DESIGNING *P*-OPTIMAL ITEM POOLS IN COMPUTERIZED ADAPTIVE TESTS WITH POLYTOMOUS ITEMS

By

Xuechun Zhou

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2012

ABSTRACT

DESIGNING *P*-OPTIMAL ITEM POOLS IN COMPUTERIZED ADAPTIVE TESTS WITH POLYTOMOUS ITEMS

By

Xuechun Zhou

Current CAT applications consist of predominantly dichotomous items, and CATs with polytomously scored items are limited. To ascertain the best approach to polytomous CAT, a significant amount of research has been conducted on item selection, ability estimation, and impact of termination rules based on polytomous IRT models. Few studies investigated the optimal pool characteristics for polytomous CAT implementation. Using the generalized partial credit model (GPCM) (Muraki, 1992), this study aims to identify an optimal item pool design for tests with polytomous items by extending the *p*-optimality method (Reckase, 2007).

The extension includes the definition of $a\theta$-bin to describe polytomous items succinctly for pool design, and item generation strategy using constrained nonlinear optimization method. Optimal item pools are generated using CAT simulations with and without practical constraints. Because items are not generated during the CAT process, a loosely defined bootstrapping approach is proposed for the simulations. The item pool characteristics under each condition are summarized and their performance is evaluated against an extended operational item pool.

The results indicated that the practical constraints of the *a*-stratified exposure control and content balancing do not affect pool size to a large extent. However, the *a*-stratified control affects the pool characteristics greatly: the items included in the simulated pools with the control

have larger ***a***-parameter and provide higher maximum information on average. On the other hand, the content balancing applied in this study has little impact on pool design.

The evaluation results of the pool performance are closely related to the pool characteristics. When the *a*-stratified exposure control applied, the consistent results include: 1) the average test information is lower than that without the constraint; 2) RMSE is higher and the correlation between the true and estimated abilities is lower; 3) the percentage of the correct classification for the highest achievement level is lower. However, for all the simulated optimal pools, the test information is consistently above the target level 10.0, it is thus concluded that the *a*-stratified method resulted in an efficient use of the less discriminative items with small decreases in measurement precision.

With regard to the item pool usage, the percentage of items that are fully used, well used, rarely used, and never used are quite comparable for the pools designed with constraints. In addition, compared with the extended operational pool, when the *a*-stratified method applied, the conditional test overlap rate of the simulated optimal pools is consistently lower.

For the pool blueprint, because the normal ability distribution is assumed, more items that are informative at the middle of the ability scale are included for all the simulated optimal pools. The distributions of the ***a***- and ***b***- parameters are similar under the four simulating conditions. The threshold parameters, d_j, are orderly distributed from the easiest to the hardest, as demonstrated in the operational items. When the *a*-stratified method applied, there are fewer items in the first stratum and the number of items in the second and third stratum is nearly identical. In addition, with the content balancing control, the number of items in the first content area is slightly less than the second one.

ACKNOWLEDGEMENTS

This dissertation would not be completed without the tremendous help from many people. First, I am deeply grateful to my advisor Dr. Mark Reckase who has provided me with great insight and guidance that motivated me to pursue this topic. His profound knowledge of psychometrics and professionalism have inspired me to become a better psychometrician during the course of my doctoral study. I also thank my dissertation committee members Dr. Edward Roeber, Dr. Richard Houang, and Dr. Yuehua Cui who have provided constant support along the way.

I am also indebted to Dr. Michelle Williams who has provided me an opportunity to work as a statistician and psychometrician on the two science education projects funded by NSF. I also thank Dr. Kimberly Maier for her support during my doctoral study.

My sincere gratitude to many other professors including Dr.Tenko Raykov, Dr. Ken Frank, Dr. Yiming Xiao and Dr. Lifeng Wang for their superb instructions of psychometrics and statistics. I also thank many graduate students and friends who have enriched my time and studies including Tingqiao Chen, Cheng-Hsien Li, Ke Li, Jessica Hall, Liyang Mao, Joi Merritt, Bing Tong, Qiu Wang, Changhui Zhang, and Yisu Zhou.

I could not have completed my doctoral study without the support and unconditional love of my parents. Lastly, my heartfelt gratitude for my husband and my daughter for your love, which has been an endless source of encouragement during the past five years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The advantages of computerized adaptive testing (CAT) over traditional paper-and-pencil tests (P&P) have been widely acknowledged. The most recognized strengths of CAT include: 1) it increases test efficiency by decreasing test length; 2) it improves measurement precision by administering items tailored to individuals' ability levels; 3) it enables faster processing of test data and score reporting; and 4) a wider range of item types can be used (Wainer, 2000; Weiss & Schleisman, 1999). Despite the challenges of the calibration of item pools, test security, and test validity, CAT shifted from research to the implementation stage during the 1990s. Many operational programs have been put into applications such as the Armed Services Vocational Aptitude Battery (ASVAB), the National Council Licensure Examination (NCLEX), the Graduate Record Examination (GRE), and Computerized Adaptive Placement Assessment and Support Services (COMPASS). As high-speed computers become widely available in schools, the prospects of administering educational assessments by CATs are promising.

Current CAT applications consist of predominantly dichotomous items, and CATs with polytomously scored items are limited to the medical and psychological assessments (Boyd, Dodd, & Choi, 2010). However, including performance-oriented items into CAT programs for cognitive assessments is already on the agenda. For instance, in the adaptive summative assessments of English Language Arts and Mathematics that will be administered by the SMARTER Balanced Assessment Consortium (SBAC) from 2014, 18% of the assessment will be traditional constructed-response items and 18% will be performance assessments for Grades 3-8 and Grades 9-12 (SBAC, 2011).

On the other hand, the technology advances in automated text scoring implies the feasibility of including constructed-response questions with multiple categories in CAT programs. So far, automated text and essay scoring and automated scoring of mathematical graphs and expressions have been applied in CAT programs (Bennett, Morley, & Quardt, 2000; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997; Shermis & Burstein, 2003). With the development of technology-enhance items (TEI) that support automated scoring, the prospects of including partial-credit scoring items in CATs are promising (Parshall, Davey, & Pashley, 2002; SBAC, 2012). As a consequence, a substantial amount of research has been conducted on item selection and ability estimation methods for CATs using various polytomous item response theory (IRT) models. This is referred as polytomous CAT in this study. However, these studies assume that there is already an item pool supporting the CAT implementation, few studies have been conducted to investigate the desired qualities of item pools and their design.

An operational CAT program is typically described with four major components: a precalibrated item pool, an item selection algorithm, an ability estimation method, and a stopping rule (Boyd et al., 2010; Reckase, 1989). Given an IRT model, the statistical and psychometric features of the items included in the pool decide measurement precision. Also, the test specification and practical security concerns determine the pool size. When practical constraints such as content balancing and item exposure control are implemented, larger pools are required correspondingly (Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Pastor, Dodd, & Chang, 2002). In addition, item pool information distribution needs to match the ability distribution of the target population of examinees (Dodd, Koch, & De Ayala, 1993)

With the item selection algorithm, CATs select items that satisfy the defined constraints to maximize measurement accuracy. The most widely used item selection methods are the maximum information (Lord, 1980) and Bayesian methods (van der Linden, 1998). When issues of content validity and test security arise, practical constraints addressing the issues are embedded in the item selection process.

The ability estimation method provides updated estimates based on an examinee's responses to administered questions during the process and a final estimate. Essentially, there are two approaches to estimate ability in CATs (Boyd et al. 2010). The first one uses the likelihood function given an IRT model, and the ability estimates solve the function for the zero if it exists. In cases of extreme scores at the beginning of a CAT administration, a step size procedure is usually adopted. The second one is based on a posterior distribution, and the estimates are usually the expected value or mode of the distribution. A mixture of the two approaches can be adopted during the estimation process.

The termination rules include fixed and variable test length for a CAT application. When a fixed-length rule applies, CAT administrations stop after a fixed number of items are used. For a variable-length rule, CATs terminate when predetermined stopping criteria are fulfilled. Three major criteria used for the variable-length stopping rules are the standard error, minimum information, and the sequential probability ratio test (SPRT) (Boyd et al. 2010).

While a significant amount of research has been conducted on the four components to ascertain the best approach to CAT applications, few studies investigated the optimal pool characteristics for a specific CAT program implementation. Undoubtedly, without an item pool that contains high-quality items covering a wide span of ability levels, CAT programs cannot function as anticipated (Flaugher, 2000). Since developing an item pool with the desired qualities is an arduous effort with respect to the cost and time spent on writing, revising and pretesting items, it is efficient to understand pool characteristics at the beginning to guide item writing and pool development efforts. This goal is achieved through a method of optimal item pool design using simulations. Put simply, given a defined CAT program and all of its predetermined constraints, CAT simulations are carried out for a population of examinees. The resulting item pool is optimal in a sense that all desired features required for the program are fulfilled. The main product, an optimal blueprint, summarizes the item and pool attributes such as item distributions and pool size. The blueprint is then used to guide item writing efforts for building and managing item pools on a continuous basis.

There are two major methods to address the optimal item pool design. One uses the technique of integer programming (van der Linden, Veldkamp, & Reese, 2000; van der Linden, 2005). Another is the *p*-optimality methodology proposed by Reckase (2007). The latter one is used in this study.

The integer programming model formulates the test assembly as a constrained optimization problem with desired qualities being expressed as constraints (Veldkamp & van der Linden, 2000). The constraints encompass all the qualitative and quantitative features during a CAT implementation process. This approach is flexible in simulating constrained CATs under any IRT model with the objective function vary as desired. However, it does not calculate the number of items needed for a pool (Veldkamp & van der Linden, 2000), and the programming is more complicated compared with the *p*-optimal method.

The *p*-optimality method was proposed to approximate an optimal pool of smaller size with little loss of measurement precision. To be more specific, when two ability estimates ($\hat{\theta}$s) differ slightly, while two optimal items are needed, one of them is adequate, supposing the desired characteristics of the two items vary negligibly for the two $\hat{\theta}$s. Consequently, this method divides the continuous ability scale into many discrete units to describe items based on the proportional optimality required. This method has been successfully used in designing item pool blueprints under various CAT situations using dichotomous IRT models (Gu, 2007; He, 2010; Reckase, 2007).

Because the implementations of CATs rely on specific IRT models, results regarding item pool design from dichotomous models are not directly applicable to the polytomous context. For instance, compared with item pools based on dichotomous items, item pools consisting of polytomous items can function well with substantially smaller sizes (Dodd et al., 1993; Dodd, De Ayala, & Koch, 1995). The scenario becomes more complicated taking into consideration the practical constraints required by each CAT procedure. It is thus necessary to investigate the properties of optimal item pools under realistic contexts, and apply the results to guide the item writing effort to construct high-quality pools.

Using the generalized partial credit model (GPCM) (Muraki, 1992), one of the most investigated polytomous IRT models in CAT, this study aims to identify an optimal item pool design for tests with polytomous items by extending the *p*-optimality method. Practical constraints of content balancing and item exposure control are considered. Optimal item pools are generated using CAT simulations with and without practical constraints. Specifically, the following research questions were addressed in this study:

1. How do practical constraints such as content balancing and item exposure control affect the optimal item pool design and its performance?

2. For each combination of the constraints, what does the blueprint display with regard to the characteristics and distribution of the items, item pool information distribution, and pool size for a modeled CAT procedure?

The results obtained from this study will promote a better understanding of characteristics of optimal item pools and how they will impact CAT performances when polytomous models are used. The implications for informing item writing efforts and developing item pools are also discussed.

CHAPTER 2
LITERATURE REVIEW OF POLYTOMOUS CAT

This chapter first presents the GPCM and its information function, along with a discussion of the relationship between information function and item parameters. It then provides an overview of polytomous CAT research on item selection and ability estimation methods. Practical considerations of content balancing and item exposure control are also discussed.

2.1 Generalized Partial Credit Model and Its Information Function

Numerous studies have been conducted over the last several decades to investigate how CAT can be applied for assessment including items with multiple response categories. Polytomous IRT models that are most investigated include the nominal response model (NSM; Bock, 1972), the rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Maters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the graded response model (GRM; Samejima, 1969) (Boyd et al., 2010). Based on how a response category is derived, Thissen and Steinberg (1986) described the GRM as a "difference" model, and the others as the "divide-by-total" models. In this study, the GPCM is used to model the interaction between test-takers and polytomous items. This section introduces the GPCM and its information function.

2.1.1 Generalized Partial Credit Model

Muraki (1992) developed the GPCM by extending the PCM with a varying slope, which is an extension of the two-parameter logistic model (2PLM) to polytomous items. With this model, for item *j* with (m_j +1) possible categories, the probability for a given θ to receive a score *k* is denoted as the category response function (CRF) shown below

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=0}^{k} D\mathrm{a}_{\mathrm{j}}(\theta - b_j + d_{jv})\right]}{\sum_{c=0}^{m_j} \exp\left[\sum_{v=0}^{c} D\mathrm{a}_{\mathrm{j}}(\theta - b_j + d_{jv})\right]} \tag{2.1}$$

where *D* is a scaling factor, and a_j is a slope parameter that “indicates the degree to which categorical responses vary among items as *θ* levels change” (see Muraki, 1992). b_j is the location parameter indicating the overall difficulty of the item, and $d_{v=1}^{k}{}_{jv}$ is a threshold parameter that “is interpreted as the relative difficulty of step *k* in comparing other steps within item *j*” (see Muraki, 1992). d_0 is defined as 0 arbitrarily because it is cancelled as a common factor. An illustration of the CRF for an item of six response categories is shown in Figure 2.1.

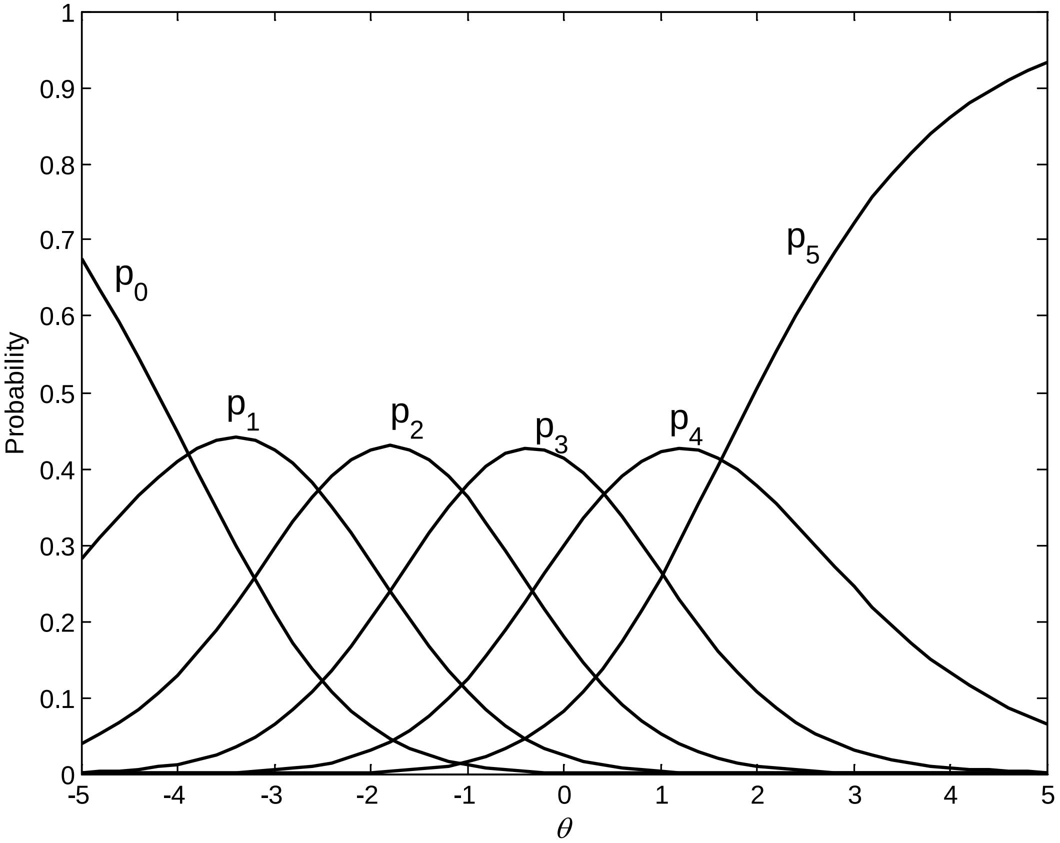


Figure 2.1 Item category response probability curves for *a* = 0.79, *b* = -1.09, d_s = 2.80, 1.45, -0.03, -1.50, -2.72

Under the GPCM, items sharing a set of threshold parameters are defined as a block, the sum of the d_{jv} parameters is constrained to 0 for an unique parameter estimation.

2.1.2 Item Information Function of the GPCM

Item information is the basis for item selection in adaptive testing. Item information function (IIF) for polytomous models may be estimated at the category and item level (Dodd et al. 1995). For the GPCM, the information function for item *j* at a given θ is denoted (Donoghue, 1994)

$$I_j(\theta) = D^2 a_j^2 \left[\sum_{k=0}^{m_j} k^2 P_{jk}(\theta) - \left(\sum_{k=0}^{m_j} k\, P_{jk}(\theta)\right)^2\right] \tag{2.2}$$

where $P_{jk}(\theta)$ is defined in function 2.1. An illustration of the item information curve (IIC) is displayed in Figure 2. 2. Test information is the sum of item information of all items included in a test.

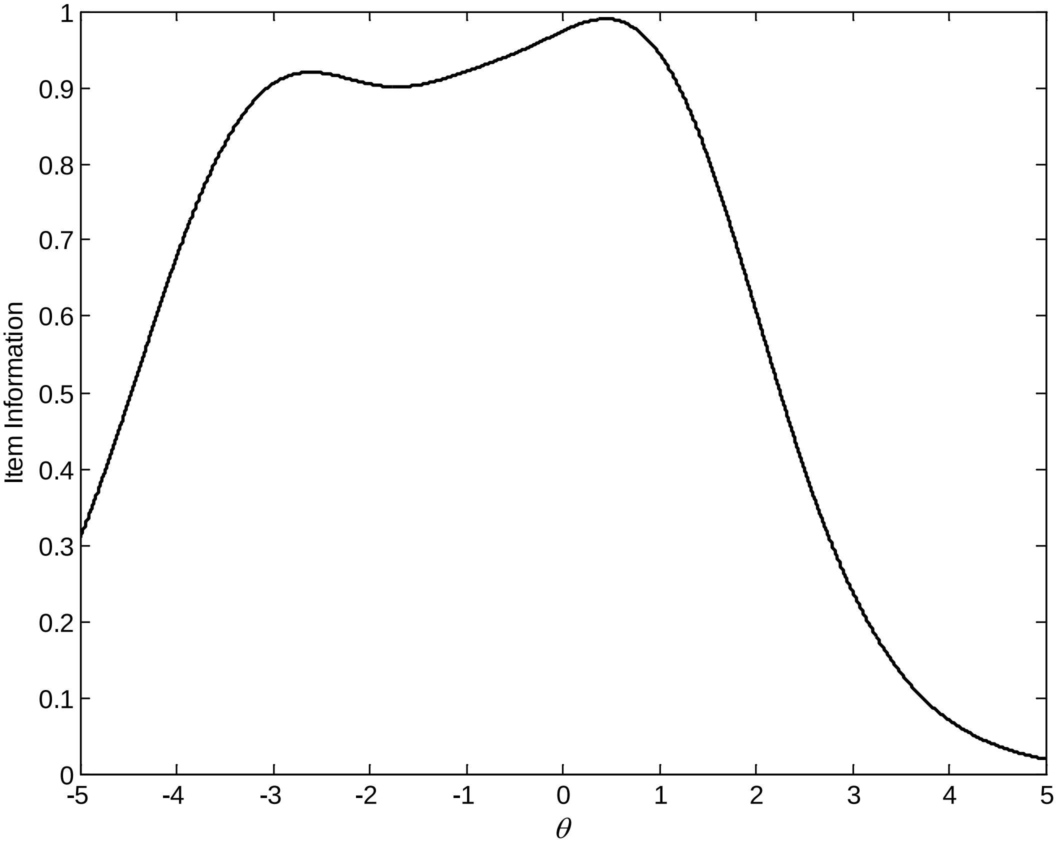


Figure 2.2 Item information curve for *a* = 0.79, *b* = -1.09, and d_s = 2.80, 1.45, -0.03, -1.50, -2.72

2.1.3 Effects of Item Parameter Variations on Item Information under the GPCM

For the GPCM, the total amount of information provided by items with the same number of steps is the same across the ability continuum (Dodd & Koch, 1987). Understanding the relationship between item parameters and IIC is critical for this study for two reasons. One is that it affects the magnitude of an item's maximum information and its location on the θ-scale, and the other is that it helps to understand the characteristics of operational items to guide subsequent item generation. The variations relevant to the IIC shape include the magnitude of *a*-parameter, the distance between the first and last threshold parameters, the ordering of the threshold parameters, and the proximity of two adjacent threshold parameters (Akkermans & Muraki, 1997; Dodd & Koch, 1987). The relationships are discussed and demonstrated separately below.

Firstly, there is a quadratic relationship between item information and discrimination parameter as shown in function 2.2. In other words, items with high discrimination parameters provide more information when holding other variables constant. The impact of higher discrimination on item information can be observed in Figure 2.3 in which two items differ in *a*-parameter by 0.2 while other parameters remains identical. The item with the higher *a* provides more information consistently across the ability continuum except on the upper end. However, although highly informative items are always desired during CAT administrations, it is neither realistic nor efficient to assume an item pool consisting exclusively of items with large *a*-parameters.

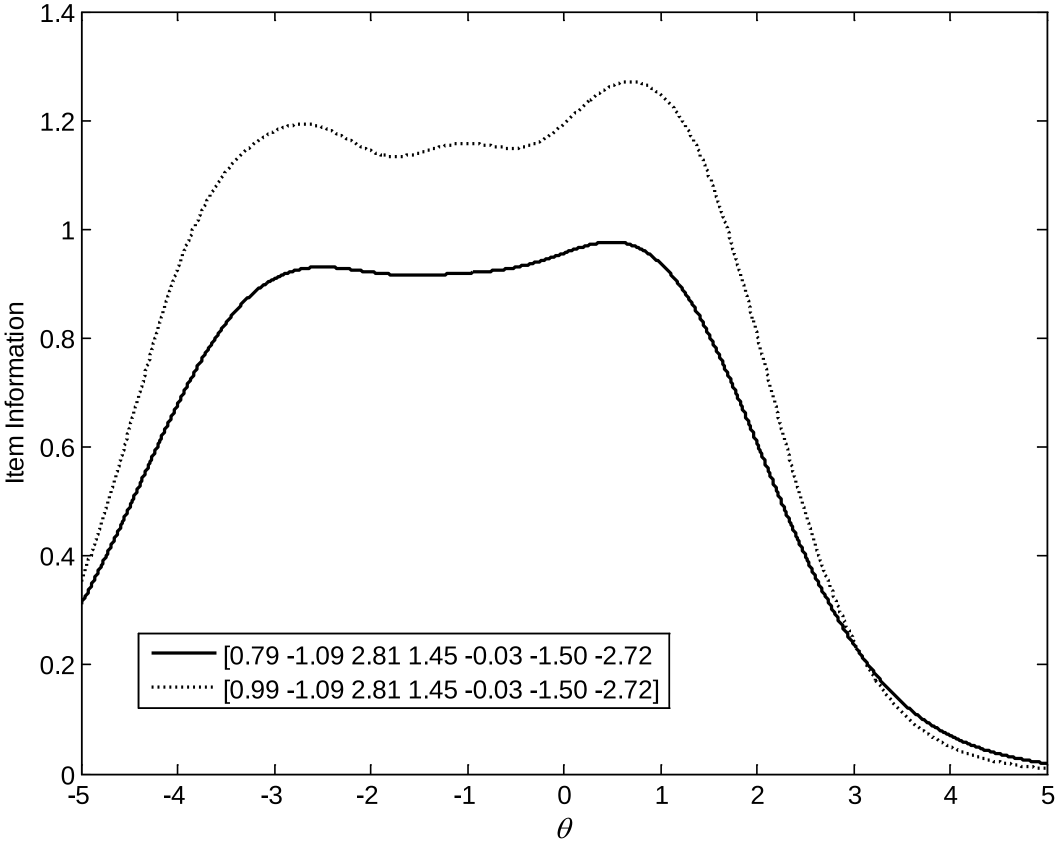


Figure 2.3 Item information curves for b = -1.09, d_s = 2.81, 1.45, -0.03, -1.50, -2.72, and a = 0.79 (solid curve) and 0.99 (dashed curve)

Secondly, the proximity of the first and last threshold parameter affects the shape of the IIC as well. Particularly, when other variables are held constant, the item with a shorter distance between the first and last threshold parameters provides more peaked information (Dodd & Koch, 1987). For the items of the same number of response categories, narrowing the distance also suggests that the item information shrinks to a shorter ability range. An example of information provided by three items with varying distances between the first and last threshold parameter is shown in Figure 2.4, in which the ranges between the first and last threshold parameters are 4.5, 5.5 and 6.5 on the ability scale from the top to the bottom.

