in the simulation study.

There are several limitations in the current simulation design. One limitation is that the results and conclusions from the simulation study are limited within the specific simulation design. Although an effort was made to better mimic a practical test setting, the performance of the methods under other settings still merits investigations. Another limitation is that the current simulation study fixed the sample sizes for all pretest items. In the future, termination rules based on measurement accuracy can be developed. Lastly, the pretest items selected according to these item-centered rules could stand out from the overall adaptive trend of item difficulty levels. However, examinees' judgments on the item difficulty levels can be inaccurate and widely varied (Vispoel, Clough, Bleiler, Hendrickson, & Ihrig, 2002; Vispoel, Clough, & Bleiler, 2005). So whether this poses an actual problem in practice is yet to be empirically investigated.

8.1 Future Directions

Below are a few thoughts on possible future directions following this study.

Other criteria for OIRPI This study only adopts one information criterion for the OIRPI framework — the D-optimal criterion. In fact, many other information criteria originally developed for *multidimensional CAT* can be borrowed into online calibration. The formulations of these criteria can be simply adjusted by replacing the examinee ability vector with the item parameter vector. Some of the item selection methods for multidimensional CAT are the *Kullback-Leibler information index* (Veldkamp & van der Linden, 2002; Wang, Chang, & Boughton, 2010; Wang & Chang, 2011), the *continuous entropy method* (Wang & Chang, 2011), the *mutual information method* (Mulder & van der Linden, 2009; Wang & Chang, 2011), and the *Bayesian Kullback-Leibler information method* (Mulder & van der Linden, 2009; Wang & Chang, 2011).

Termination rules A variety of termination rules can be developed, and their impacts on the ultimate sample sizes and their interactions with item selection methods and estimation methods can be studied.

For example, the following three termination rules can be compared: (1) sample size rule: a pretest item is exported from the pretest phase when its sample size reaches a predetermined critical value. (2) Standard error rule: a pretest item is exported when all of its item parameters have standard errors below their corresponding critical values. A maximum sample size can also specified — once a pretest item reaches the maximum sample size, it is exported regardless of its standard errors. (3) Stabilization rule: there are many possible ways to define whether the estimates are stabilized. The most intuitive way is to compare the absolute difference between the last two updates with the critical values. It is also reasonable to use the last several updates. A maximum sample size can also be specified for this rule.

Polytomous models Both the pretest item selection methods and statistical estimation methods can be extended to polytomous IRT models.

Multidimensional IRT and cognitive diagnostic models There have been recent developments in the statistical estimation methods for online calibration with multidimensional IRT and cognitive diagnostic models (e.g., Chen et al., 2012). Future studies can also extend the pretest item selection methods proposed in this thesis to multidimensional IRT and cognitive diagnostic models.

Hybrid seeding locations This study only compared three simple choices for seeding locations. However, more sophisticated designs can also be implemented, such as (1) start from the first 1/3 of the test in the early stage of sampling and switch to the middle 1/3 after receiving a certain number of samples and then switch to the last 1/3 after receiving more samples, and (2) start from the last 1/3 and switch to the middle and then the first 1/3 as the sample size accumulates. The rationale for these two designs resembles that of the a-stratification method (H.-H. Chang & Ying, 1996). Further theoretical investigation will be conducted if significant difference is found among these options.

Other research questions There are many other possible research questions surrounding online calibration, such as

- How can the proposed designs be adapted to multistage testing?
- How can online calibration be utilized to help build vertical scales?
- How can calibration efficiency be balanced with test-taker experience?
- Is it helpful or distracting to include halfway done pretest items in estimating examinee ability levels?
- How does the online calibration design interact with termination rules of variable-length CAT?

The past, current, and future studies on online calibration could lay the foundation for the implementation of these strategies in operational testing programs. In the future, online calibration could become standard in the maintenance of item banks in long-term CAT programs.

References

Ali, U., & Chang, H.-H. (2011). On-line calibration design for pretesting items in adaptive testing. In *the 11th international and the 76th annual meeting of the psychometric society.* Hong Kong, China.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis (2nd ed.).* New York, NY: Wiley.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Armitage, P. (2002). *Statistical methods in medical research (4th ed.).* Bodmin: MPG Books.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* (Second ed.). New York: Marcel Dekker, Inc.

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-line Pretest Item—Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *38*(3), 191–212. doi: 10.1111/j.1745-3984.2001.tb01123.x

Berger, M. P., & Wong, W. K. (2005). *Applied optimal designs.* England: John Wiley & Sons Ltd.