As demonstrated, the most peaked information curve is for the item with the smallest range, and the flattest information curve for the one with the widest range. Furthermore, the information is more spread out for the item with the larger range, as shown at both ends of the ability continuum.

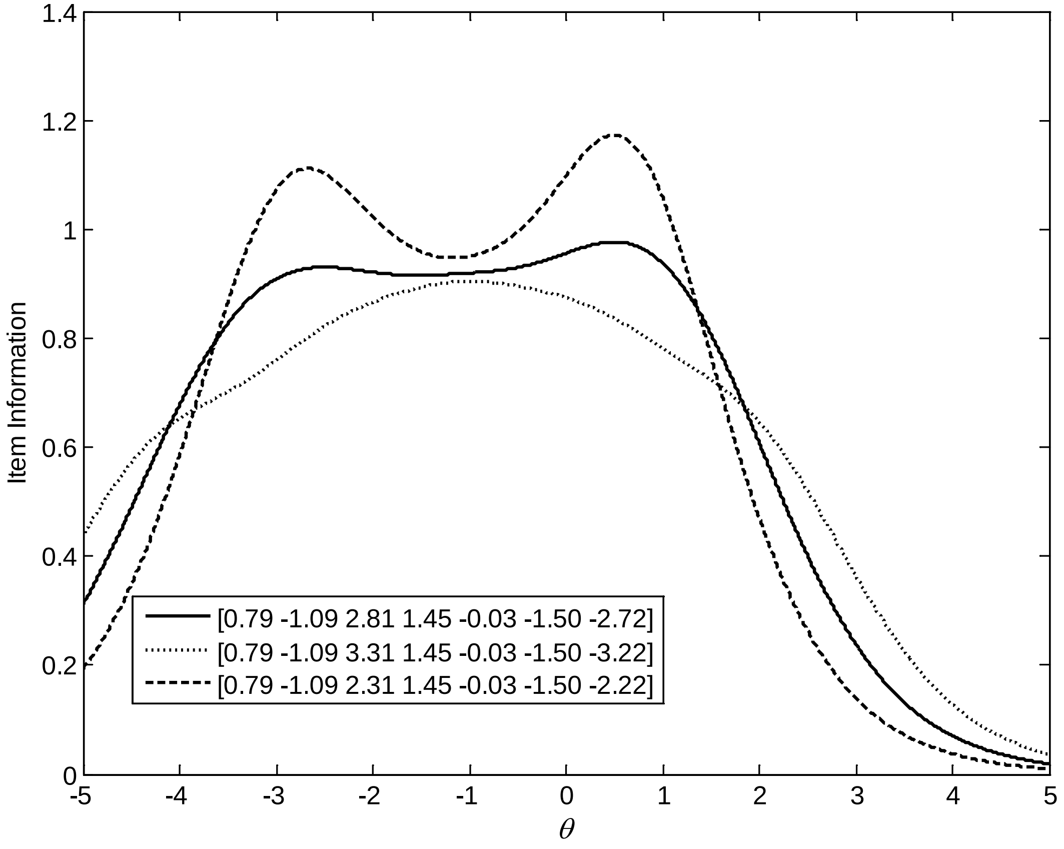


Figure 2.4 Item information curves for *a* = 0.79, *b* = -1.09, d_s = 2.81, 1.45, -0.03, -1.50, -2.72 (*solid curve*), d_s = 3.31, 1.45, -0.03, -1.50, -3.22 (*dashed curve*), and d_s = 2.31, 1.45, -0.03, -1.50, -2.22 (*dotted curve*)

Finally, the sequential order of the threshold parameters and magnitude of the distance between two adjacent ones have an impact on the item information as well. With the GPCM, Dodd and Koch (1987) found that for items of four category responses, when holding the distance between the first and last step parameter constant, if the step parameters are in an order from the easiest to the most difficult, the information curve is the flattest. Using this pattern as a reference, they also concluded that the more the step parameters are out of sequential order and the greater the disposition is, the more peaked the information curve. An examination of the generated items shows that their findings also hold for the six-category items. For instance, if threshold parameters, $[d_1\ d_2\ d_3\ d_4\ d_5]$, are in an ascending order in difficulty, greater disposition such as switching d_1 with d_5 results in a more peaked information curve than switching d_2 with d_5. When the total distance remains unchanged, there are 11 different ways of reordering the threshold parameters except for the one in an ascending order. While displaying all combinations in one graph is unnecessary, the effect of reordering the threshold parameters on item information is selectively demonstrated in Figure 2.5.

The two information curves with the greatest peaks occur when the largest disposition is made through switching the first, d_1, and last threshold parameter, d_5. Furthermore, if d_2, d_3, and d_4 are also switched, all the threshold parameters are out of order compared with the reference pattern, the IIC is thus the most peaked as shown in Figure 2.5. The flattest is for the item where the threshold parameters are in sequential order.

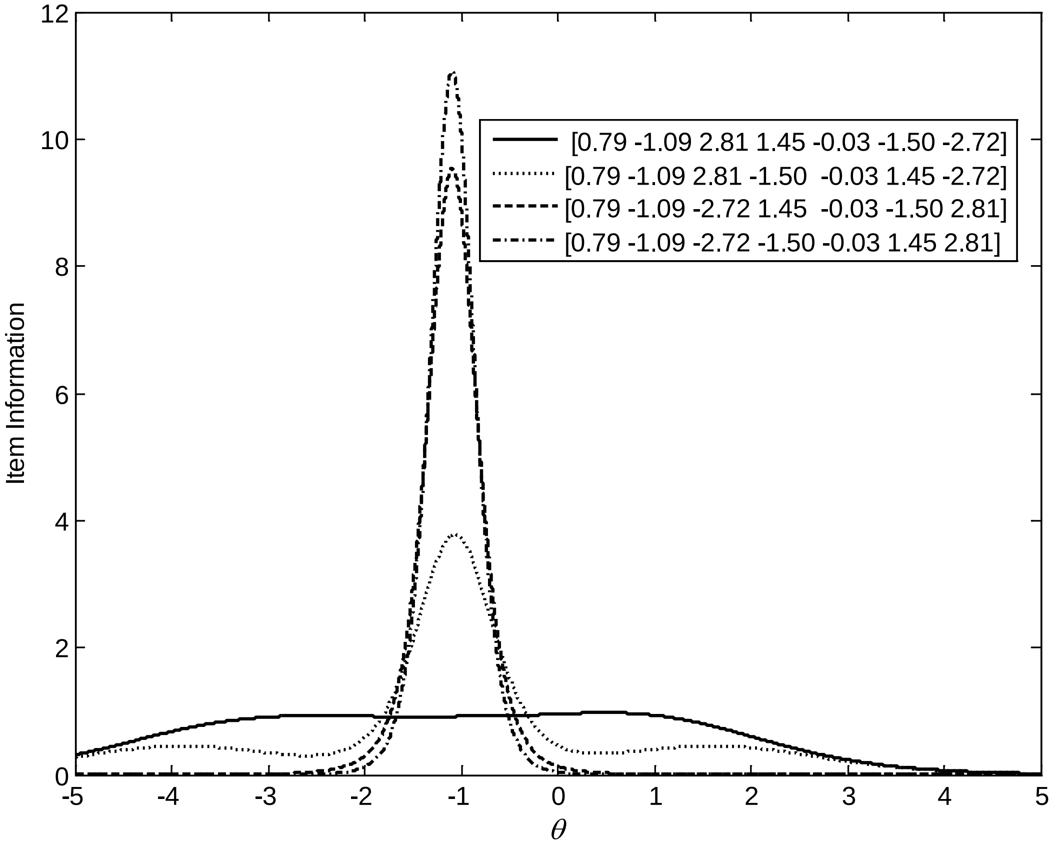


Figure 2.5 Item information curves for *a* = 0.79, *b* = -1.09, d_s = 2.81, 1.45, -0.03, -1.50, -2.72 (*solid curve*), d_s = 2.81, -1.50, -0.03, 1.45, -2.72 (*dashed curve*), d_s = -2.72, 1.45, -0.03, -1.50, 2.81 (*dotted curve*), and d_s = -2.72, -1.50, -0.03, 1.45, 2.81 (*dash-dot curve*)

In addition, the distance between two adjacent threshold parameters relates to the modality of the IIC. When the distance is large enough, the information curve is not unimodal (Akkermans & Muraki, 1997; Muraki, 1993). For instance, for items with three categories, the IIC is bimodal when *Da* $(d_2 - d_1) >$ 4In2, and unimodal otherwise (Akkermans & Muraki, 1997). While this inequality cannot be generalized to items with more than three categories, the effect of the proximity of two adjacent threshold parameters on the item information curve can be demonstrated in Figure 2.6.

The solid black IIC is for an operational item whose IIC is unimodal with maximum information located at 0.71 on the ability scale. With all other parameters identical, increasing d_2 and decreasing d_4 by 0.3 units, the IIC becomes bimodal as shown by the dotted curve.

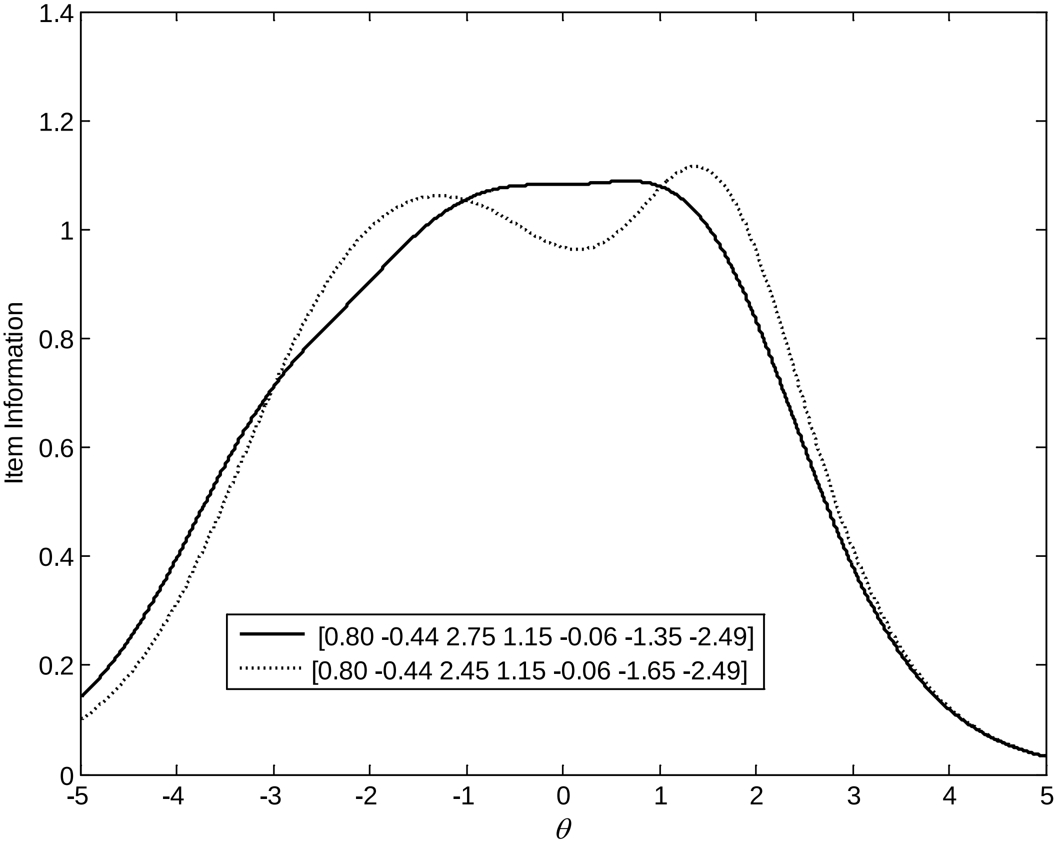


Figure 2.6 Item information curves for a = 0.80, b = -0.44, d_s = 2.75, 1.15, -0.06, -1.35, -2.49 (*solid curve*), and d_s = 2.75, 1.45, -0.06, -1.65, -2.49 (*dotted curve*)

2.2 Item Selection Methods for Polytomous CAT

CAT programs are typically described with four components: an item pool, an item selection algorithm, an ability estimation method, and a stopping rule (Dodd et al., 1995; Reckase, 1989). Each component, as well as practical concerns regarding test validity, security and cost, affects the outcomes of a specific CAT procedure (Chang, Qian, & Ying, 2001; Kingsbury & Zara, 1989). These considerations result in constraints embedded in CAT implementation such as content balancing and item exposure control. In addition, model selection taking into account factors such as data type, model assumptions, model fit, and model parsimony has an impact on the results (Dodd et al, 1995). Because this study concentrates on item pool design with a fixed-length CAT program without a predetermined pool, findings from previous polytomous CAT research investigating item selection and ability estimation methods are summarized to guide the CAT algorithm selection.

During the CAT process, an item is selected adaptively to address specific psychometric and practical specifications that a CAT program requires. In other words, items satisfying the predetermined conditions in these two aspects are assumed to be optimal and efficient to administer. Two major approaches are adopted in CAT practice. One approach is based on item information and the other is the Bayesian approach. A brief introduction to these two methods and a summary of the results from prior studies are presented in the following sections.

2.2.1 Item Information Approach

For this approach, measurement precision is achieved using the Fisher information. In other words, item selection is based on the amount of information an item contributes at a provisional θ point. Such procedures include maximum information (MI; Lord, 1980), Kullback-Leibler information (KLI; Chang & Ying, 1996; Lima Passos, Berger, & Tan, 2007; Veldkamp, 2003), and general weighted information (GWI; Veerkamp & Berger, 1997; Choi &Swartz, 2009; van Rijn, 2002).

Maximum information (MI). MI is the most widely used method for item selection to ensure measurement precision. It aims to select an item from a pool that provides the maximum Fisher information at a provisional θ estimate. This selection criterion is repeated until the predetermined stopping rule is satisfied. For the GPCM, the Fisher information is calculated using the function 2.2. MI is favored for its asymptotical property and straightforward calculation compared with other information functions used in different methods. Asymptotically, the standard error of θ estimates, $\hat{\theta}$, is

$$Var\left(\hat{\theta}\middle|\theta\right) = \frac{1}{I(\theta)} \tag{2.3}$$

where $\hat{\theta}$ is the maximum likelihood estimation (MLE) of θ, and $I(\theta)$ is the test information, the sum of item information (Lord, 1980).

There are several disadvantages of using the MI criterion. First, the selected item maximizes the information at the $\hat{\theta}$ instead of the true θ. Second, due to the quadratic relationship between item information and *a*-parameter, items with large *a* values tend to be selected more frequently than those with low *a* values, which causes concerns about test security and inefficient use of informative items at early stages of a CAT when θ estimates are not stable.

Kullback-Leibler information (KLI). Chang and Ying (1996) argued that when θ estimates are not close to true θ values at early stages of a CAT implementation, the local information that MI method provides at a specific θ estimate might not be efficient for item selection. They proposed KLI as a global information approach, which is defined as

$$K_j(\theta\ ||\theta_0)\ \equiv\ E_{\theta_0}\log\left[\frac{L_j(\theta_0;u_j)}{L_j(\theta;u_j)}\right] \tag{2.4}$$

where $K_j(\theta\ ||\theta_0)$ denotes the KLI of item *j* with response u_j for any θ given true parameter θ_0, E_{θ_0}represents expectation over u_j, and $L_j(\theta;u_j)$ is the likelihood function for item *j*.

Test information is the sum of the item information. Because θ_0 is unknown and θ is unspecified, KLI as an item selection index is defined by the function below (Chang & Ying, 1999)

$$K_j\left(\hat{\theta}_n\right) = \int_{\hat{\theta}_n-\delta_n}^{\hat{\theta}_n+\delta_n} K_j(\theta||\hat{\theta}_n)d\theta \tag{2.5}$$

where $\hat{\theta}_n$ is the θ estimate based on the responses to *n* items. This index represents the area under the KLI function over a specified interval of $\hat{\theta}_n$ for item *j*: ($\hat{\theta}_n - \delta_n$, $\hat{\theta}_n + \delta_n$). The item with the maximum *K*$\left(\hat{\theta}_n\right)$ is selected to be administered.

General Weighted Information (GWI). When there are no finite solutions or multiple solutions to the likelihood functions, Veerkamp and Berger (1997) proposed the GWI as an alternative item selection criterion for CAT. This criterion is defined as

$$GWI_j(\theta) = \int_{-\infty}^{\infty} W_n(u_n;\theta)I_j(\theta)\mathrm{d}\theta \tag{2.6}$$

where W_n is a weight function based on the *n* items that have been administered, u_n is the response vector to the *n* items, and I_j is the Fisher information given θ. Numerical approximation is usually adopted to solve the function. As formulated in 2.6, GWI criterion selects an item that provides the maximum weighted average information for a specified ability range.

Based on the GWI criterion, the MI is a special case of the function 2.6 when the weight function is 1 and θ equals the θ estimate. The other two variations of the GWI are the interval and likelihood weighted information (Veerkamp & Berger, 1997), which are formulated in functions 2.7 and 2.8 respectively

$$GWI_j(\theta) = \int_{\theta=\hat{\theta}_L}^{\hat{\theta}_R} I_j(\theta)\mathrm{d}\theta \tag{2.7}$$

where $\hat{\theta}_L$ and $\hat{\theta}_R$ are the left and right limits of a confidence interval of θ. This selection criterion represents the area under the GWI function between $\hat{\theta}_L$ and $\hat{\theta}_R$, and an item with the maximum value is selected.

$$GWI_j(\theta) = \int_{-\infty}^{\infty} L_n(\theta; u_n) I_j(\theta)\mathrm{d}\theta \tag{2.8}$$

where $L_n(\theta; u_n)$ is the likelihood function of the *n* items with a response string u_n. Because this criterion calculates the area under the function across the entire ability levels, it does not use the updated θ estimates during a CAT process. Hence, item selection is not influenced when there are multiple local maxima or the interim θ estimates are infinite.

2.2.2 Bayesian Approach

In this approach, items are selected using information functions that incorporate a weight function based on a posterior ability distribution with the information function. This approach mainly includes maximum posterior weighted information (MPWI; van der Linden, 1998), maximum expected information (MEI; van der Linden, 1998), and minimum expected posterior variance (MEPV; van der Linden, 1998). Van der Linden (1998) and van der Linden and Pashley (2000) presented a comprehensive illustration of item selection using Bayesian criteria.

Maximum Posterior Weighted Information (MPWI). MPWI was proposed to address the uncertainty of interim θ estimates. Mathematically, it is formulated in function 2.9 (van der Linden, 1998),

$$MPWI_j = \int_{-\infty}^{\infty} I_{u_n}(\theta) g(\theta|u_n)\mathrm{d}\theta \tag{2.9}$$

where I_{u_n}equals the Fisher information based on the response string u_n, and $g(\theta|u_n)$ is the posterior distribution of θ derived from function 2.10,

$$g(\theta|u_n) = \frac{L(\theta|u_n)g(\theta)}{\int_{-\infty}^{\infty} L(\theta|u_n)g(\theta)d\theta} \tag{2.10}$$

where g(θ) denotes the prior distribution. An item with the maximum MPWI is selected for administration.

Maximum Expected Information (MEI). MEI criterion selects an item using the posterior predictive distribution (van der Linden, 1998) and Fisher information. In particular, for an item to be selected, *j*, its probability function after *n* administered items is

$$p_j\left(U_j = u_j \middle| u_1, \dots, u_{\mathrm{n}}\right) = \int_{-\infty}^{\infty} p_j\left(U_j = u_j \middle| \theta\right) g(\theta|u_1, \dots, u_{\mathrm{n}})\mathrm{d}\theta \tag{2.11}$$

where θ is estimated from the response string u_{n}. For dichotomous items, MEI is then calculated as

$$\begin{aligned} MEI_j = {} & p_j\left(U_j = 0 \middle| u_1, \dots, u_{\mathrm{n}}\right) I_{u_{\mathrm{n}}, u_{\mathrm{j=0}}}\left(\hat{\theta}_{u_{\mathrm{n}}, u_{\mathrm{j=0}}}\right) \\ & + p_j\left(U_j = 1 \middle| u_1, \dots, u_{\mathrm{n}}\right) I_{u_{\mathrm{n}}, u_{\mathrm{j=1}}}\left(\hat{\theta}_{u_{\mathrm{n}}, u_{\mathrm{j=1}}}\right) \end{aligned} \tag{2.12}$$

where I($\hat{\theta}$) is updated with possible responses to item *j* included and other notations are the same as defined previously. An item maximizes MEI is selected as the next item.

Minimum Expected Posterior Variance (MEPV). When the information function in 2.12 is replaced with the posterior variances of θ estimates, the MEPV criterion is formulated as

$$MEPV_j = p_j\left(U_j = 0 \middle| u_1, \dots, u_{\mathrm{n}}\right) Var\left(\theta \middle| u_1, \dots, u_{\mathrm{n}}, u_{\mathrm{j=0}}\right) \qquad (2.13)$$

$$+ p_j\left(U_j = 1 \middle| u_1, \dots, u_{\mathrm{n}}\right) Var\left(\theta \middle| u_1, \dots, u_{\mathrm{n}}, u_{\mathrm{j=1}}\right)$$

For these item selection methods, numerical approximation is used during a selection process. Penfield (2006) described the implementation of the MPWI and MEI in detail.

2.2.3 Previous Research Results

Numerous studies have been conducted to compare the performance of various item selection methods based on both dichotomous and polytomous IRT models. With relevance to this study, results based on polytomous models are presented below.

Veldkamp (2003) compared the performance of the three methods using item information (i.e., MI, KLI and GWI) with the GPCM for item pools of various lengths and discrimination ranges. Regarding item overlap rate and ability estimation precision, he found that MI performed equally well as the other two methods when the test length was fixed at 20. The author concluded that taking into account the asymptotic properties of the Fisher information and the ease of its computation, MI was recommended in polytomous CAT.

Van Rijn, Eggen, Hemker, and Sanders (2002) compared the performance of MI and GWI with regard to θ estimation precision using the GPCM. They also concluded that MI performed equally well as GWI.

Ho (2010) studied the performance of four item selection methods based on the GPCM (i.e., MI, MPWI, MEI, and MEPV) in constrained and unconstrained CATs. The results indicated that none of the four methods outperformed to others based on the evaluation criteria. The four methods produced comparable θ estimates across the ability continuum, regardless of ability estimation methods. No significant difference was found in terms of item exposure rate and pool utilization. He thus recommended MI selection method for CAT based on the GPCM.

2.2.4 Practical Considerations in Item Selection

In practice, item selection depending solely on the statistical properties of the models might cause concerns about test validity and security. For instance, some content areas may be underrepresented, or items overexposed are memorized. These considerations are addressed by imposing constraints on item selection process such as item exposure control and content balancing. A brief summary of the constraints is provided below.

2.2.4.1 Item Exposure Control

There are three types of methods for controlling item exposure: randomization item selection, conditional item selection and stratification approach (Way, 1998).

According to Way (1998), for randomization selection, the item administered is selected from a set of optimal items randomly. For conditional selection, the probability that a selected item will be used is conditional on the probability that the item is selected from a specific population of examinees. For each type of method, there are various procedures for its implementation. For instance, randomization selection includes the randomesque selection procedure (Kingsbury & Zara,1989) and the progressive-restricted method (Revuelta & Ponsoda, 1998). The Sympson-Hetter procedure (S-H; Sympson & Hetter, 1985) is representative of the conditional procedures.

In regard to stratification approach, an item pool is divided into several strata based on a single or multiple item parameters, and items are selected correspondingly from each level. With the unconstrained MI selection method, there is a concern that items with high *a* values tend to be overexposed during CAT administrations. Besides test security concern, this also results in an inefficient use of items with high discrimination parameters at early stages of a CAT when θ estimate is unstable and imprecise. Chang and Ying (1999) proposed the *a*-stratified (AS) method to control the item exposure rate. The AS method for dichotomous IRT models is briefly described below:

1) Divide the item pool into *K* strata in ascending order of *a*-parameters;

2) Divide the test into *K* stages accordingly;

3) For administration at each stage, select items from the corresponding *k*th stratum until a stopping rule is satisfied.

The number of strata is determined by the structure of the item pool (Hau, Wen, & Chang, 2002). According to Chang and Ying (1999), there are several factors to consider when deciding the number of strata. One is the variation of *a*-parameters: few strata are applied when *a*-parameters display little variation and a larger number of levels can be used when greater variations appear. Another factor is the item pool size: the larger the pool is, the more strata can be adopted. In addition, stratification needs to take into account the distribution of *b*-parameters. That is, each level after the stratification should contain items that are informative for a wide range of ability levels.

Pastor, Dodd, and Chang (2002) compared five item exposure control methods in CATs using the GPCM: AS, S-H, enhanced AS, conditional S-H, and conditional enhanced AS. They found that the AS method, compared to the no-exposure control, reduced item exposure and overlap rate and increased pool utilization with little loss in measurement accuracy. Regarding item overlap and exposure control, the AS method did not perform as well as the other four methods. However, among the five methods, the AS method is the least restrictive and simplest to implement, and it shows advantages in reducing the number of nonconvergent cases.

Yi, Wang, and Wang (2003) compared three item exposure control methods in CATs of 8, 12 and 20 items using the GPCM: AS, AS with *b*-blocking (AS_B), and MI_S-H. They used the location parameters, *b*, to divide the pools first, and stratified the pools into four levels. The results indicated that compared to the AS_B, the AS method only performed slightly worse with respect to the overall average test overlap rate. The conditional indices showed that the AS and AS_B methods performed similarly regarding the item exposure rate, test overlap rate, bias, and standard error. The authors also suggested that the characteristics of the items included in the pools had an impact on the performance of the various methods.

2.2.4.2 Content Balancing

To ensure that each CAT administration adheres to the content specification, content balancing is applied in item selection. This constraint also enables the test scores obtained from CATs to be relatively comparable (Stocking & Swanson, 1993). Two methods to achieve the content balancing control have been studied for polytomous CAT: constrained CAT (CCAT; Kingsbury & Zara, 1989) and rotation method (Boyd et al., 2010).

CCATs select items from the least represented content area. More specifically, the percentage of items administered in each content domain is compared to the target percentage as delineated in test specification, and the one with the largest discrepancy is selected. In the rotation method, items are selected in a fixed order from content-specific subsets (Segall, Kathleen, & Hetter, 1997). It is the simplest method, and it performs well if the content areas are evenly distributed (Boyd et al., 2010). The content subsets are assumed to be undimensional and the ability estimation is a single score.

Yi and Chang (2003) proposed an AS with a content blocking method that incorporates content balancing with the item pool stratification. For this method, when dichotomous IRT models are used, an item pool can first be partitioned by the content areas, then sorted by item difficulty and discrimination parameters.

For polytomous CAT research, Davis (2004) applied the CCAT method with six item exposure control procedures using the GPCM: two randomization, two conditional, and two stratification methods. Although this study focused on comparing the performance of the exposure control procedures, it indicated the feasibility of applying content balancing in polytomous CATs.

2.3 Ability Estimation Methods for Polytomous CAT

Ability estimation is crucial in CAT because it determines the item selection during the CAT process and the final test scores. This estimation process typically involves three stages: initial, interim, and final (Van der Linden & Pashley, 2000). The initial value is usually set equal to the mean of the population distribution. The interim value is estimated and upgraded using a specific method given an examinee's responses to items administered, which is then used to inform item selection. The final estimate is obtained based on the examinee's responses to all items. It is unnecessary to use the identical estimation method during the process. In current CAT practice, ability estimation is primarily based on either the likelihood function or the posterior distribution. The most widely used methods include maximum likelihood estimation (MLE; Lord, 1980; Birnbaum, 1968; MML; Bock & Aitkin, 1981; WLE; Warm, 1989) and Bayesian estimation (Owen's method; Owen, 1969, 1975; EAP; Bock & Aitkin, 1981; MAP; Samejima, 1969). A specific method, weighted likelihood estimation, derived from maximum likelihood is presented below.

2.3.1 Weighted Likelihood Estimation

Assuming items are locally independent, the likelihood function for a response vector u_{jk} is

$$L(U|\theta) = \prod_{j=1}^{n}\prod_{k=0}^{m_j}\left(p_{jk}(\theta)\right)^{u_{jk}} \tag{2.14}$$

where $p_{jk}(\theta)$ is the probability to respond to category k for item j at a given θ. u_{jk} =1 if the examinee's response falls in the category k, and u_{jk} = 0 otherwise. The MLE of θ is the solution that maximizes $L(U|\theta)$ obtained by solving the function below using an approximation method

$$\frac{\partial}{\partial\theta} l(U|\theta) = 0 \tag{2.15}$$

where $l(U|\theta)$ = In$L(U|\theta)$ is the log-likelihood function for a given model.

MLE is favored due to the attractive asymptotic properties (Samejima, 1969; Hambleton & Swaminathan, 1985). Asymptotically, MLE of θ is consistent and efficient of true θ. Additionally, $\hat{\theta}_{\text{MLE}}$ is normally distributed with the mean equal to true θ, and the variance equal to the reciprocal of the test information. However, $\hat{\theta}_{\text{MLE}}$ is biased outward (Lord, 1983), and the magnitude of the bias positively correlates with θ. Bias is small over the middle θ range, and large at extreme values: larger at negative than positive θ values (Lord, 1983). This is undesirable in a criterion-referenced test where precision at the cut-score is critical (Wang, Hanson, & Lau, 1999).

To reduce the first-order bias of MLE, Warm (1989) derived the weighted likelihood estimation (WLE) method using the 3-parameter logistic model (3PLM) model to deal with the bias. Essentially, bias is reduced by modifying the likelihood function with a weight function. WLE is obtained by solving the formula below

$$\frac{\partial l(u|\theta)}{\partial\theta} - \; Bias\left(MLE(\theta)\right) * I(\theta) = \; 0 \tag{2.16}$$

where $l(u|\theta)$ is defined in function 2.14 and, $I(\theta)$ is the test information function. Samejima (1998) extended WLE to include polytomous responses. The MLE bias function for the GPCM is shown below (Wang & Wang, 2001)

$$-\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}}{P_{jk}(\theta)} \tag{2.17}$$

$$= -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j} D^3\, a^3 P_{jk}(\theta)\left(k - \; \sum_{c=0}^{m_j} cP_{jk}(\theta)\right)$$

$$* \left[k^2 \; - \; 2k\sum_{c=0}^{m_j} cP_{jk}(\theta) + 2(\sum_{c=0}^{m_j} cP_{jk}(\theta)^{2} - \sum_{c=0}^{m_j} c^2 P_{jk}(\theta)\right]$$

2.3.2 Previous Research Results

Studies on ability estimation methods in CAT based on polytomous IRT models have compared the performance of MLE, WLE, EAP, and MAP (Chen, Hou, Fitzpatrick, & Dodd, 1997; Chen, Hou, & Dodd, 1998; Wang & Wang, 2001; Ho, 2010). In general, the results suggested that EAP and MLE performed similarly when normal or uniform prior distribution was assumed.

With regard to the ability estimation methods in the GPCM-based CATs, Wang and Wang (2001) examined the performance of the four estimation methods in the context of varying test termination rules. They concluded that for the fixed-length test, WLE showed smaller bias, standard error, and root mean squared error (RMSE) than MLE, and it had smaller bias than EAP and MAP. For the variable-length test (i.e., fixed reliability), WLE and MLE performed similarly in terms of all overall indices.

Ho (2010) compared item selection approaches using four ability estimation methods in CATs using the GPCM (i.e., MLE, WLE, EAP with normal prior distribution, EAP with positively skewed distribution). He found that the performance of the four methods were comparable on both constrained and unconstrained CAT. WLE is recommended for the reason that it resulted in significantly fewer non-convergent cases.

When an extreme response category is observed at the beginning of a CAT implementation, WLE might fail to converge (Chen & Cook, 2009; Gorin, Dodd, Fitzpatrick, & Shieh, 2005). A variable step size procedure can be used to update θ estimates until responses in nonextreme categories appear. In the program designed for polytomous CAT simulations, Chen and Cook (2009) used the step size as 0.5* ($\hat{\theta}$ + d_j), where $\hat{\theta}$ is the initial or previous interim ability estimate, d_j represents the first step parameter for the lowest response category, and the last for the highest category.

CHAPTER 3
p-OPTIMALITY METHOD FOR ITEM POOL DESIGN AND EXTENSIONS TO POLYTOMOUS CAT USING GPCM

This chapter first introduces the desired properties of an optimal item pool. It then presents the *p*-optimality method for describing an item pool and its application to item pool design in previous research using the 1PLM and 3PLM. Finally, the extensions of the method to the CAT item pool design based on the GPCM are discussed in detail with respect to the definition of *aθ*-bin, item generation strategy, and master pool generation.

3.1 Optimal Item Pool Design

3.1.1 Desired Properties of an Optimal Item Pool

An optimal item pool is defined as one that can always provide optimal items that satisfy the desired characteristics of a CAT program during implementation (Reckase, 2007). To achieve this, an item pool must have a sufficient number of items and a distribution matching the target population (Boyd et al., 2010; Lima Passos, Berger, & Tan, 2007; Reckase, 2007; Veldkamp & van der Linden, 2000). Throughout this study, item pools are deemed optimal as one of an adequate number of items resulting from CAT simulations with all predetermined psychometric, statistical and practical specifications satisfied.

Item pool size is shaped from a variety of CAT program features such as test length, number of examinees taking the test, content balancing, item exposure rate, and stopping rule (Way, 1998). When content balancing and item exposure need to be controlled for high-stake testing, a larger item pool is necessary to guarantee test validity and measurement precision. For instance, if there are few items in a specific content domain, overexposure of the items threatens the test validity and security. Also, while item exposure control might result in limited use of high informative items, measurement precision decreases when the test length is fixed (Stocking & Lewis, 2000).

With regard to item distribution, information shape of an item pool plays a critical role in CAT implementation. The shape of target information curve should reflect the purpose of a CAT program (Lord, 1980). In the context of a criterion-referenced test, the curve should peak around the cut-point score to ensure desired measurement accuracy (Parshall, Spray, Kalohn, & Davey, 2002). In addition, the information distribution of the item pool needs to match the ability distribution of the target population to avoid nonconvergent trait estimation cases (Dodd et al., 1993; Gorin et al., 2005).

As for the actual pool size, Pastor et al. (2002) found that when the exposure rate was controlled in the GPCM-based fixed-length CAT simulations with 1000 simulees, at least 100-120 items were required to maintain measurement precision. It can be anticipated that when content balancing is incorporated and examinees increase, a larger item pool is necessary to achieve the desirable precision. When items were unevenly distributed among different content domains, overexposure occurred even when the item pools contained 149 and 210 items (Burt, Kim, Davis, & Dodd, 2003).

3.1.2 Optimal Item Pool Design

Based on the desired properties of the optimal item pools, pool design simulation aims to produce a blueprint to guide test assembly based on the predetermined CAT characteristics (Veldkamp & van der Linden, 2000). In other words, optimality is realized when the desired characteristics such as measurement precision, content balance, or exposure rate are fulfilled. A blueprint describes the attributes of items in the pool and the number of items for each combination of constraints for a CAT implementation (Reckase, 2007; Veldkamp & van der

Linden, 2000). The blueprint resulting from the simulations not only guarantees measurement efficiency but also helps item writing effort for item pool construction and management.

Previous studies have used computer simulation of CAT procedures for both blueprint design and post hoc pool performance evaluation. For CAT using dichotomous models, two major techniques were examined for optimal item bank design. One is the mathematical integer programming (Theunissen, 1985; van der Linden, Veldkamp, and Reese, 2000) and another is the *p*-optimality method (Reckase, 2007). The *p*-optimality method has been successfully extended to the 3PLM (Gu, 2007) and to the constrained CAT (He, 2009). This study aims to extend this method to polytomous CAT using the GPCM with content balancing and AS constraints imposed.

3.2 *p*-Optimality Method

3.2.1 *p*-Optimality Method

A well functioning item pool is more than an assembly of available items. For an optimal item pool, there should always be an informative item available for each θ estimate using a specified item selection method. Because θ is defined on a continuous scale, even when θs differ as slightly as .001, different items are needed in an absolute sense of optimality. On the other hand, the desired characteristics of items such as item information provided by the two items across a short θ interval might vary quite negligibly. Including in an item pool a large quantity of items that function similarly is impractical as it greatly increases financial cost yet barely improves measurement precision.

The *p*-optimality method was introduced to approximate an optimal pool of smaller size with little loss of specified characteristics (i.e., item information). More specifically, items included in a pool are classified using a *p*-optimality criterion: *p*-proportional of the desired characteristics with *p* representing an acceptable level of a loss of the characteristics such as 98% of the maximum information. The concept of *p*-optimality and its implementation in CAT using different IRT models are described in detail below.

3.2.2 Applying the *p*-Optimality Method to Optimal Item Pool Design

When the *p*-optimality method is applied, an optimal item pool is generated under the constraints of predetermined CAT characteristics using a Monte Carlo method with an IRT model. Items are first generated for each updated θ estimate to maximize the desired characteristics. For instance, an item that yields the maximum information. However, not all items generated will be added to the final pool: they are added into the final optimal pool based on the *p*-optimality criterion. A blueprint summarizes the resulting pool with respect to item distribution and pool size. The key concept of applying the method -- "bin" and the general simulation procedures are explained below.

3.2.2.1 Information Unit "Bin'

Because items providing *p*-proportion or more of the specified characteristics are considered equally optimal for θ estimates within a defined interval, the application of the *p*-optimality criterion describes items in terms of item sets based on the amount of information items provide. The concept of "bin" was introduced to stand for such sets, and items from the same bin can be used interchangeably to satisfy the item selection criterion (Reckase, 2007). Using the MI as the item selection method, the bin unit under the Rasch model and 3PLM is illustrated below.

Under the Rasch model, defining bins on the θ scale is straightforward because an item's maximum information is located at $\hat{\theta} = b$. For an item with $b = 0$, and with $p = 0.95$, an illustration of a bin unit is shown in Figure 3.1.

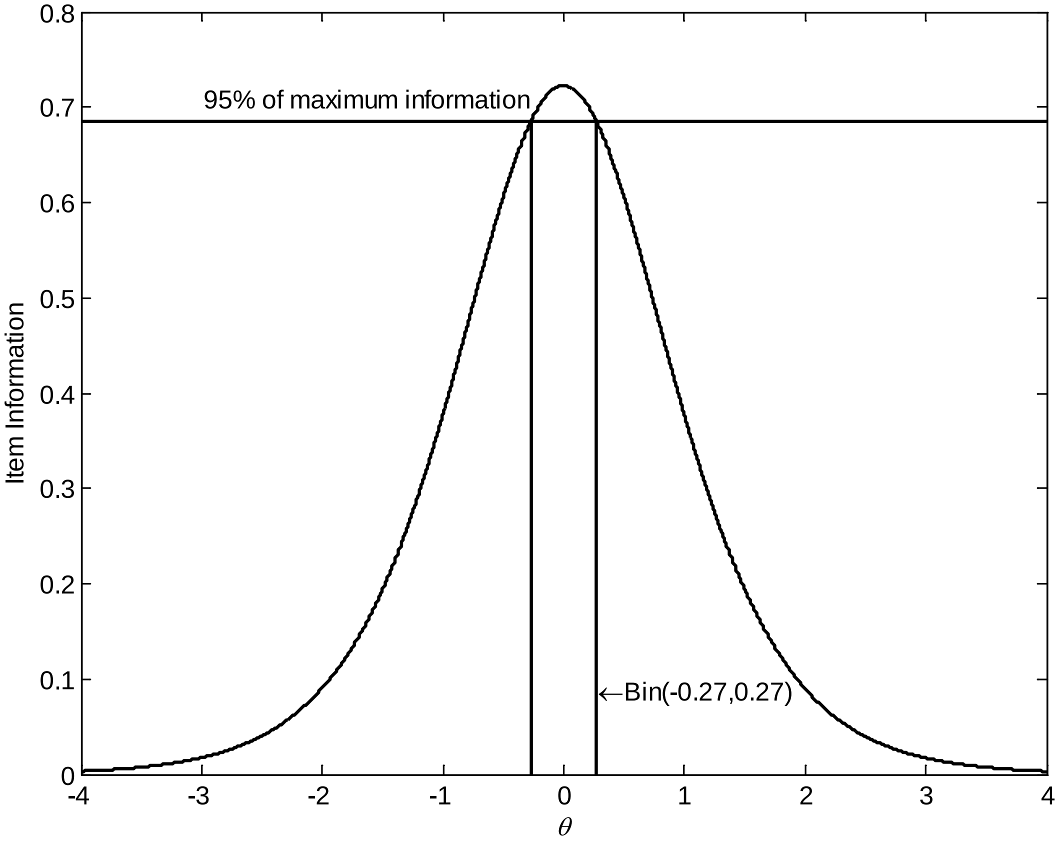
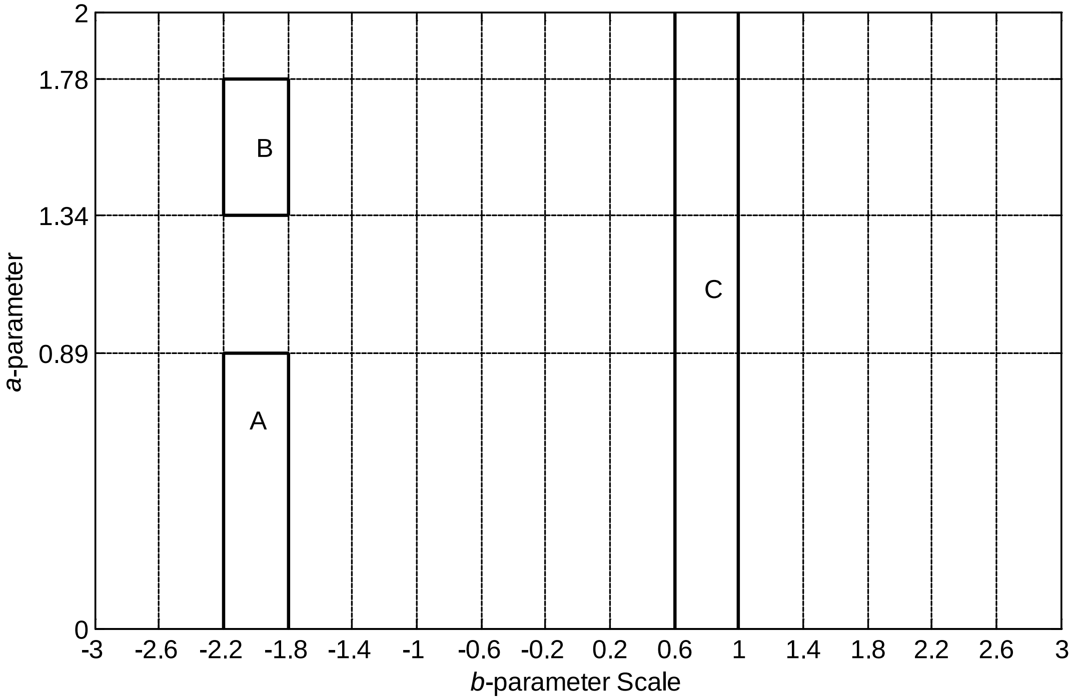


Figure 3.1 Illustration of bin unit for Rasch model

As the figure indicates, the maximum information is located at θ = 0, and the 95% interval of the maximal falls approximately between -0.27 and 0.27 on the θ scale. For $\hat{\theta}s$ within this range, items with *b*-parameters in this range belong to the same bin and they are considered as equally informative using the MI selection method. Moreover, during CAT simulation processes, if there is already an item included in the bin, no additional item needs to be added unless item exposure control is applied.

The concept of bin is extended to the 3PLM (Gu, 2007; Gu & Reckase, 2007). Because the information provided by an item conditional on θ depends mainly on *a*- and *b*-parameter, the bins are defined by *a*- and *b*-parameter, *ab*-bins. In their studies, the *b*-intervals are of equal distance on the θ metric, and *a*-intervals are determined by the predetermined acceptable amount of information change. Therefore, the width of the *a*-intervals is flexible. A graphic illustration of *b*-bins, *ab*-bins, and the change of maximum information provided by items within *b*-bins adapted from Gu (2007) is shown in Figure 3.2.



A: ab-bin with maximum information change of 0.4

B: ab-bin with maximum information change of 0.2

C: b-bin

Figure 3.2 Illustration of *b*-bins and *ab*-bins for 3 PLM

Furthermore, with the 3PLM, an item (i.e., j_1) providing its maximum information at a θ point can be less informative than another item (i.e., j_2) even if its information does not peak at that point. Both items are included in the pool because the simulation algorithm uses the maximum information only. Supposing the pool is then applied in practice, at the θ point, if no more informative items are accessible, item j_2 is selected over j_1, which suggests that j_1 can be removed from the original simulated pool. In their study, this is handled using a post-simulation adjustment before a final blueprint is obtained. When item exposure rate is considered, the post-hoc adjustment also addresses the exposure rate control. Namely, an additional item is added into the pool if an existing item reaches the predetermined item exposure rate.

3.2.2.2 Appling p-Optimality Method to Optimal Item Design Using 1PLM and 3PLM

Reckase (2007) and Gu (2007) have applied the *p*-optimality method to design optimal item pools and evaluate pool performance using computer simulations under the 1PLM and 3PLM. The goal is to investigate the property of an optimal pool when predetermined characteristics such as measurement precision, content balance, or exposure rate are fulfilled. In general, the optimal pool design uses a Monte Carlo method and is implemented by the following steps (Gu, 2007; Reckase, 2007):

1) Identify items' categorical qualities such as content areas and divide the pool into subsets based on model assumptions such as dimensionality and practical considerations such as exposure control.

2) Specify a CAT program's characteristics explicitly such as the expected ability distribution of the target population, test length, item types, item selection method, ability estimation method, item exposure control, and stopping rule.

3) Randomly sample a simulee from the expected ability distribution, and generate the first optimal item using a specified initial value (i.e., the mean of the ability distribution) based on the relationship between item parameters and θ that maximizes the item information.

4) Generate a response to the item using a selected IRT model and update the $\hat{\theta}$ with an ability estimation method.

5) Generate the next optimal item based on the updated $\hat{\theta}$.

6) Classify the resulting item into "bin/*ab*-bin" units.

7) The iterative process of ability estimation and item generation for one simulee repeats until the stopping rule is satisfied.

8) Repeat steps 3 to 7 for another simulee. However, additional items will only be included in the final pool unless no items are accessible or the existing ones have reached a preset exposure rate.

9) After an expected number of simulees are administered the test, the union of items in all bins forms the optimal pool if no adjustments are made.

Based on the final pool, a blueprint is produced summarizing the pool size and item distribution. The resulting pool is deemed optimal because it supports the CAT implementation with each desired characteristic satisfied. The blueprint is thus instructive for CAT programs with the identical requirements. In addition, simulation programs can be modified to meet varying CAT programs.

3.3 Extending the *p*-Optimality Method to Polytomous CAT Using GPCM

In this study, polytomous items with six categories are used to mirror items in an operational achievement test. As shown in equation 2.2, item information is determined by discrimination, location and threshold parameters. To extend the *p*-optimality method to design an item pool for a CAT implementation with the GPCM, strategies to define the bin unit, generate optimal items, and simulate an item pool with the constraints of content balance and item exposure control are presented in detail below.