Berger, M. P. F. (1991). On the Efficiency of IRT Models When Applied to Different Sampling Designs. *Applied Psychological Measurement*, *15*(3), 293–306. doi: 10.1177/014662169101500310

Berger, M. P. F. (1992). Sequential Sampling Designs for the Two-Parameter Item Response Theory Model. *Psychometrika*, *57*(4), 521–538. doi: 10.1007/BF02294418

Berger, M. P. F., King, J. C. Y., & Wong, W. K. (2000). Minimax D-optimal designs for item response theory models. *Psychometrika*, *65*(3), 377–390. doi: 10.1007/BF02296152

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Buyske, S. (2005). Optimal Design in Educational Testing. In M. P. F. Berger & W. K. Wong (Eds.), *Applied optimal designs* (pp. 1–19). England: John Wiley & Sons, Ltd.

Chang, H.-H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*(4), 333–341.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. doi: 10.1177/014662169602000303

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*(3), 1466–1488. doi: 10.1214/08-AOS614

Chang, Y.-c. I. (2011). Sequential estimation in generalized linear models when covariates are subject to errors. *Metrika*, *73*, 93–120. doi: 10.1007/s00184-009-0267-y

Chang, Y.-c. I., & Lu, H.-Y. (2010, August). Online Calibration Via Variable Length Computerized Adaptive Testing. *Psychometrika*, *75*(1), 140–157. doi: 10.1007/s11336-009-9133-0

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012, February). Online Calibration Methods for the DINA Model with Independent Attributes in CD-CAT. *Psychometrika*, *77*(2), 201–222. doi: 10.1007/s11336-012-9255-7

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369–383.

Clyman, S., Melnick, D., & Clauser, B. (1999). Computer-based case simulations from medicine: assessing skills in patient management. *Innovative simulations for assessing professional competence*, 29–41.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, *5*(8).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

Jones, D. H., & Jin, Z. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika*, *59*(1), 59–75. doi: 10.1007/BF02294265

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. doi: 10.1111/j.1745-3984.2006.00021.x

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In *the gmac conference on computerized adaptive testing.* Retrieved from http://iacat.org/sites/default/files/biblio/cat09kingsbury.pdf

Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*, 259–267.

Lord, F. M. (1980). *Applications of item response to theory to practical testing problems.* Lawrence Erlbaum.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. , *22*, 224–236.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mislevy, R. J., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika*, *65*(2), 149–156.

Mulder, J., & van der Linden, W. J. (2009, June). Multidimensional Adaptive Testing with Optimal Design Criteria for Item Selection. *Psychometrika*, *74*(2), 273–296. doi: 10.1007/s11336-008-9097-5

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, *16*(2), 159–176.

Nering, M. L., & Ostini, R. (2011). *Handbook of polytomous item response theory models.* New York, NY: Taylor & Francis.

Ren, H., & Diao, Q. (2013). Item Utilization in a Continuous Online Calibration Design. In *the annual meeting of the national council on measurement in education.* San Francisco, CA.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement.*

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* American Psychological Association.

Silvey, S. D. (1980). *Optimal Design.* London, New York: Chapman and Hall.

Stocking, M. L. (1988, May). *Scale drift in online calibration* (Tech. Rep. No. RR-88-28-ONR). Princeton, NJ: Educational Testing Service.

Swanson, L., & Stocking, M. L. (1993). A Model and Heuristic For Solving Very Large Item Selection Problems. *Applied Psychological Measurement*, *17*(2), 151–166.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*, 973–977.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York, NY: Springer.

van der Linden, W. J., & Ren, H. (2014). Optimal Bayesian Adaptive Design for Test-Item Calibration. *Psychometrika*.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*(3), 273–291.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional Adaptive Testing with Constraints on Test Content. *Psychometrika*, *67*(4), 575–588. doi: 10.1007/BF02295132

Vispoel, W. P., Clough, S. J., & Bleiler, T. (2005). A Closer Look at Using Judgments of Item Difficulty to Change Answers on Computerized Adaptive Tests. *Journal of Educational Measurement*, *42*(4), 331–350.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., & Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement*, *39*(4), 311–330.

Wainer, H. (2000). Introduction and History. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (pp. 171–199). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Mislevy, R. J. (2000). Item Response Theory, Item Calibration, and Proficiency Estimation. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Lawrence Erlbaum.

Wang, C., & Chang, H.-H. (2011, May). Item Selection in Multidimensional Computerized Adaptive Testing–Gaining Information from Different Angles. *Psychometrika*, *76*(3), 363–384. doi: 10.1007/s11336-011-9215-7

Wang, C., Chang, H.-H., & Boughton, K. A. (2010, November). Kullback–Leibler Information and Its Applications in Multi-Dimensional Adaptive Testing. *Psychometrika*, *76*(1), 13–39. doi: 10.1007/s11336-010-9186-0

Ying, Z., & Wu, C. F. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica*, *7*, 75–92.

Zheng, Y., Wang, C., Culbertson, M., & Chang, H.-H. (2014, April). Overview of Test Assembly Methods in MST. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications.* New York, NY: CRC Press.

Zhu, R. (2006). *Implementation of Optimal Design for Item Calibration in Computerized Adaptive Testing (CAT).* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.