3.3.1 Extending the "Bin" Concept

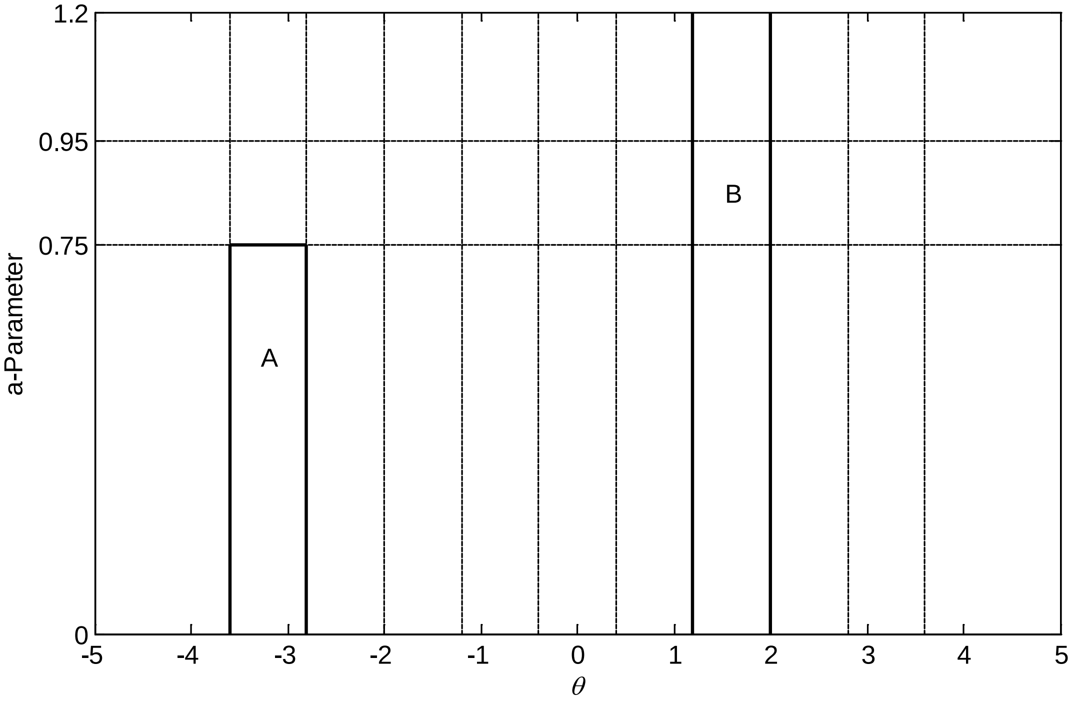
When the MI selection method is used, items are optimal in the sense that item information is maximized at a particular θ point. During a simulation to design an item pool, given an initial θ or an interim θ estimate, a set of item parameters needs to be generated based on their mathematical relationship with the maximum information. Mathematically, the relationship is deduced in two steps. First, maximize the item information function $I(\theta)$ with respect to θ. That is, take the derivative of $I(\theta)$ with respect to θ, set this derivative equal to zero, and solve for θ. Second, substitute the θ solution into the function $I(\theta)$ and solve for item parameters sequentially. For the GPCM with six response categories, the analytic solution of the θ that maximizes $I(\theta)$ is unavailable. Consequently, mathematical function indicating the relationship between maximum information and item parameters cannot be derived. Furthermore, with seven parameters determining the amount of information, defining the bin unit using item parameters cannot describe the item pool characteristics explicitly as what was conducted with the 1PLM and 3PLM. While the *b*-parameter in the GPCM is commonly used as a counterpart of the difficulty parameter in the 1PLM and 3PLM, inspection of the operational items revealed that the θ point corresponding to the maximum information point could deviate from the location parameter to a great extent. For example, when *b*-parameter falls in (-1, -0.5), θ values correspond to the maximum information point might fall in a wide range between -2 and 1.5.

On the other hand, given item parameters, there always exists a unique θ corresponding to where an item reaches its maximum information. The θ point is determined by the whole set of item parameters. Describing items based on the θ value captures a critical item characteristic without referring to location and threshold parameters directly. Furthermore, it enables polytomous items with a set of parameters to be graphically presented in a succinct manner, which is also essential for interpreting final blueprints. Consequently, the "bin" concept is extended as $a\theta$-bin with θ representing where an item provides maximum information. To present the information unit, the procedure to identify the boundary of a-bin and θ-bin is clarified, and the $a\theta$-bin concept is illustrated.

The boundary of ***a***-parameter is set up in a different way from what was used with the 3PLM (Gu, 2007; Gu & Reckase, 2007). With the 3PLM, the relationship between the change of the maximum information and *a*-parameter can be approximated analytically, which, however, is not applicable with the GPCM. In prior studies of AS and AS_B method in polytomous CAT, item banks were stratified according to the values of ***a***-parameter (Yi & Chang, 2003; Yi, Wang, & Wang, 2003). This approach is adopted to define the *a*-bin boundary.

The width of the *θ*-bin is determined by the predetermined proportion of the maximum information items provide. To determine the width for the 98-optimality criterion, 100 operational and generated items with the θs at the maximum information distributed across the ability continuum are examined. Since the average distance is 0.81 on the θ scale and the initial value to start the CAT is usually set to zero, the *θ*-bins center on zero with a fixed width of 0.8 except at both ends.

The range of *a*-bin and *θ*-bin defines the *aθ*-bin, as graphically illustrated in Figure 3.3. There are A × B cells in total with A denoting the number of stratum and B denoting the number of the θ intervals. The magnitude of A and B depends on the item pool size, item characteristics, and desired measurement precision. Particularly, when item pools are large and *a*-parameters demonstrate adequate variations, more *a*-bin can be adopted ((Yi & Chang, 2003). If higher measurement precision is desired, narrower *θ*-bin width should be applied. In this case, the final pool is composed of 3 × 11 cells for each content area, and items in each cell are assumed equally informative for the θ estimates in that interval. To record the cells simply, cells are indexed as $a\theta_{ab}$ (a = 1, 2, 3; b = 1, 2, ... ,11), with B ordered from the leftmost. For example, $a\theta_{16}$ refers to the 6th *θ*-bin (-0.4: 0.4) at the first stratum.



A: aθ-bin for items in the 1st stratum with MI in θ-interval (-3.6, -2.8)

B: θ-bin for items with MI in θ-interval (1.2, 2)

Figure 3.3 Illustration of $a\theta$-bin for GPCM

Although 98% is adopted to determine the θ-bin width, the proportion of maximum information is inconsistent in the item pool. This is because the 98% intervals for all items vary under the GPCM and the distance from each item's θ point to the θ-bin's center differs slightly. For instance, the 98% interval for an item with a flat information curve ranges across 1.2 ability units, but the interval for another item is probably 0.6 units. Also, items selected from one end of the θ-bin might deviate from the 98-optimality more than items close to θ-bin center. However, an examination of the item information curves indicates that the deviation is not severe.

3.3.2 Generating Optimal Items Using Constrained Nonlinear Optimization

To a large extent, an item pool is optimal because it consists of optimal items. When the *p*-optimality method is applied using the 1PLM and 3PLM, optimal items that are the most informative for $\hat{\theta}$s are generated simultaneously and sequentially during the CAT simulation. This is fulfilled through the use of the analytic solution of θ that maximizes the information function, $I(\theta)$. Under the GPCM, such an analytic solution is unavailable for items with six response categories, it is thus impractical to generate items during the CAT process. Instead, optimal items targeting the entire ability continuum are generated prior to CAT simulations.

Generating optimal items is realized through a mathematical optimization algorithm. The relationships discussed in section 2.1, along with their characteristics displayed in the operational items, are used to formulate the constraint inputs. In addition, a location constraint, $\sum_{v=1}^{m} d_v = 0$, is imposed (Muraki, 1992). This is used in item parameter estimation under the GPCM to ensure the comparison between the location parameters over the blocks (Muraki, 1992). That is to say, with the location constraint, item parameter invariance holds when the ability scale is fixed. For items of six categories, this constraint suggests that threshold parameters are somewhat symmetric around zero.

The application of constrained nonlinear optimization algorithm and some adjustments during the optimization are presented below.

First of all, the optimization problem is formulated as maximizing the item information function by systematically selecting item parameters from the allowable sets subject to the constraints (Snyman, 2005). In general, the constraints in mathematical optimization problems are represented in linear, nonlinear functions and/or in bounds. For item generation in this context, a linear equality constraint and bound constraints are applied. Items that provide maximum information across the ability continuum form the final pool. The mathematic form of the optimization problem is formulated as

arg maximize $f(\boldsymbol{x})$ (3.1)

subject to

$\boldsymbol{Aeq} * \mathbf{x} = Beq,$ (3.2)

Lower bound (LB) $\leq \mathbf{x} \leq$ upper bound (UB), (3.3)

where the objective function $f(\boldsymbol{x})$ is the information function. The design variable, $\boldsymbol{x}$, is the threshold parameters, $d_{v=1}^{m}$. In 3.2, ***Aeq*** is a coefficient vector of the threshold parameters, *Beq* is a vector of equality constraint, and **Aeq*x =** *Beq* represents the linear equality constraint, $\sum_{v=1}^{m} d_v = 0$. The inequality formula 3.3 quantifies the bound constraints imposed on the discrimination, location and threshold parameters.

Secondly, according to the functions 3.3, constraint input values are derived. As discussed in section 2.1, the constraints included are the magnitude of *a*-parameter, the proximity of the first and last threshold parameters, the ordering of threshold parameters, and the distance between two adjacent threshold parameters. For all the constraints, their input values were determined through analyzing the operational item calibration results.

Finally, adjustments are made to include randomness in item generation within the allowable sets specified by the constraints. This step is conducted due to the little variation among the items generated under the identical bound constraints. For example, *a*-parameter remains at the upper bound, and the distance between the two adjacent threshold parameters is identical. On one hand, merely duplicating items is undesirable for designing an optimal item pool. On the other hand, the complex relationship between the item information and item parameters suggests that an item's maximum information and its corresponding θ point vary greatly even with an identical bound constraint. The final item generation is thus conducted as described below:

1) Identify the bounds of *a*-parameter for each stratum, and randomly generate *a*-parameter as $a \sim U(a_{LB}, a_{UB})$.

2) θ is randomly generated as θ ~ *N* (0, 1) to set up the scale. To achieve a balanced distribution, *b*-parameter is randomly generated over intervals equally spaced on the ability scale. Because threshold parameters are usually between -4 and 4, large *b*-parameter values such as -2 or 2 would shift item response curves greatly. In practice, this would result in inadequate observations in extreme categories for item calibration. On the other hand, items with moderate *b*-parameter can be informative over a wide range of ability. Therefore, for *θ*-bins except at both ends, *b*-parameter intervals are commonly set as (-1, -0.5), (-0.5, 0), and (0, 0.5).

3) Formulate the bound constraints for the threshold parameters. Because of the linear equality constraint, $\sum_{v=1}^{m} d_v = 0$, restriction in the total range of θ, [-5, 5], and the constraint in distance between two adjacent threshold parameters, threshold parameters show a consistent pattern in magnitude. For instance, when ***b***-parameter falls into (-0.5, 0), the lower and upper bound for $d_{v=1}^{m}$ can be set up as (3, 1.5, -0.5, -2, -4) and (4, 3, 0.5, -1, -2.5) respectively. The bounds for the threshold parameters are adjusted slightly when the location parameters are changed to ensure the step parameters are in the range [-5, 5].

Under the GPCM, items sharing a set of threshold parameters are defined as a block (Muraki, 1992). With the method describe above, item parameter invariance is achieved over the blocks when the ability scale is fixed and the location constraint applies. Also, items generated demonstrate the characteristics of the operational items.

After items are generated, they are classified into $a\theta$-bins based on the *a*-parameter and the θ value where the MI is located. When information curves are bimodal, items are labeled according to where the larger MI is. Items are then divided into corresponding $a\theta$-bins as described in section 3.3.1 for two content areas. The final pool consists of the union of items in all $a\theta$-bins. In one sense, the pool is optimal because of the large quantity of items available for a wide coverage of θ with varying ***a***-parameter in two content areas.

CHAPTER 4
METHODS AND PROCEDURES

This chapter first introduces how a master item pool is developed. It then describes the CAT simulation procedures to obtain optimal item pools. An extended operational item pool of equal size to an optimal pool is obtained using the item parameter replication method. Finally, it presents the item pool performance evaluation criteria.

4.1 Generating the Master Item Pool

Because items are not generated during the CAT process, a loosely defined bootstrapping approach is proposed for CAT simulations instead. The rationale behind this bootstrapping approach is that optimal item pool distribution can be approximated by a resampling method. To be more specific, supposing there exists an optimal item pool that contains a sufficient number of the informative items for a CAT program, CAT simulations can be viewed as a process of resampling items with replacement from the pool. If the original pool is deemed optimal and all the desired CAT characteristics are satisfied during the simulation, the resulting item pools thus approximate desirable optimal pools. This original pool is referred to as the master pool, and the procedure to generate it is described below.

4.1.1 Identifying the Boundary for $\boldsymbol{a\theta}$-bins

At first, based on the $a\theta$-bin approach discussed in Chapter 3, it is necessary to identify a reasonable range of a-parameter values, the number of stratum, and the θ-bin width. In regards to a-parameter values, polytomous items administered in a large-scale achievement test were used as the benchmark. The assessment measures students' language proficiency, which consists of multiple-choice and open-ended items. Four open-ended items are included in each test form and each is scored on a six-point scale. A composite score is finally derived and the final score is reported on a five-point performance scale. 16 items over a 4-year period were used. For each form, approximately 90,000 examinees were randomly sampled. The items were calibrated using PARSCALE 4.1 with its default item parameter estimation method of marginal maximum likelihood (MME) (Muraki & Bock, 1999).

An examination into the operational items showed that the *a*-parameter varies slightly from 0.8 to 1.1. It is desirable to increase the variation for several reasons. Firstly, as CAT is viewed as individualized versus group testing in the P&P form (Wainer, 2000), a CAT item pool needs to include items of greater variation. Secondly, the master pool should be much larger than what was used in previous research. While an item pool requires a sufficient number of items, more stratification levels can be employed (Chang & Ying, 1999). Namely, a wider range of *a* values is needed. To understand the range of the *a*-parameter in operational polytomous items, 62 polytomous items from the National Assessment of Educational Progress National Reading Assessment (National Center for Education Statistics; NCES, 2011) are referenced. Of the 62 items, 48 have three categories and 14 have four, all of which were calibrated on a scale of normal standard distribution. For the 62 items, the ***a***-parameter falls into a greater range from 0.3 to 1.2. Taking into account the factors discussed above, the ***a***-parameter values used for the master pool generation was adjusted to 0.55 -1.10.

The number of strata depends on the variation of *a*-parameter (Chang & Ying, 1999). When the values of the *a*-parameter are similar, fewer levels are necessary compared with an item bank covering a wide range of *a*-parameter. Based on the range of the *a*-parameters proposed above, three strata are thus used with the boundary set from (0.55, 0.75), (0.75, 0.95), and (0.95, 1.1) respectively for simulating and describing item pools.

As discussed in section 3.3, for the 98-optimality criterion, the width of the θ-bin is around 0.80. Because items that are informative at both tails are usually less frequently used, the width of $\theta_{.1}$- and $\theta_{.11}$-bin at the ends is increased to 1.4. In particular, $\theta_{.1}$-bin contains items that are informative between the ability levels -5 and -3.6, and $\theta_{.11}$bin for those between 3.6 and 5. As a result, a total of 11 θ-bins are adopted for generating the pools.

4.1.2 The Master Pool

After the $a\theta$-bins properties are specified, the number of items in each $a\theta$-bin is determined. To mirror the operational achievement test, items are also generated for two content areas, so the master pool is the union of the items in the 33 $a\theta$-bins of each content area.

Because the master pool is used as the original sample for CAT simulations, it is supposed to be much larger than operational pools. Previous research on the CAT using the GPCM was based on item pools with varying sizes between 60 and 210 (Burt et al., 2003; Davis & Dodd, 2003; Pastor, Dodd & Chang, 2002). Considering that item pools tend to be larger when practical constraints are imposed, approximately 3,000 items are planned to be generated for the master pool.

According to Chang and Ying (1999), the number of items included at each *a*-level needs to be proportional to the items administered to ensure the item exposure rate to be comparable. Also, the number of items administered from different *a*-levels should be equal with the exception at the first stratum. Therefore, an equal number of items are generated for each stratum. Considering that the CAT simulation starts from the first stratum in content area one, more items are generated for that stratum. Additionally, because item distribution depends on the ability distribution, the number of items in each θ-bin is roughly proportional to the density function of the normal distribution.

The items are generated using the method presented in Chapter 3 based on the operational item characteristics. After an item is generated, it is classified into a corresponding $a\theta$-bin. Eventually, a total of 3,150 items are generated, which is approximately 15 times greater than the largest item pool used in previous research. The number of items included in the master pool is shown in Table 4.1.

Table 4.1 Number of Items in the Master Pool

Stratum	Content Area 1	Content Area 2
First	560	518
Second	518	518
Third	518	518

Except for the first stratum in content area one, the number of items in each θ-bin is presented in Table 4.2.

Table 4.2 Number of Items in θ-bin

	***aθ*-bins**										
	$a\theta_{.1}$	$a\theta_{.2}$	$a\theta_{.3}$	$a\theta_{.4}$	$a\theta_{.5}$	$a\theta_{.6}$	$a\theta_{.7}$	$a\theta_{.8}$	$a\theta_{.9}$	$a\theta_{.10}$	$a\theta_{.11}$
Number of items	4	5	10	50	110	160	110	50	10	5	4

4.2 Designing *p*-Optimal Item Pool for Polytomous CAT

Although the operational test is a mixed form of multiple choice and polytomous items, to fully investigate the characteristics of an item pool consisting of polytomous items, the simulated CATs include polytomous items exclusively.

4.2.1 CAT Characteristics

To specify the characteristics of the CAT simulations, item selection and ability estimation methods are determined based on previous research results. The item pools are prepared for each simulation condition from the master pool. In addition, a stopping rule needs to be decided. Because the effect of the termination rule is not emphasized in this study, along with the consideration of practical constraints, the CATs to be simulated are set as with a fixed test length. Because the test is criterion-referenced, measurement precision is desired. Based on the operational items, the test length to achieve measurement accuracy is determined through the process described below.

Assuming ability follows the normal distribution, the score points from real administrations are matched on the ability scale based on the score distribution (Hambleton & Xing, 2006; Kim, 2010). To be specific, the cutoff points on the θ scale correspond to the percentile for each score point are identified. According to the final reports of the past four years, on average, 85% of the examinees are classified as possibly qualified, 70% as qualified, 50% as well qualified, and 25% as proficient, the cutoff points matched on the θ scale are then approximately -1, -0.5, 0, and 0.7. That is to say, if the standard error of measurement of 0.32 or smaller is desired, the test information needs to be larger than 10 over the ability continuum from -1 to 0.7. Plotting the test information curves of tests consisting of 8, 10, and 12 operational items as shown in Figure 4.1, it can be observed that a test of 12 items is sufficient to meet the measurement precision criterion. However, it should be noted that due to the tailored nature of CATs, the accuracy of ability estimation will be greater than what is displayed here.

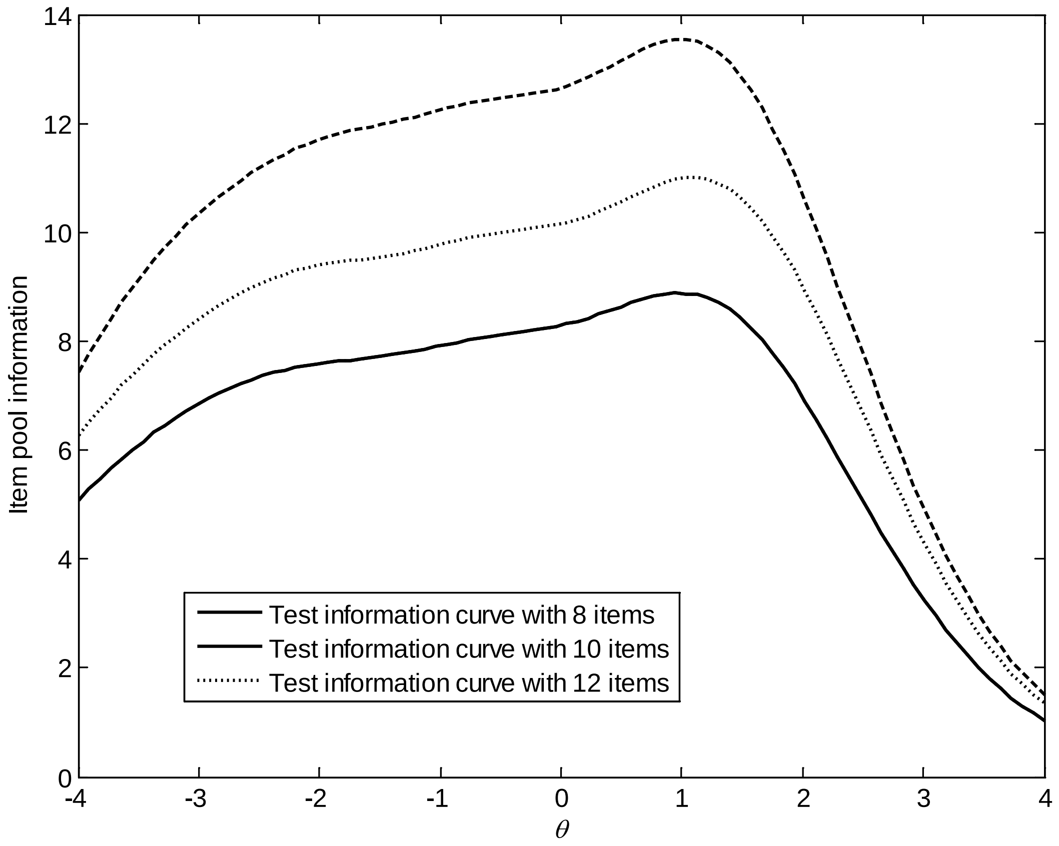


Figure 4.1 Test information curves of tests with 8, 10 and 12 items

CAT simulations conducted were based on the psychometric and practical considerations previously discussed. Unidimensionality of the two content areas is assumed. For all simulations, prior ability distribution was assumed to be normal *N*(0,1). Maximum item exposure rate is set to 0.20. Ability estimate after the last item was the final θ score. The general modeled CAT procedures are summarized in Table 4.3.

Table 4.3 Summary of Modeled CAT Simulation Design

CAT Components		Procedures
Item pool	Pool size	Master item pool: 1596 in Content 1, 1554 in Content 2
	Item parameters	Simulated parameters
Item selection	Initial selection	*U*(-0.4, 0.4)
	Interim selection	Fisher information
	Content balancing*	Rotation
	Exposure control*	Restricted maximum exposure rate of 0.20 *a*-Stratified
Ability estimation	Initial, interim, and final ability estimation	WLE and variable step size
Stopping rule	Fixed length	Maximum number of items: 12

Note. *Content balancing and exposure control apply as needed.

There are four conditions for simulations: CAT with content balancing and *a*-stratified constraints, CAT with content balancing control, CAT with *a*-stratified constraint, and unconstrained CAT. They are consistently referred to as Condition 1 to Condition 4 hereafter.

Condition 1 is the most constrained CAT simulation. Under this condition, content balancing was achieved by the rotation method (Davis et al., 2003). Combining the content balancing with the AS-MI method, item were administered as

C1S1, C1S1, C2S1, C2S1, C1S2, C1S2, C2S2, C2S2, C1S3, C1S3, C2S3, C2S3

where C denotes the content area and S represents the stratum.

For Condition 2, items were administered as

C1, C1, C1, C1, C1, C1, C2, C2, C2, C2, C2, C2,

For Condition 3, items were administered as

S1, S1, S1, S1, S2, S2, S2, S2, S3, S3, S3, S3.

4.2.2 Simulation Procedure

Simulations were used to design the optimal pool based on the design summarized in Table 4.3. MATLAB R2009 was used for programming. The simulation procedure is described below:

1) Generate the true θ values from *N*(0,1).

2) Select an initial ability estimate, $\hat{\theta}$, randomly from *U*(-0.4, 0.4). The first item is chosen as the most informative one from the corresponding $\theta_{.6}$ -bin.

3) Generate the simulees' responses to item *j* using the method as presented by Dodd and Koch (1987). Specifically, category responses, $P_{jk}(\theta)$, is calculated based on the true θ value. The CRF values are then summed cumulatively (i.e., $p_{j0}, p_{j0} + p_{j1}, p_{j0} + p_{j1} + p_{j2}$, $p_{j0} + p_{j1} + p_{j2} + p_{j3}, p_{j0} + p_{j1} + p_{j2} + p_{j3} + p_{j4},\ p_{j0} + p_{j1} + p_{j2} + p_{j3} + p_{j4} + p_{j5})$, which are compared to a random number generated from a uniform distribution [0, 1]. The score corresponds to the response category where the sum is larger than the random number.

4) Given the response to the item, $\hat{\theta}$ is updated using the WLE method. For the extreme category responses, a variable step size procedure proposed by Chen and Cook (2009) is used to modify $\hat{\theta}$. With the provisional estimate, the second item is selected from the remaining items in the corresponding subset using the AS-MI selection method if AS method applies. Otherwise, MI selection method is used.

5) The process of ability estimation and item selection repeats for one simulee until 12 items are administered. The final $\hat{\theta}$, items used, and test information are recorded.

During the process, if an item reaches the exposure rate 0.20, it is then excluded from administration. The simulation was conducted with a sample of 6000 simulees, and five replications were implemented for each simulation condition.

4.2.3 Determining Item Pool Size

During the simulation process, items that are most informative among the remaining items in *aθ*/*θ*-bins are selected. Because items from the *aθ*/*θ*-bin are supposed to be exchangeable based on the *p*-optimality criterion, post-simulation adjustment is conducted to ensure that redundant items from the same *aθ*/*θ*-bins were removed from the final pool. Resulting item pools are therefore optimal in a sense that a minimum number of items satisfying the CAT characteristics are included. The process to determine the number of items for the simulated pools is described below:

1) Recode a simulation record of items' unique indices into a corresponding *aθ*/*θ*-bins table;

2) Calculate the total number of items needed in each *aθ*/*θ*-bin under different simulation conditions. Based on the table obtained in the first step, a new item is added only when there is no adequate number of items for a simulee and the item exposure rate is smaller than 0.20. For instance, one simulee uses two items from a specific *aθ*/*θ*-bin, if there is already an item included in that bin, one new item is added if the existing item has not reached the predetermined maximum exposure rate of 0.20, otherwise two items are added.

3) According to the table obtained from the second step, items included in each *aθ*/*θ*-bin are determined based on their frequency of administration. For instance, if five items are needed for a specific *aθ*/*θ*-bin, the five most frequently used items are then selected and the less frequently used ones are trimmed out.

4) Post-simulation adjustment was made for each replication, and the results were then averaged across the replications. The items included formed the final optimal pools that are used for subsequent pool performance evaluation. In addition, the blueprint is a summary of the final pools.

The post-simulation procedure described above determines item pool sizes using the *p*-optimality criterion for designing item pools. In other words, items are included in the final pools in a way that they are informative for a given θ interval rather than for a specific θ point. Therefore, the resulting pools are optimal in the sense of *p*-optimality. Also, it can be clearly seen that the θ-bin width impacts the number of items to a large extent. More items are required if θ-bins are narrower and fewer if θ-bins are wider.

4.3 Extending Operational Item Pool

As mentioned earlier, only 16 operational items are available for calibration. Since it is much smaller than the simulated item pools, it is necessary to expand the operational pool so that both pools are of similar sizes. Item parameter replication (IPR) method for polytomous items is used for this purpose (Raju, Fortmann-Johnson, Kim, Morris, Nering & Oshima, 2009). The IPR procedure is described below:

1) Using the item parameter estimates and variances from the PARSCALE output, obtain two matrices of item parameter estimates $\boldsymbol{M}$ and their covariance matrix $\boldsymbol{V}$. Item parameter correlation matrix, $\boldsymbol{R}$, is then derived from the covariance matrix.

2) Decompose $\boldsymbol{R}$ as $\boldsymbol{R}$ = T'T.

3) Generate a vector $\boldsymbol{X}$ with the length equal to the number of response categories randomly from $N(0,1)$.

4) Obtain **Y** as Y = T'X.

5) Denote ***D*** as the diagonal matrix consisting of the diagonal elements in ***V***, replicated item parameter ***Z*** are obtained through ***Z*** = $\sqrt{D}$ ****Y*** + ***M***.

Repeat steps 3 to 5 as many times as replications are needed. In this case, each item was replicated eight times. The extended operational pool (EOP) consists of 144 items in total. When the EOP is used to compare its performance with the simulated pools, items are classified based on the varying simulation conditions accordingly.

4.4 Evaluating Item Pool Performance

Simulations were used to evaluate the performance of the simulated optimal pools (SOPs) and extended operational pool (EOP). The simulation procedure is the same as the one used for designing the item pool except that the item selection is no longer limited to the *aθ*-bins. Two types of distribution were used: 1) 6000 simulees randomly sampled from *N*(0,1) to evaluate the pool performance in general, and 2) 500 simulees at each of the 41 *θ* points from -4 to 4 in increments of 0.2 for an evaluation at a conditional level.

The comparison between operational and simulated pools reveals how they perform with respect to ability estimation precision and pool utilization. The evaluation criteria for ability estimation include Pearson product-moment correlation between the true (θ_i) and $\hat{\theta}_i$, bias, and root mean squared error (RMSE). The bias and RMSE are denoted

$$\text{Bias} = \sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)/n \qquad (4.1)$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i\,)^2/n} \qquad (4.2)$$

where n is the sample size.

In addition, for the operational test, students' overall performance falls into five levels, and the highest level is critical for course decisions. Therefore, classification accuracy for Level 5 is obtained to examine how different pools impact the results, which is measured by the percentage of the correct classification, the false-positive (FP) errors and the false-negative (FN) errors. Because the cutoff point for Level 5 is 0.7, treating the generated ability as the true ability, θ, the correct classification suggests that both the θs and $\hat{\theta}_s$ are higher than 0.70, FP errors occurred when the θs are lower than 0.70 but the $\hat{\theta}_s$ are higher than 0.70, and FN indicated that the θs are higher than 0.70 but the $\hat{\theta}_s$ are lower than 0.70.

For item pool utilization, the evaluation criteria are overall pool usage, percentage of items with varying exposure rate and test overlap rate. As Chang and Ying (1999) proposed, the efficiency of overall item pool usage can be measured by the discrepancy between the observed and expected item exposure rate. It follows χ^2 distribution and is denoted as

$$\chi^2 = \sum_{j=1}^{N} \frac{(r_j - L/N)^2}{L/N} \tag{4.3}$$

where r_j is the observed exposure rate for item j, L is the test length, N is the number of items in the item pool. A low χ^2 value implies that most of the items are fully used.

Item exposure rate is the ratio of the number of item administrations to the total number of examinees. The rate higher than 0.3 is regarded as overexposed (Segall, Moreno, & Hetter, 1997), and one lower than .02 is considered underexposed (Gu, 2007). Test overlap describes item exposure as well, and it has been used as item pool security index. Overlap rate is defined as the average proportion of items that two randomly selected simulees have in common (Way, 1998). For a fixed-length CAT, it is obtained by the formula (Chen, Ankenmann, & Spray, 2003)

$$\mathrm{T}_{\mathrm{overlap}} = \frac{\mathrm{n}\sum_{j=1}^{\mathrm{N}} \mathrm{r}_{\mathrm{j}}^{2}}{k(\mathrm{n}-1)} - \frac{1}{\mathrm{n}-1} \tag{4.4}$$

where r_j is the exposure rate of item *j*, n is the total number of simulees, *N* is the total number of items, and *k* is the number of items in the test. Test overlap lower than 15% is desired.

CHAPTER 5
ITEM POOL CHARACTERISTICS AND PERFORMANCE

This chapter first presents a summary of the extended operational pool (EOP) and simulated optimal item pools (SOP). The item pool characteristics under each condition are summarized and their performance is evaluated.

5.1 Item Pools Characteristics

Figure 5.1 shows the pool information curves of the EOP and SOPs. The impact of the practical constraints on the pool information is evident: 1) the pools designed without the *a*-stratified constraint are more informative than the pools with it across the entire ability continuum, and 2) when the *a*-stratified constraint was applied, the pool without content balancing control is slightly more informative than the one with the control.

Compared with the EOP, the pool information curves of the SOPs are smooth and somewhat bell shaped, indicating the items in the SOPs are more evenly distributed based on the maximum information they provide. The EOP is most informative at the left tail for the ability levels between -4 and -2.5, and least informative roughly from 0.4 to the right end. Additionally, the EOP is more informative than the SOPs designed with the *a*-stratified constraint for the ability levels below 0.4, but less informative above 0.4. The EOP is less informative than the SOPs designed without the *a*-stratified constraint for the ability levels above -2.5.

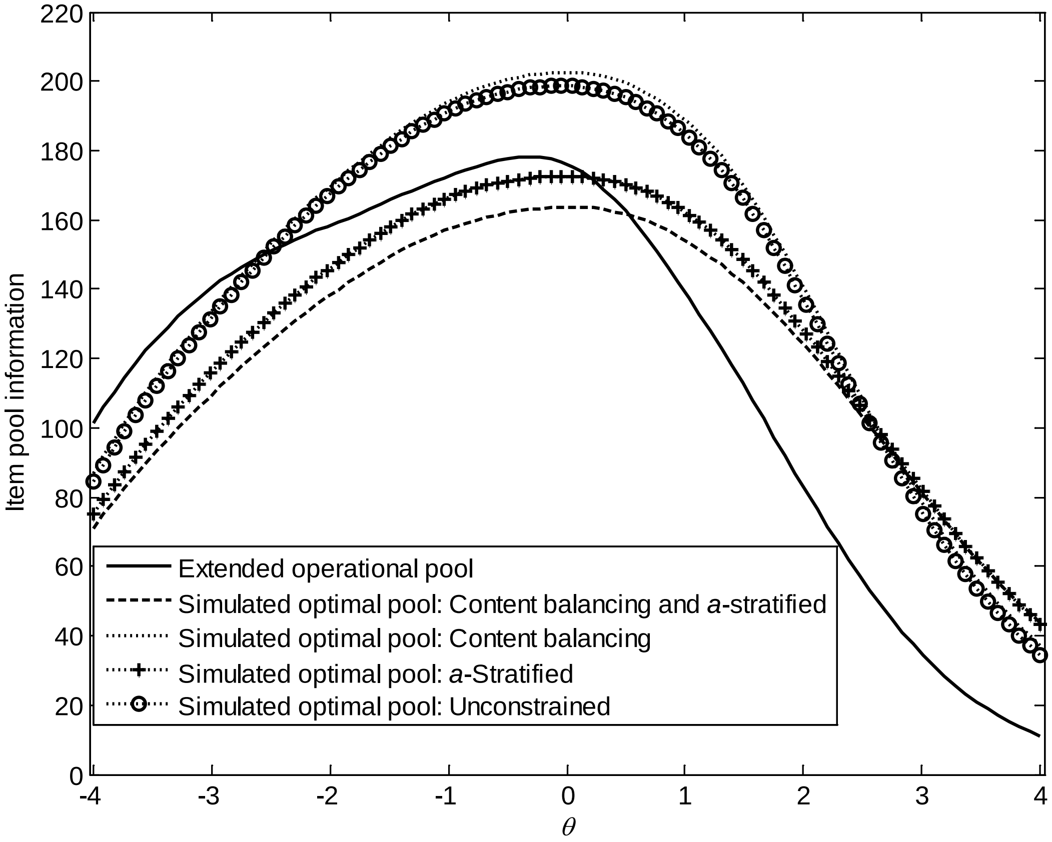


Figure 5.1 Pool information curves of the EOP and SOPs

Table 5.1 shows the descriptive statistics of the ***a***- and ***b***-parameter. The distribution of the ***a***-parameter of the EOP and the SOPs with the *a*-stratified constraint is similar. The SOPs without the constraint have larger ***a***-parameter on average, but the variations are comparable. The mean of the ***b***-parameter in the EOP is much smaller than those in the SOPs, and the ***b***-parameter in the SOPs shows greater variations. This is because the SOPs include more easy and difficult items. The maximum information the items provide is also summarized in Table 5.1. On average, the pools designed with the *a*-stratified constraint have the lower MI, and those designed without the constraint have the higher MI. The EOP contains the most informative items.

Table 5.1 Statistics of Discrimination, Location Parameters, and Maximum Information

			a				***b***			**Maximum Information**			
Pools	**Pool Size**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Extended Operational Pool	144	0.89	0.14	0.64	1.14	-1.08	0.40	-1.83	-0.50	1.30	0.43	0.71	2.37
Simulated Optimal Pool: Condition 1*	144	0.91	0.14	0.67	1.10	-0.25	1.05	-1.91	1.97	1.24	0.27	0.75	1.83
Simulated Optimal Pool: Condition 2*	150	1.02	0.12	0.56	1.10	-0.42	0.98	-1.88	2.00	1.46	0.28	0.60	1.83
Simulated Optimal Pool: Condition 3*	151	0.90	0.14	0.59	1.10	-0.32	1.01	-1.90	1.98	1.24	0.26	0.80	1.83
Simulated Optimal Pool: Condition 4*	147	1.01	0.13	0.55	1.10	-0.43	0.95	-1.88	1.96	1.46	0.28	0.54	1.83

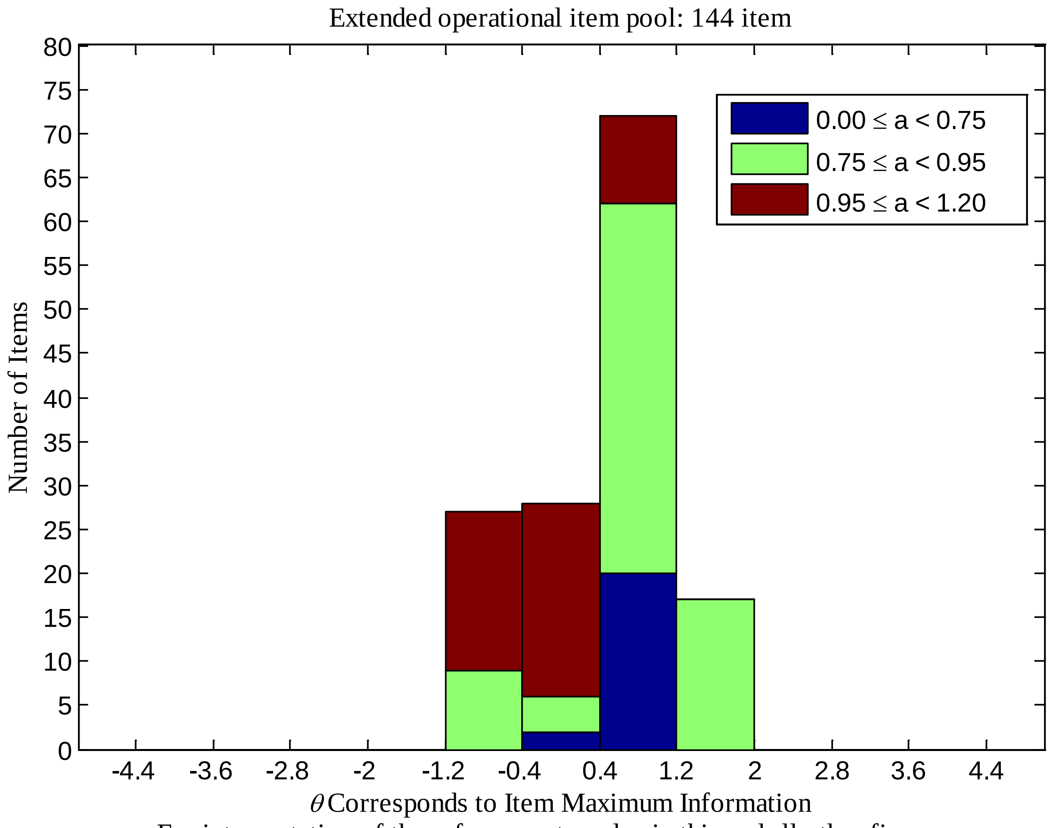
Note. *Condition 1: Content balancing and *a*-stratified constraints.

Condition 2: Content balancing control.

Condition 3: *a*-Stratified constraint.

Condition 4: Unconstrained.

While the items included in the SOPs are not the same under the four conditions, the EOP remains the same. To compare their performances, the EOP is split into different subsets based on the conditions the simulated pools are designed. Figure 5.2 shows the item distribution of the EOP based on the ***a***-parameter and where the MI is located. Because the operational test is not an adaptive one, Figure 5.2 illustrates that the items reach their maximum information for the ability levels ranging from -1.2 to 2. In addition, approximately 15% of the items have the ***a***-parameter smaller than 0.75, 50% have the ***a***-parameter ranging from 0.75 to 0.95, and 35% have the ***a***-parameter larger than 0.95. The detailed item distributions and descriptive statistics of the item parameters are presented in Appendix A.1 and A.2.



For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation

Figure 5.2 Item distribution of the EOP

5.2 Item Pool Designed with Content Balancing and *a*-Stratified Control

5.2.1 Item Pool Characteristics

Figure 5.3 shows the item distribution for each content area of the SOP under Condition 1. The detailed item distribution and descriptive statistics of item parameters are summarized in Appendix A.3 to A.6.

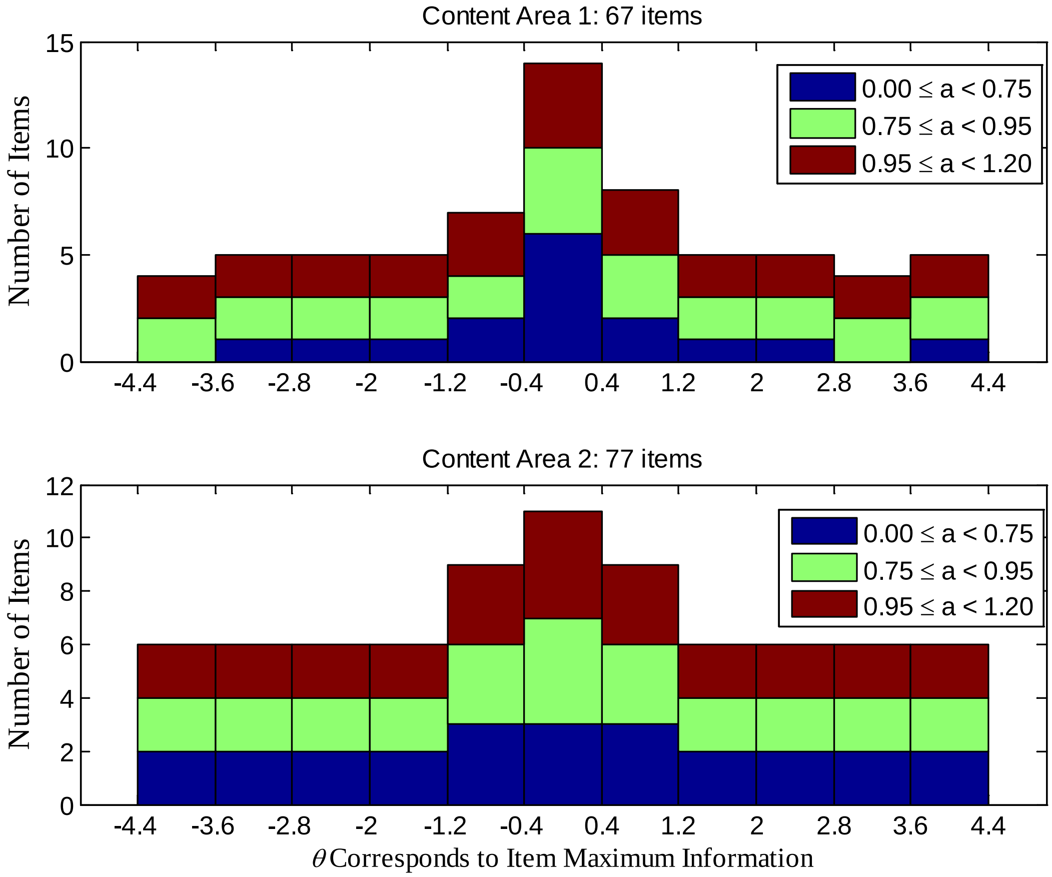


Figure 5.3 Item distributions for the SOP under Condition 1

As can be seen in Figure 5.3, the items are evenly distributed across the θ levels between -4.4 and 4.4. Furthermore, the item distributions in two content areas are not the same. To start with, since the simulated CAT administration starts from content area one with an initial ability randomly selected from *U*(-0.4, 0.4), more items which are informative for this θ range are included, especially in the first stratum. In addition, in content area one, there are fewer items in the first stratum than in other two strata. For content area two, the items are quite evenly distributed across the ability scale with respect to where their maximum information is located. For both content areas, slightly more items are included between the ability levels -1.2 and 1.2, which is consistent to the specified ability distribution of the target population. Also, as there are 15% items with low discrimination in the EOP, comparatively, there are 24% such items in the SOP.

5.2.2 Item Pool Performance

To compare the pool performance under the most constrained situation, the EOP was first divided into three strata with the *a*-parameter range identical to that of the SOP. Within each stratum, an equal number of items are randomly grouped into two contents.

The evaluation results of the ability estimates and pool utilization for the two pools under Condition 1 are presented in Table 5.2. Table 5.3 shows the classification results for Level 5 and the distribution of the true θ and θ estimates for each classification category. The ability estimates from the EOP show positive bias, and those from the SOP are negative. However, the magnitudes of the bias are negligible. RMSE from the SOP is slightly smaller than that of the EOP. The correlation between the true and estimated θ is the same, and both are significant. The SOP also resulted in higher average test information.

As Table 5.3 shows, the SOP yielded a higher percentage of correct classification than the EOP did, 69% versus 66%. For this category, the distribution of the $\hat{\theta}$ is close to that of the true θ. The FP and FN errors are 16% and 15% for the SOP versus 20% and 14% for the EOP. For these two classification categories, the distributions of the $\hat{\theta}$ deviated greatly from those of the true θ.

Table 5.2 Summary Statistics of Item Pools Performance: Condition 1

Statistic	**Extended Operational Pool**	**Simulated Optimal Pool**
Bias	0.0024	-0.002
RMSE	0.37	0.36
Correlation	0.94	0.94
Test information	14.85	14.92
χ^2 of item exposure rate	10.05	10.66
Items with exposure rate equals 0.20	31%	25%
Items with exposure rate between .02 and 0.2	21%	33%
Items with exposure rate < .02	18%	27%
Items that are not used	30%	15%
Test overlap rate	0.18	0.17
Pool size	144	144

Table 5.3 Summary Statistics of Classification Results for Level 5: Condition 1

	Correct Classification	Distribution of True θ	Distribution of $\hat{\theta}$	FP Errors	Distribution of True θ	Distribution of $\hat{\theta}$	FN Errors	Distribution of True θ	Distribution of $\hat{\theta}$
EOP	66%	1.36*	1.40*	20%	0.47*	0.96*	14%	0.93*	0.48*
		(0.50)	(0.54)		(0.19)	(0.21)		(0.19)	(0.17)
SOP	69%	1.35*	1.40*	16%	0.46*	0.93*	15%	0.89*	0.48*
		(0.47)	(0.51)		(0.19)	(0.20)		(0.17)	(0.17)

Note. *The first number is the mean and in the parenthesis is the standard deviation.

As shown in Table 5.2, χ^2 value of the item exposure rate of the EOP is slightly smaller than that of the SOP. Examining the percentage of the items that are fully used, well used, and not used, it seems that this occurred because of the high percentage of the items that reached the preset maximum exposure rate of 0.20. More specifically, there are 31% of the items that are fully used in the EOP versus 25% in the SOP. The EOP has 18% items that are rarely used in comparison to 27% in the SOP. Finally, the EOP has a much larger proportion of items that is not used than the SOP, 30% versus 15%. That is to say, while there are 52% of the items in the EOP are reasonably well used, 58% are well used in the SOP. A detailed item exposure rate of the two pools is graphed below. Table 5.2 also shows that the SOP has a lower test overlap rate.

Describing items based on where the maximum information is located, Figures 5.4 and 5.5 plot the item exposure rate for the simulation with 6000 simulees. Because the maximum exposure rate 0.20 was applied along with the *a*-stratified method, the rate varies from 0.0 to 0.20. As depicted in Figure 5.4, for the EOP, the distribution of the exposure rate for each stratum is similar for the two content areas. The items are informative for the ability levels between -1.2 and 1.5, and the fully exposed items spread approximately between -1.15 and 1.15. Because there are a small number of items in the first stratum, those items are well used. Many items in the second stratum that are informative for the ability levels between 0.5 and 1.5 are rarely or never administered. 35% of the items in the EOP are classified into the third stratum, and these items in the content area one are better used than those in the content area two.

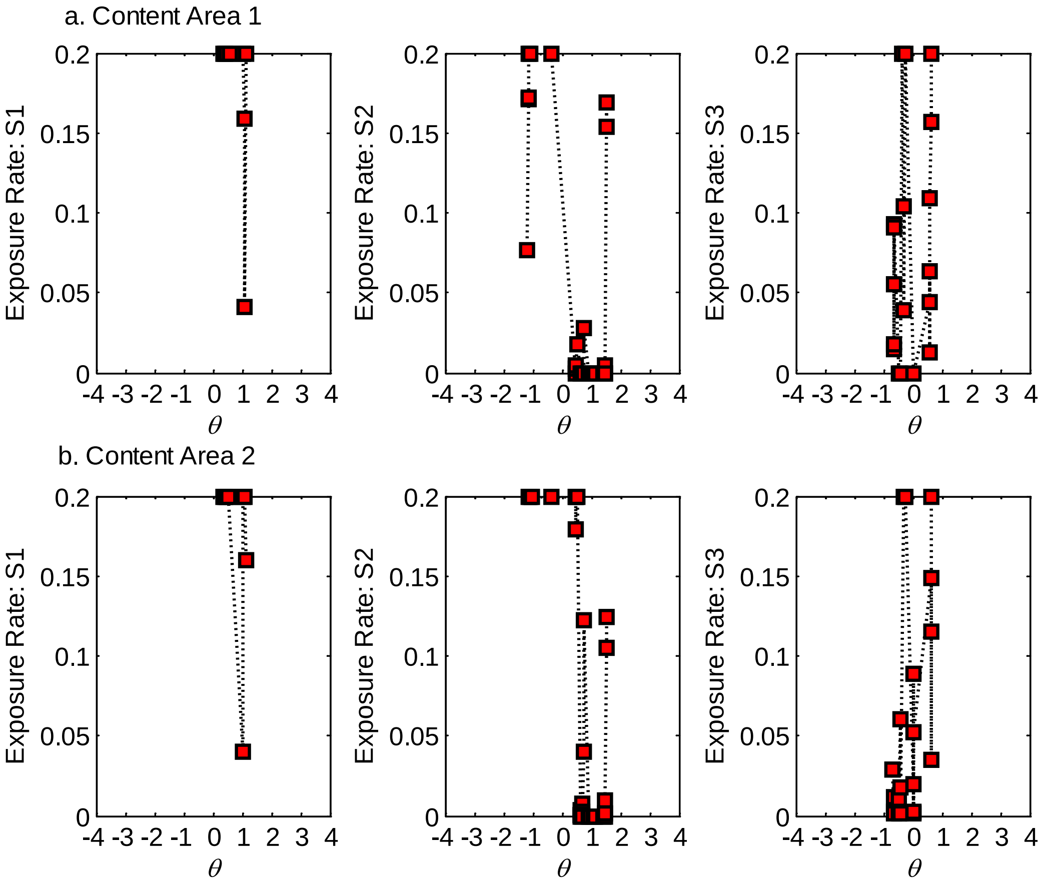
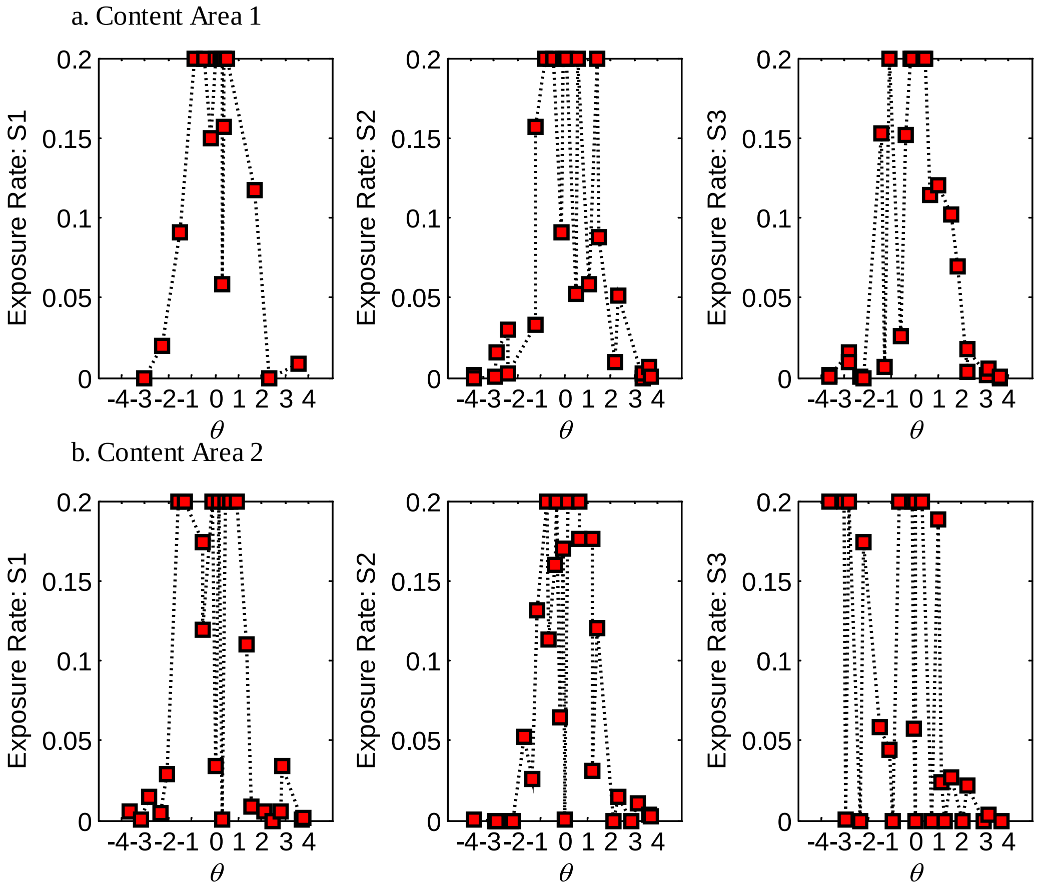


Figure 5.4 Item exposure rate of the EOP under Condition 1

Figure 5.5 illustrates that for the SOP, the distribution of the item exposure rate is similar at the first and second stratum, but differs at the third for two content areas. For content area one, the items which reach their MI between the θ levels -2.0 and 2.0 are used more than those whose MI is located at the tails. For content area two, not surprisingly, the items whose MI is located in the middle are still well used. However, a few items which are informative for the θ values between -3.6 and -2.0 are also fully used. For both content areas, the items that are informative for the extreme θ values tend to be under-exposed.


a. Content Area 1

b. Content Area 2

Figure 5.5 Item exposure rate of the SOP under Condition 1

Figures 5.6 to 5.10 plot the conditional evaluation indices of the θ estimates based on the simulation for 500 simulees on the 41 fixed θs.

Figures 5.6 and 5.7 present the conditional test information and standard error of measurement (SEM) respectively. Figure 5.6 indicates that the SOP provides test information consistently above the target information level, 10.0, even at the extreme θ values. The EOP provides test information greater than 10.0 between the θ levels -4 and 2.0, and smaller than 10.0 roughly beyond the θ value 2.0. The EOP provides higher test information between the ability levels -2.2 and 0.2. Correspondingly, Figure 5.7 demonstrates that for the θ levels above 0.60, the conditional SEM is consistently smaller for the SOP. For both pools, the test information is the greatest when the θ equals -0.4, which is desirable because it roughly corresponds to the cut-off point, -0.5, for proficiency classification level 4.

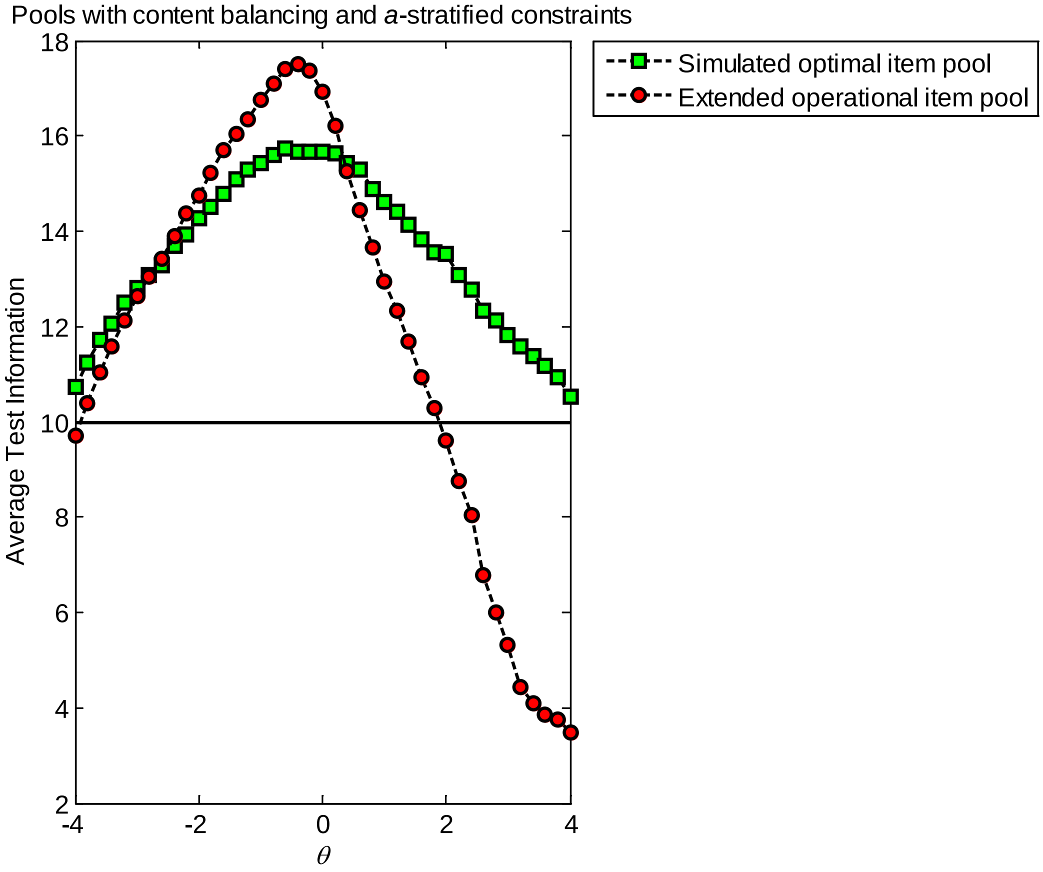


Figure 5.6 Average test information conditional on θ

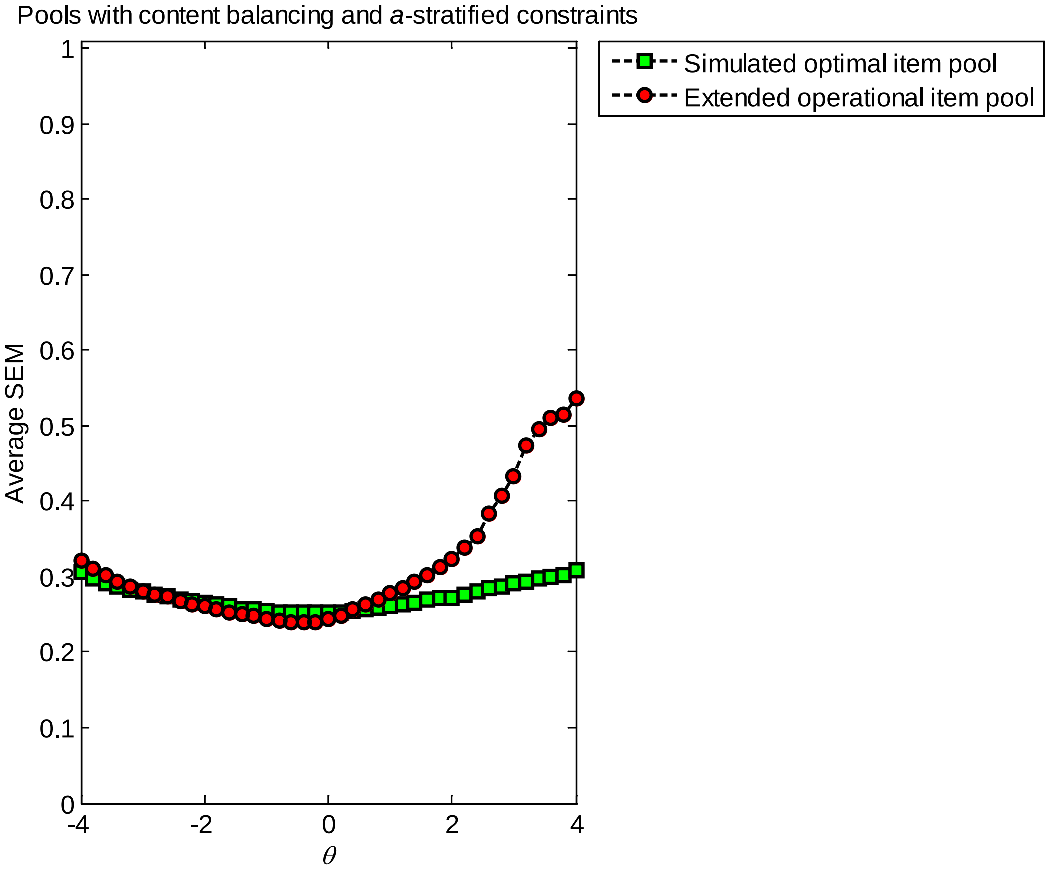


Figure 5.7 Conditional standard error of measurement (CSEM)

Figure 5.8 plots the average conditional bias for both pools. For the EOP and SOP, there is positive bias at the low θ levels approximately ranging from -4.0 to -1.8, and negative bias at the high θ levels roughly above 1.8. In general, for the θ levels between -1.8 and 1.8, positive bias is observed in the EOP while negative bias in the SOP. With regard to RMSE, Figure 5.9 shows that the SOP results in higher RMSE for the θ levels between -2.8 and 2.2, and lower RMSE for the θ values at the tails.

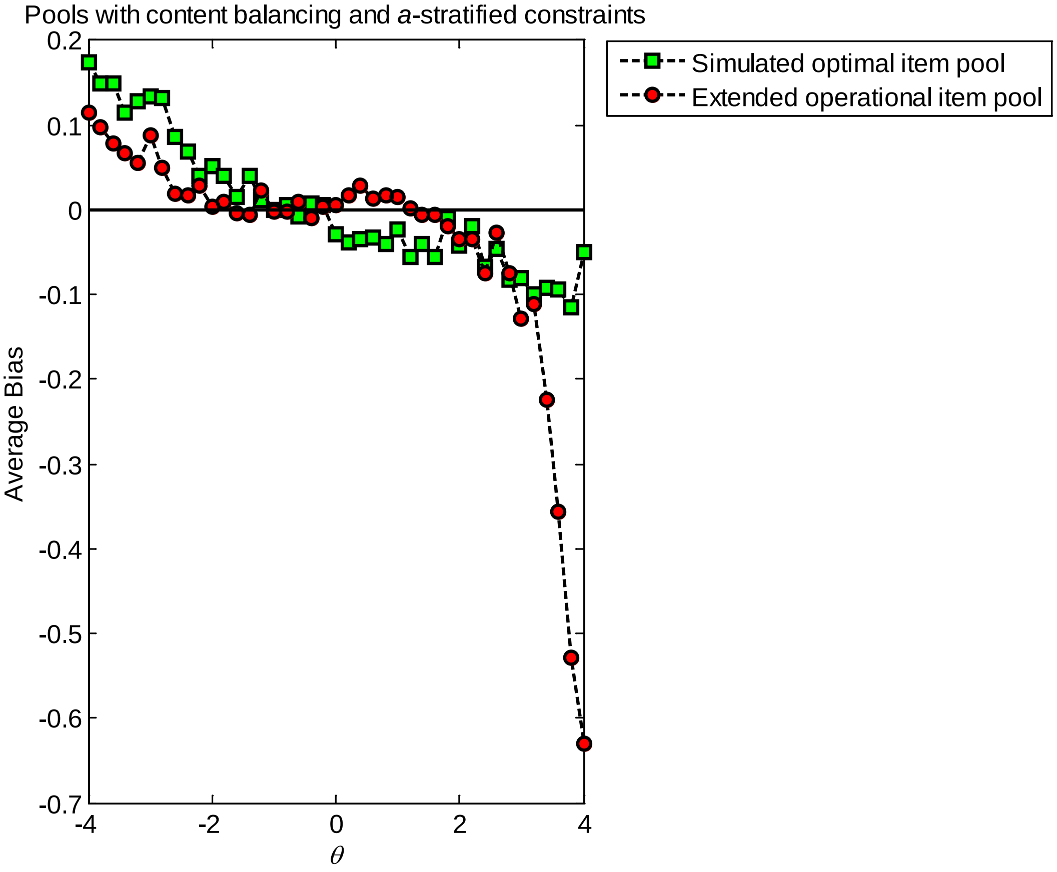


Figure 5.8 Conditional bias

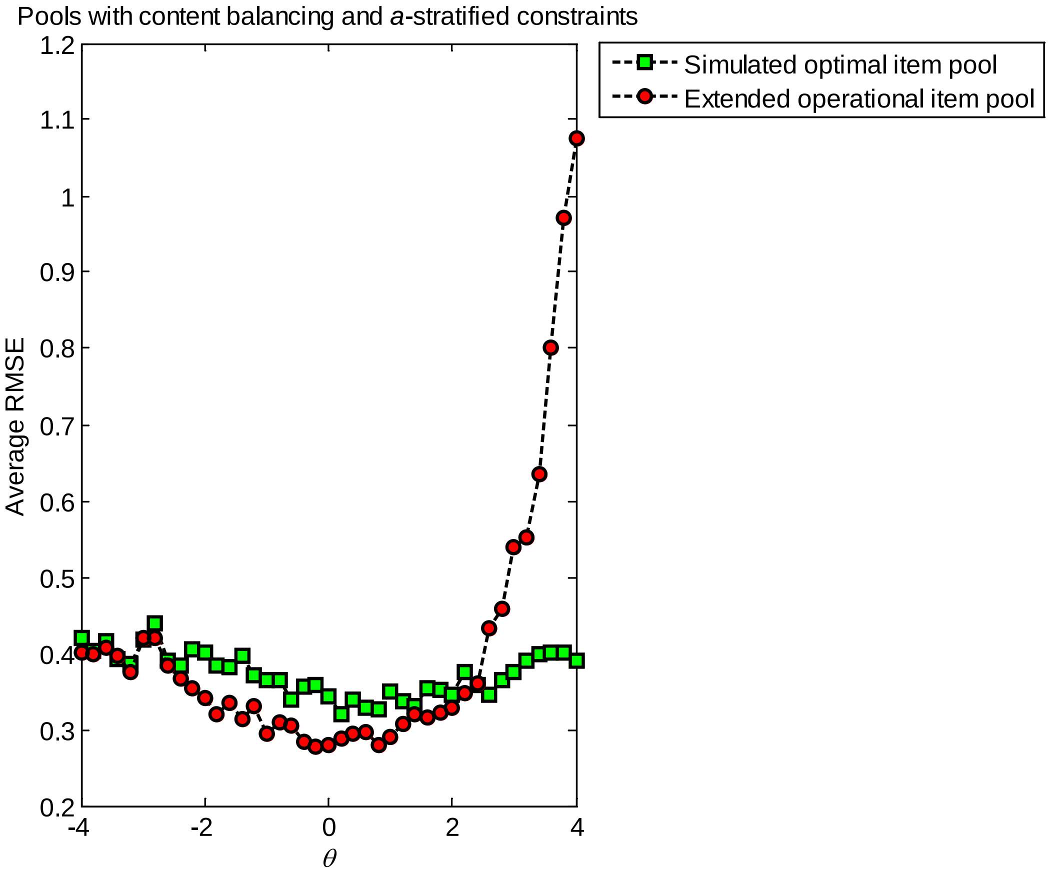


Figure 5.9 Conditional RMSE

Figure 5.10 graphs the average test overlap rate conditional on θ. Because item exposure control methods of *a*-stratified and restricting maximum exposure rate are applied, the test overlap rate is quite stable for both pools between 0.45 and 0.70 across the θ scale. The plot shows that the SOP has a consistently smaller test overlap rate across the θ levels between -3.6 and 4. Furthermore, for the EOP, the test overlap rate is the highest between the θ levels -1.8 and 0.4 and at the right end. For the SOP, the test overlap rate is the highest between the θ levels -1.4 and 0.2 and at the both ends.

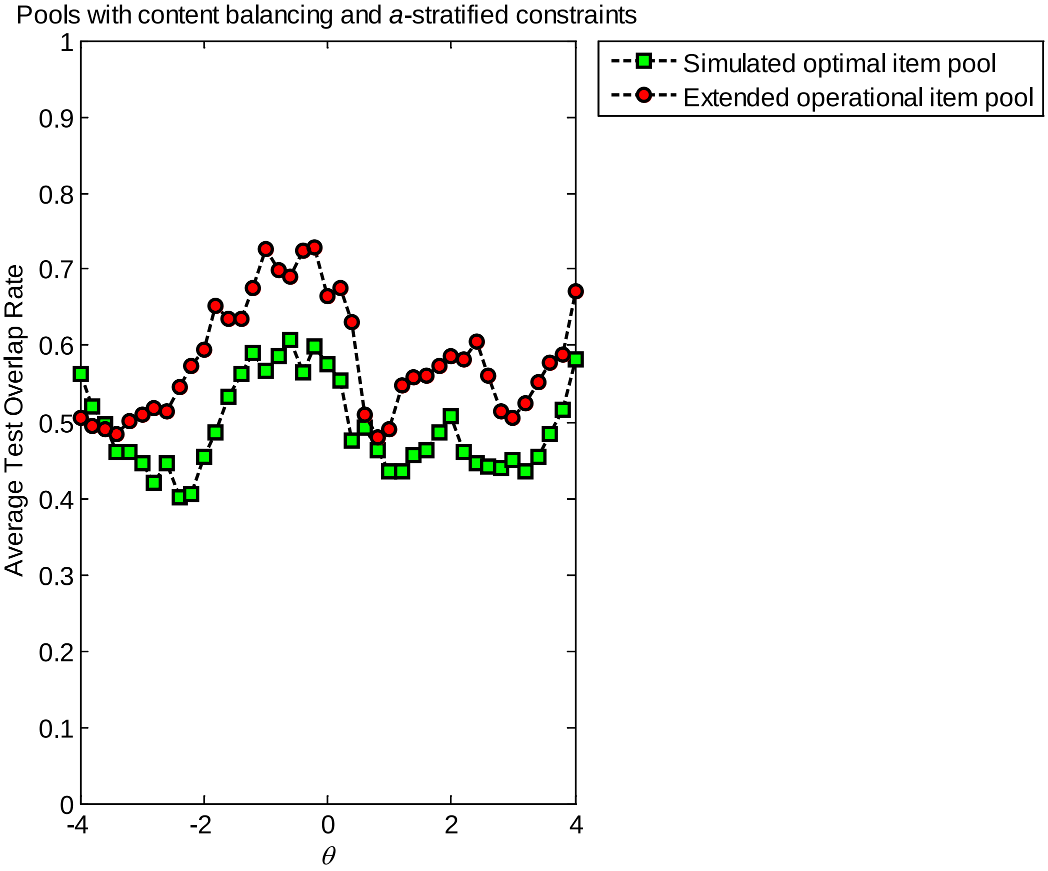


Figure 5.10 Test overlap rate conditional on θ

In summary, the results suggest that the optimal item pool designed under Condition 1 outperforms the extended operational pool in the following aspects: 1) it provides test information above the target level across the entire θ range; 2) it has fewer items that are not used; 3) it has smaller test overlap rate across the θ range.

With respect to the item pool characteristics, the EOP and SOP contain an equal number of items. The ***a***-parameter in the two pools is in the range from 0.60 to 1.15. The ***b***-parameter covers the range from -1.80 to 0.50 in the EOP, and from -1.90 to 1.97 in the SOP. In addition, the EOP has few items that are informative for the θ levels at the right end. Also, the SOP contains items that are evenly distributed based on the maximum information they provide.

Although some of the evaluation indices are similar on average, it should be noted that the SOP has more items with low discrimination, and the average maximum information the items provide is less than those in the EOP. Overall, it seems that the optimal item pool designed with the *p*-optimality method performs reasonably well even with more items of low discrimination.

5.3 Item Pool Designed with Content Balancing Control

5.3.1 Item Pool Characteristics

Figure 5.11 graphs the item distributions of the SOP under Condition 2. The detailed item distribution and descriptive statistics of item parameters are presented in Appendices A.7 to A.10.

The most notable characteristic of the pool designed under this condition is a large proportion of items with large ***a***-parameter. Specifically, 88% of the items have the ***a***-parameter ranging from 0.95 to 1.10. Also, the mean of the ***a***-parameter is greater than that of the EOP, 1.02 in comparison to 0.89. The average maximum information is larger as well, 1.46 versus 1.30. In addition, similar as the pool designed under Condition 1, the number of items in the two content areas is not the same, 71 in the first and 79 the second, but they are evenly distributed across the θ scale. Corresponding to the normal ability distribution, more items are included between the ability levels -1.2 and 1.2 for both content areas, but the number of items needed at the tails is nearly the same.

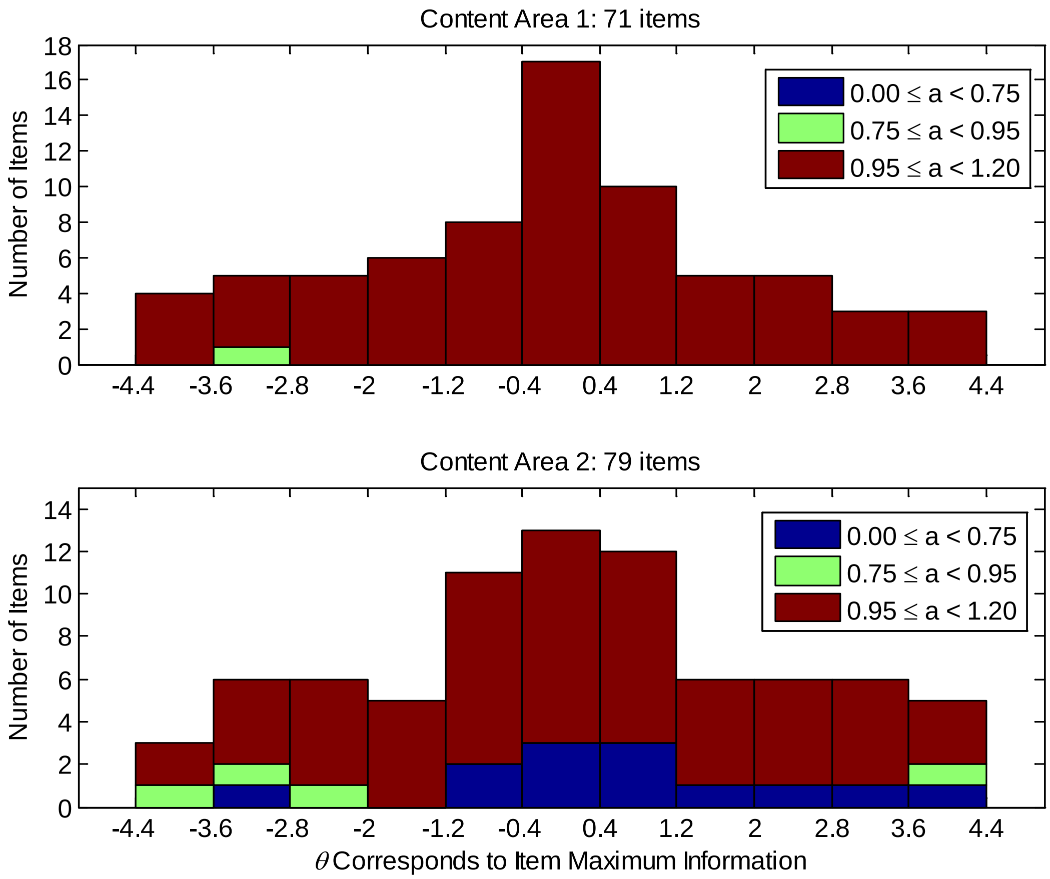


Figure 5.11 Item distributions for the SOP under Condition 2

5.3.2 Item Pool Performance

To evaluate the pool performance under this condition, the EOP was randomly grouped into two content areas of the same size. Table 5.4 and 5.5 presents the evaluation results of the two pools under Condition 2.

The ability estimates from both pools show small negative bias. The EOP resulted in a smaller RMSE. The correlation between the true θ and θ estimates for both pools does not indicate significant differences. However, the average test information from the SOP is higher than the EOP. This occurs probably because the EOP contains the most informative items.

Table 5.4 Summary Statistics of Item Pools Performance: Condition 2

Statistic	Extended Operational Pool	Simulated Optimal Pool
Bias	-0.02	-0.01
RMSE	0.27	0.31
Correlation	0.96	0.95
Test information	17.97	18.73
χ^2 of item exposure rate	6.86	9.10
Items with exposure rate equals 0.20	31%	25%
Items with exposure rate between .02 and 0.20	22%	30%
Items with exposure rate < .02	22%	28%
Items that are not used	25%	17%
Test overlap rate	0.17	0.17
Pool size	144	150

Table 5.5 Summary Statistics of Classification Results for Level 5: Condition 2

	Correct Classification	Distribution of True θ	Distribution of $\hat{\theta}$	FP Errors	Distribution of True θ	Distribution of $\hat{\theta}$	FN Errors	Distribution of True θ	Distribution of $\hat{\theta}$
EOP	74%	1.38*	1.38*	12%	0.51*	0.86*	14%	0.87*	0.52*
		(0.48)	(0.51)		(0.16)	(0.14)		(0.15)	(0.14)
SOP	72%	1.39*	1.40*	14%	0.50*	0.91*	14%	0.88*	0.50*
		(0.49)	(0.52)		(0.17)	(0.17)		(0.15)	(0.15)

Note. * The first number is the mean and in the parenthesis is the standard deviation.

With regard to the classification accuracy for Level 5, the percentage of correct classification is higher for the EOP than for the SOP, 74 versus 72. For this classification category, the $\hat{\theta}$ distribution is close to the true θ distribution. The FP and FN errors are 14% and 14% for the SOP versus 12% and 14% for the EOP. The mean of the $\hat{\theta}$ is much larger than that of the true θ for the FP category, and the mean of the $\hat{\theta}$ is much smaller than that of the true θ for the FN category.

In regards to the item pool usage, Table 5.4 shows that χ^2 value of the item exposure rate of the EOP is smaller than that of the SOP. Similar as what was observed for the pools under Condition 1, this occurred because of the high percentage of the items that met the maximum exposure rate of 0.20. In particular, while 31% of the items are fully used in the EOP, there are 25% in the SOP. Yet again, the SOP has a smaller percentage of items that are never administered. In total, 55% of the items in the SOP and 52% in the EOP have the exposure rate larger than 0.02.

Figures 5.12 and 5.13 plot the item exposure rate for the simulation with 6000 simulees. As Figure 5.12 indicates, the item exposure rate distribution of the two content areas in the EOP is similar. The items that are informative between the ability levels -0.7 and 0.7 are better used than the items at the ends.

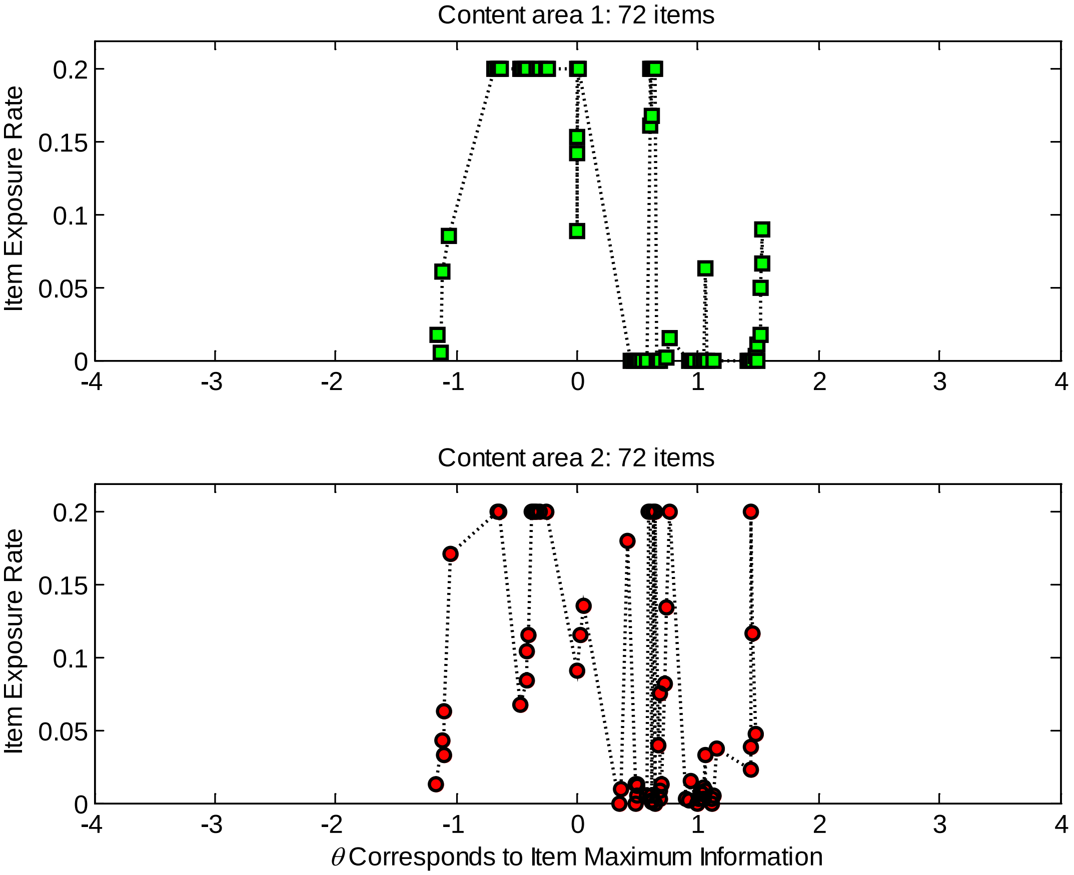


Figure 5.12 Item exposure rate of the EOP under Condition 2

Figure 5.13 shows the item exposure rate distribution of the SOP. For both content areas, the items that are informative between the ability levels -1.2 and 1 are well used. Additionally, several items that reach their MI at the lower end are highly exposed. Compared with Figure 5.12, it is apparent that fewer items are not administered.

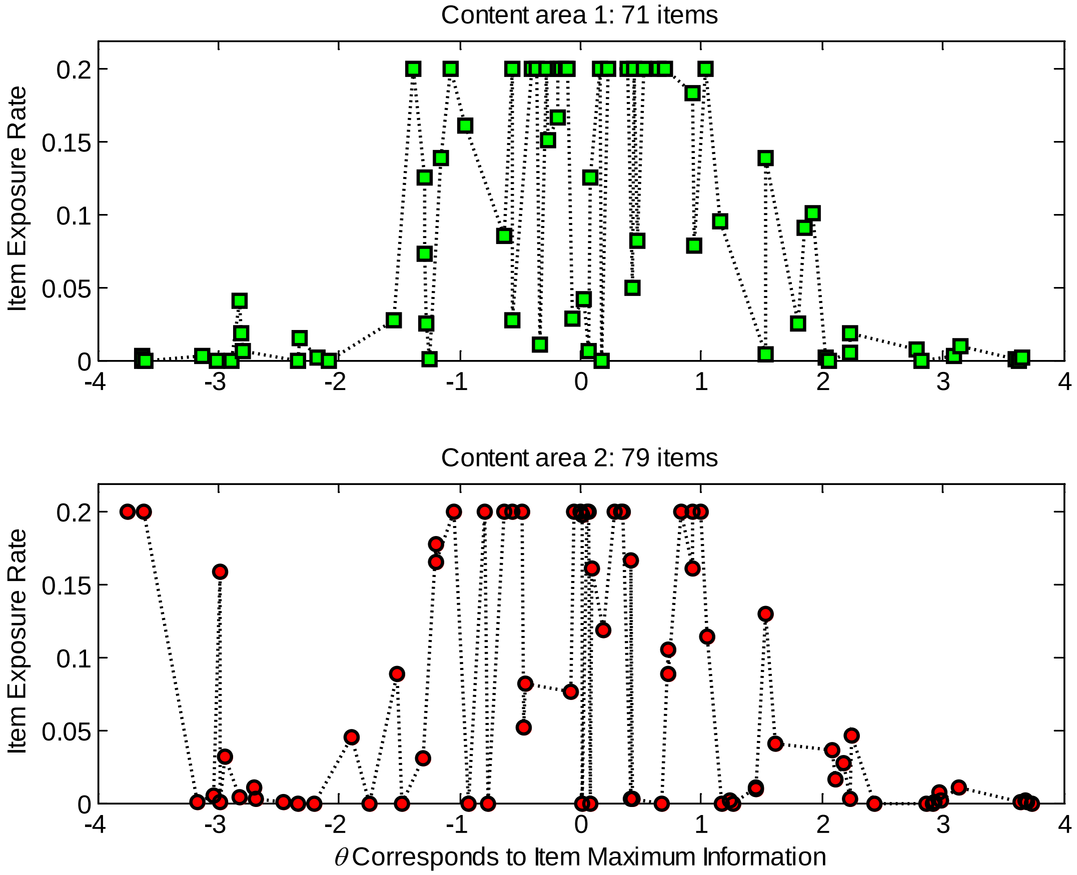


Figure 5.13 Item exposure rate of the SOP under Condition 2

Figures 5.14 to 5.18 plot the conditional indices of the θ estimates. Figures 5.14 and 5.15 display the test information and SEM. Without the *a*-stratified constraint, Figure 5.14 shows that the SOP provides test information consistently higher than the target level across the entire θ scale. Correspondingly, the CSEM is constantly smaller than 0.28, resulting in conditional reliability larger than 0.92. For the EOP, the measurement precision does not reach the desired level of 10.0 above the θ point 2.0. The EOP provides less test information than the SOP except for the ability levels between -3.4 and -0.2. Finally, the EOP did not achieve the desired measurement precision for the ability levels above 2.0.

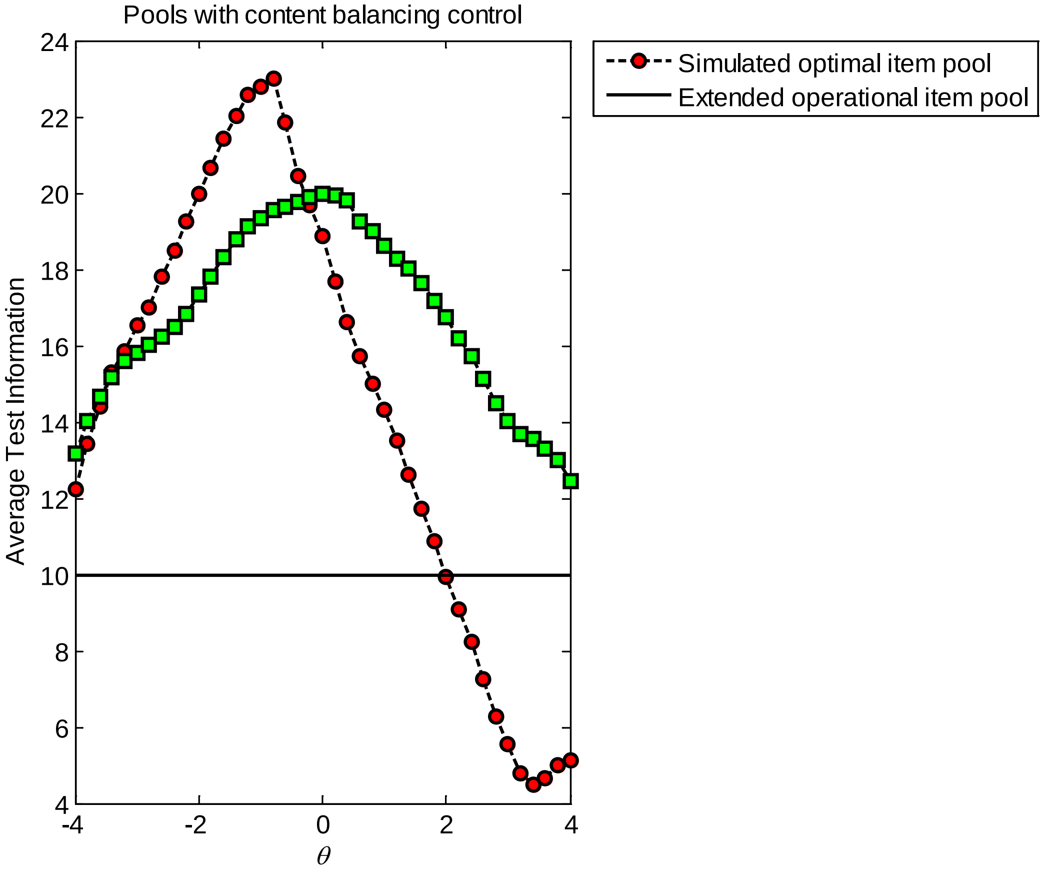


Figure 5.14 Average test information conditional on θ

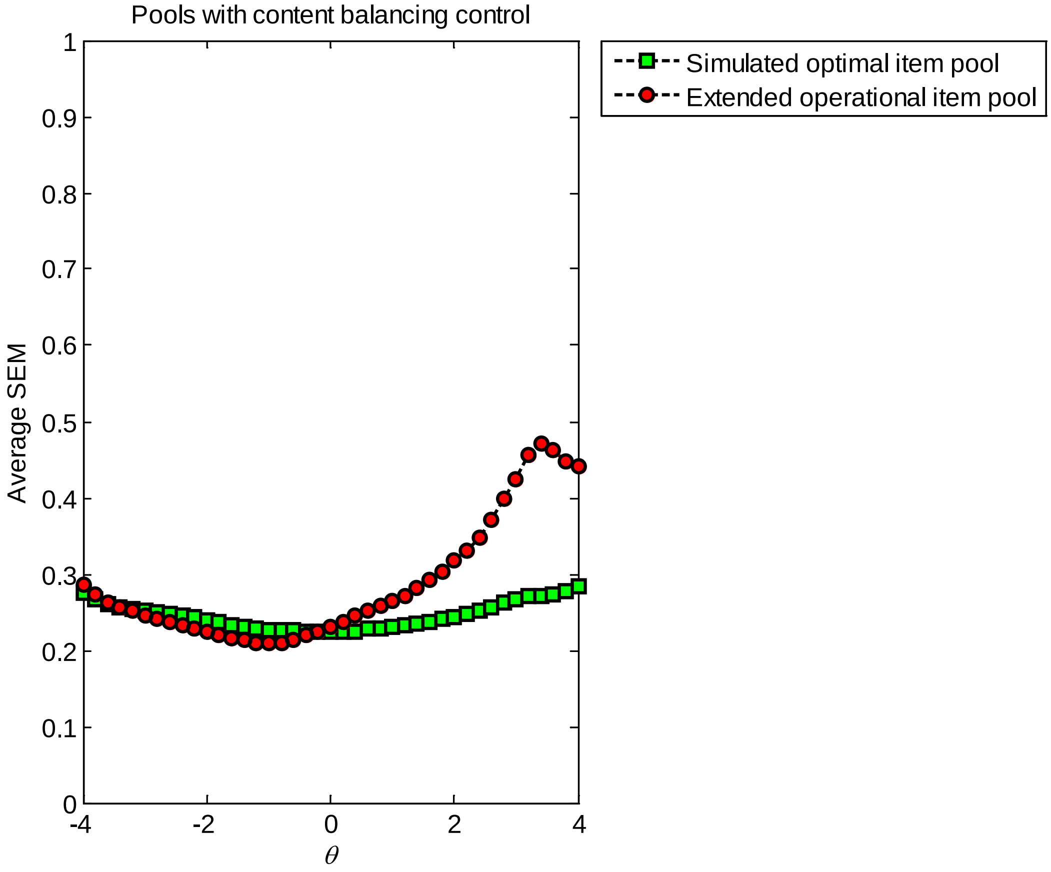


Figure 5.15 Conditional standard error of measurement

Figure 5.16 plots the average conditional bias for both pools under Condition 2. For both the EOP and SOP, there is positive bias at the ability levels roughly between -4 and -2, and negative bias above 0.2. As shown in Figure 5.17, the SOP shows consistently lower RMSE through the θ levels.

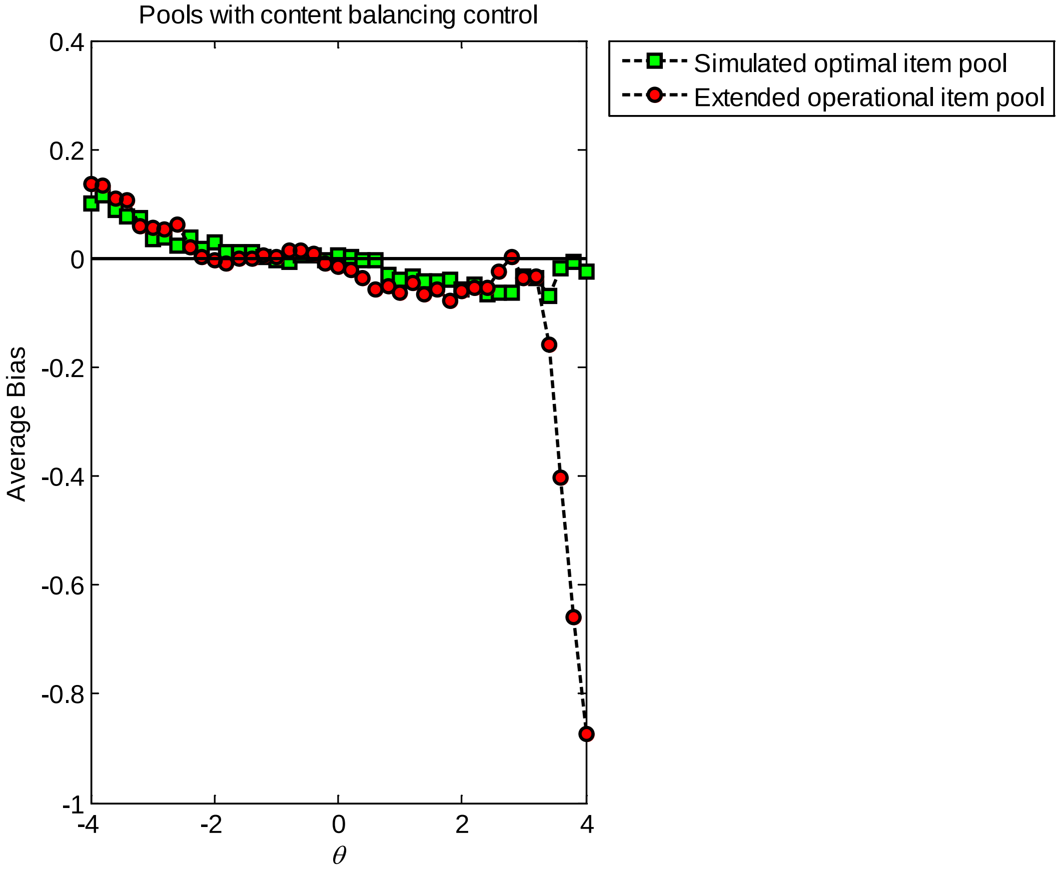


Figure 5.16 Conditional bias

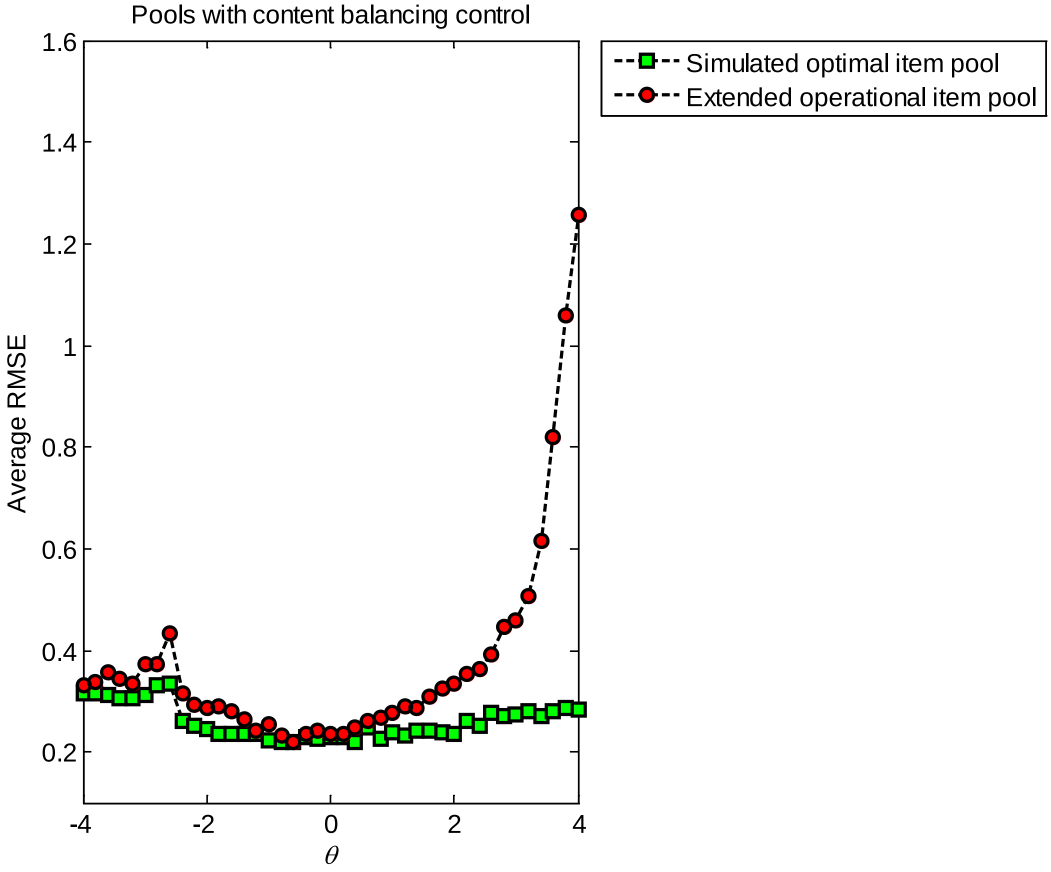


Figure 5.17 Conditional RMSE

Figure 5.18 graphs the conditional test overlap rate. Removing the *a*-stratified constraint results in a high item overlap rate across the entire ability level. Except for the ability levels between -2 and -0.6, the EOP has a smaller test overlap rate.

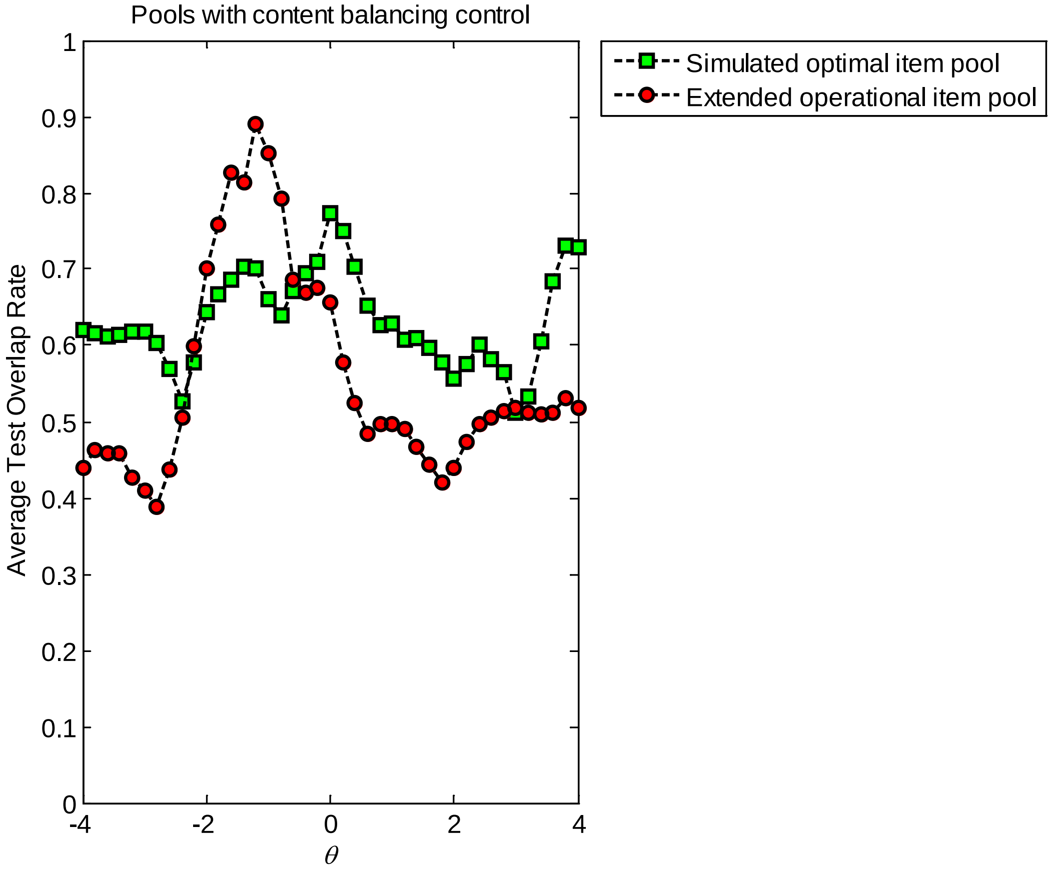


Figure 5.18 Test overlap rate conditional on θ

To summarize, the results suggest that the SOP designed under Condition 2 performs better than the EOP in several ways: 1) it provides test information higher than the target level across the ability levels; 2) it results in a consistently smaller RMSE; 3) it has fewer items that are never administered.

With respect to the item pool characteristics, the SOP has a slightly larger size than the EOP. As a result of the MI item selection method, the mean of the ***a***-parameter and MI is greater in the SOP. The ***b***-parameter in the EOP is in the range from -1.83 to -0.50, and the SOP shows a greater variation ranging from -1.88 to 2.00. This suggests that the EOP includes few informative items for the high θ levels.

5.4 Item Pool Designed with *a*-Stratified Constraint

5.4.1 Item Pool Characteristics

Figure 5.19 plots the item distributions of the SOP under Condition 3. The detailed item distribution and descriptive statistics of item parameters are summarized in Appendix A.11 and A.12.

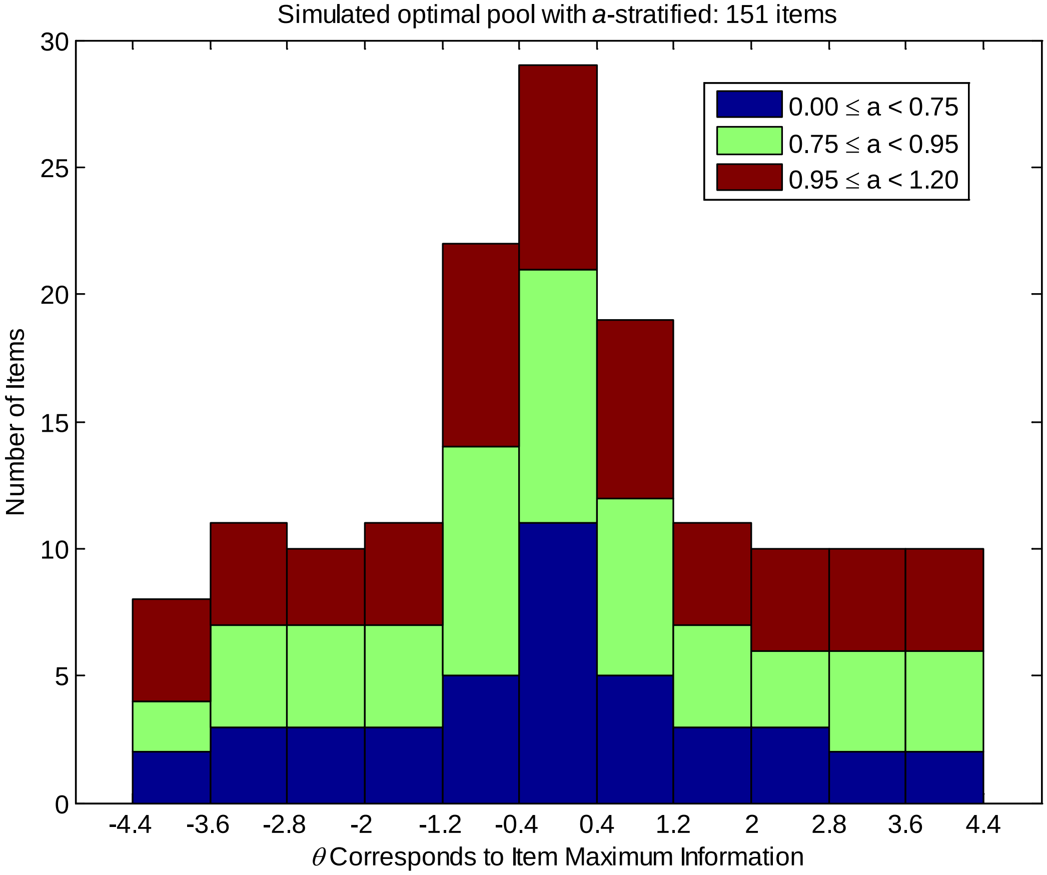


Figure 5.19 Item distributions for the SOP under Condition 3

As shown in Figure 5.19, the SOP includes items that are informative for the ability levels between -4.4 and 4.4. There are fewer items at the first stratum, 28%, in comparison to 36% in the second and third stratum. Once again, this is because the CAT simulations start from the first stratum. In addition, while more items that are informative at the middle range are needed when the normal ability distribution is assumed, the items that are informative for both tails are distributed symmetrically.

In regards to the pool characteristics of the SOP under this condition, the distribution of the ***a***-parameter is similar to that of the EOP, but the ***b***-parameter shows a greater variation with the inclusion of informative items at the right tail. With the *a*-stratified constraint, the mean maximum information provided by the items in the SOP is smaller than those in the EOP.

5.4.2 Item Pool Performance

For the purpose of the performance evaluation, the EOP was divided into three strata based on the ***a***-parameter values. Table 5.6 and 5.7 presents the evaluation results of the two pools.

Table 5.6 Summary Statistics of Item Pools Performance: Condition 3

Statistic	**Extended Operational Pool**	**Simulated Optimal Pool**
Bias	0.0082	-0.0037
RMSE	0.41	0.38
Correlation	0.92	0.93
Test information	14.85	14.98
χ^2 of item exposure rate	5.60	9.56
Items with exposure rate equals 0.2	31%	26%
Items with exposure rate between .02 and 0.20	23%	27%
Items with exposure rate $<$ .02	15%	30%
Items that are not used	31%	17%
Test overlap rate	0.18	0.17
Pool size	144	151

Table 5.7 Summary Statistics of Classification Results for Level 5: Condition 3

	Correct Classification	Distribution of True θ	Distribution of $\hat{\theta}$	FP Errors	Distribution of True θ	Distribution of $\hat{\theta}$	FN Errors	Distribution of True θ	Distribution of $\hat{\theta}$
EOP	67%	1.37*	1.43*	20%	0.41*	0.95*	13%	0.91*	0.44*
		(0.50)	(0.51)		(0.20)	(0.22)		(0.18)	(0.20)
SOP	69%	1.33*	1.42*	17%	0.44*	0.93*	14%	0.91*	0.45*
		(0.46)	(0.52)		(0.20)	(0.20)		(0.19)	(0.20)

Note. *The first number is the mean and in the parenthesis is the standard deviation.

While the ability estimates from the EOP exhibit positive bias, those from the SOP show negative bias. RMSE from the SOP is smaller than that of the EOP. Also, the SOP has a slightly larger correlation coefficient and higher test information than the EOP. The results are consistent with those from Condition 1.

Regarding the classification accuracy for Level 5, the percentage of correct classification is higher for the SOP than for the EOP, 69 versus 67. For this category, the $\hat{\theta}$ distribution is close to the true θ distribution. The FP errors are larger than the FN errors for both pools: 17% and 14% for the SOP versus 20% and 13% for the EOP.

For the item pool usage evaluation indices, χ^2 value of the item exposure rate of the EOP is smaller than that of the SOP. As what was observed under Condition 1 and 2, this was due to the large proportion of the items that are fully used, 31% in the EOP against 26% in the SOP, and small percentage of the items that are seldom used, 15% versus 30%. However, there are fewer items that are never administered in the SOP, 17% in comparison to 31% in the EOP. In addition, though the difference is negligible, the SOP displays a lower test overlap rate.

Figures 5.20 and 5.21 show the item exposure rate for the simulation with 6000 simulees. For the EOP, Figure 5.20 shows that the items in the first stratum reach the MI between the ability levels 0.35 and 1.10. Because there are a small quantity of items in the first stratum in the EOP, they were well used. For the items in the second stratum, the MI is located approximately between the ability levels -1.20 and 1.50. Since 50% of the items are in the second stratum, many of them are not used. The items in the third stratum are informative for the ability levels between -0.70 and 0.65. A few of them were fully used, but the majority of them are administered.

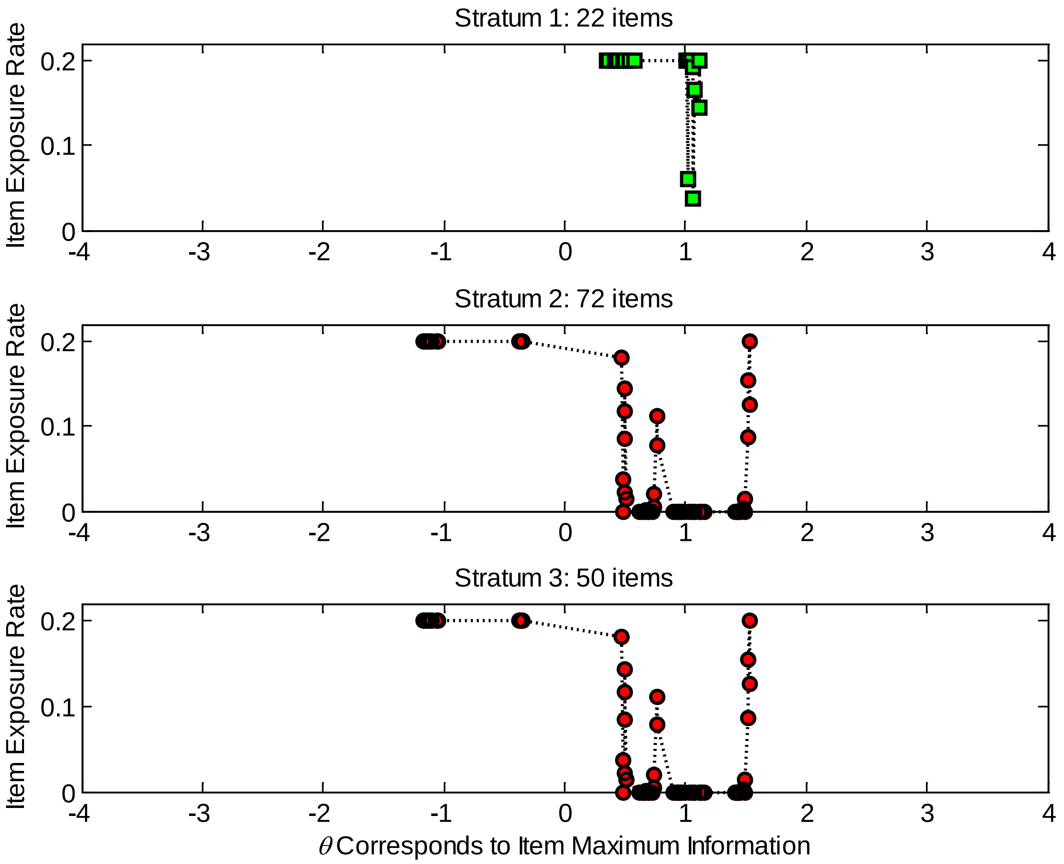


Figure 5.20 Item exposure rate of the EOP under Condition 3

For the SOP, the item exposure rate of the three strata exhibits a similar pattern. The items that are informative between the ability levels -1.5 and 1.5 tend to be well used, and the informative items at the tails are rarely used. Furthermore, the items that are fully used in the first stratum are more dispersed than those in other two strata.

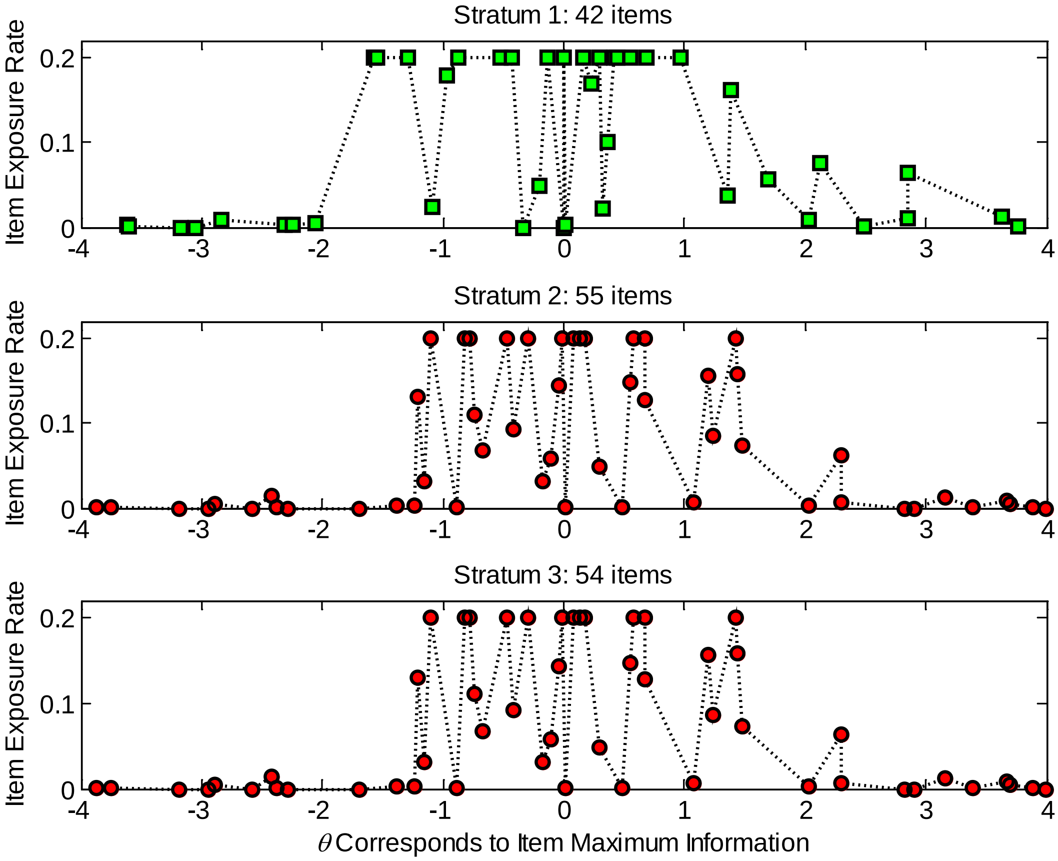


Figure 5.21 Item exposure rate of the SOP under Condition 3

Figures 5.22 to 5.26 show the graphs of the conditional evaluation indices of the θ estimates. As can be seen in Figures 5.22 and 5.23, the SOP provides test information consistently higher than the target level of 10.0 across the ability continuum. For the EOP, the test information is greater than 10.0 between the θ levels -4 and 2.0, and smaller than 10.0 above 2.0. Likewise, except for the ability levels between -2.6 and 0.2, the SOP resulted in smaller conditional SEM. Furthermore, for that interval, the measurement precision of the SOP is well above the desired level.

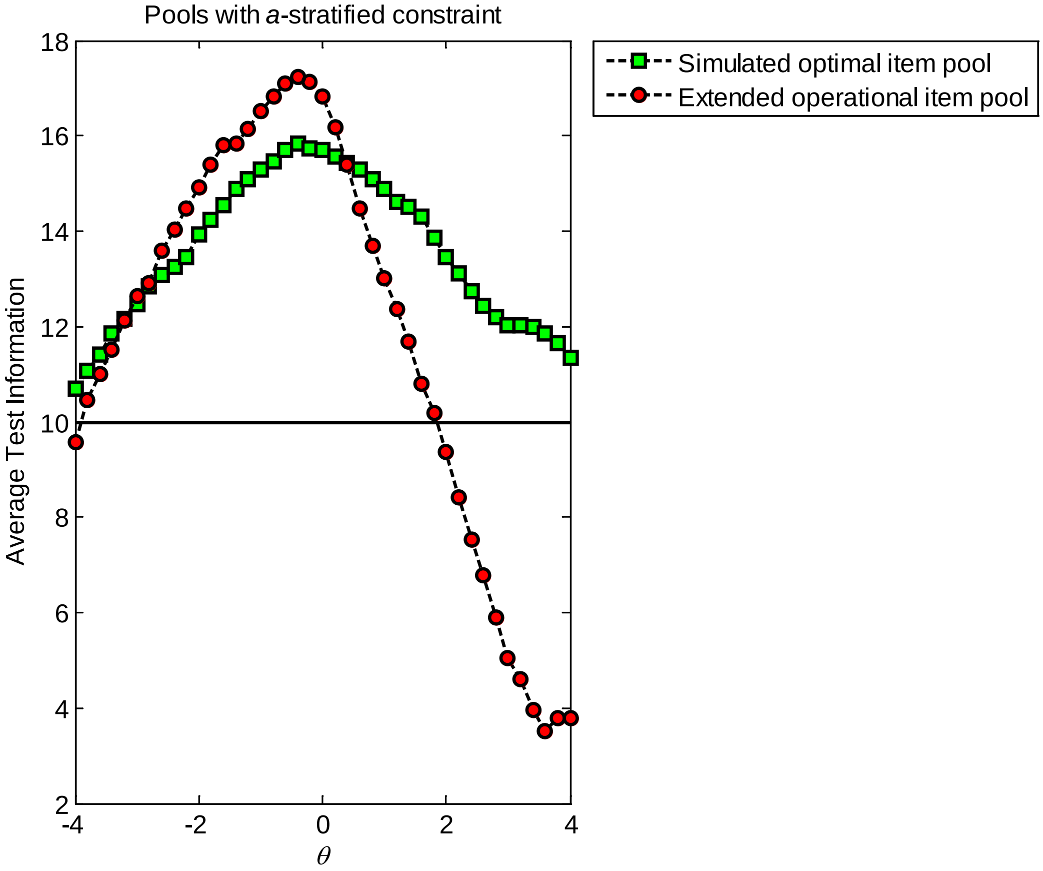


Figure 5.22 Average test information conditional on θ

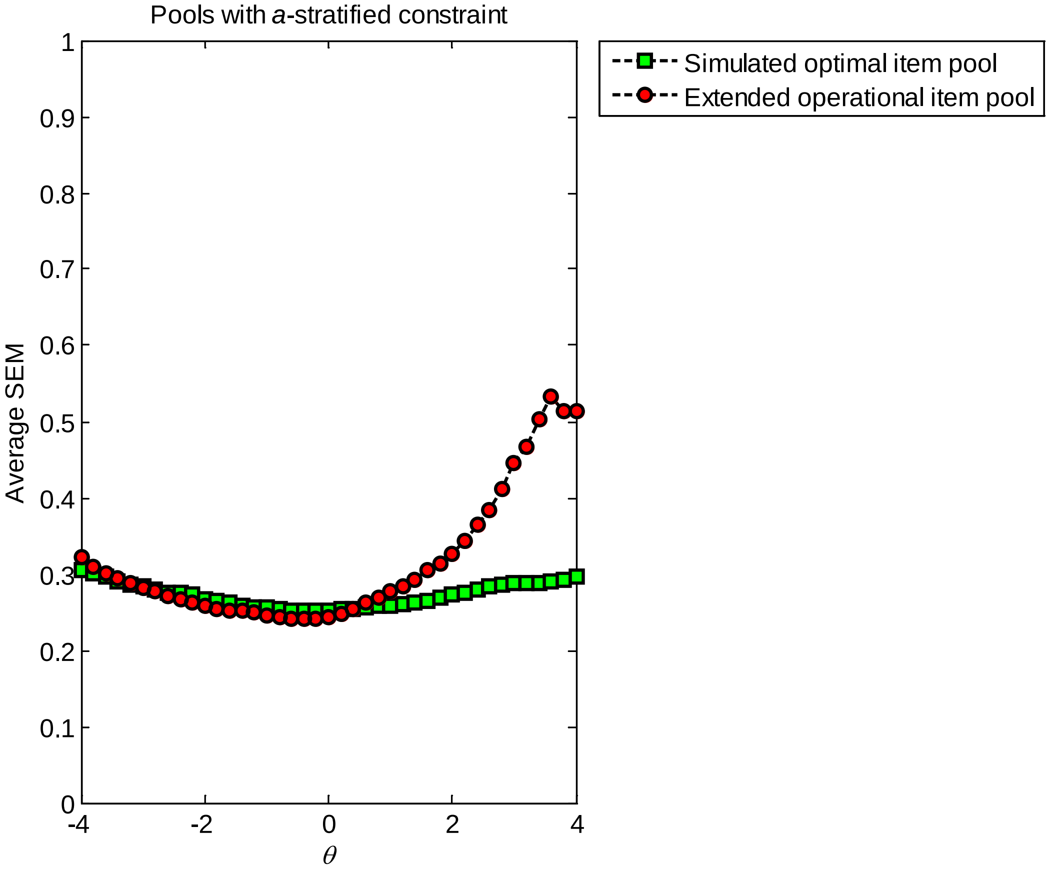


Figure 5.23 Conditional standard error of measurement

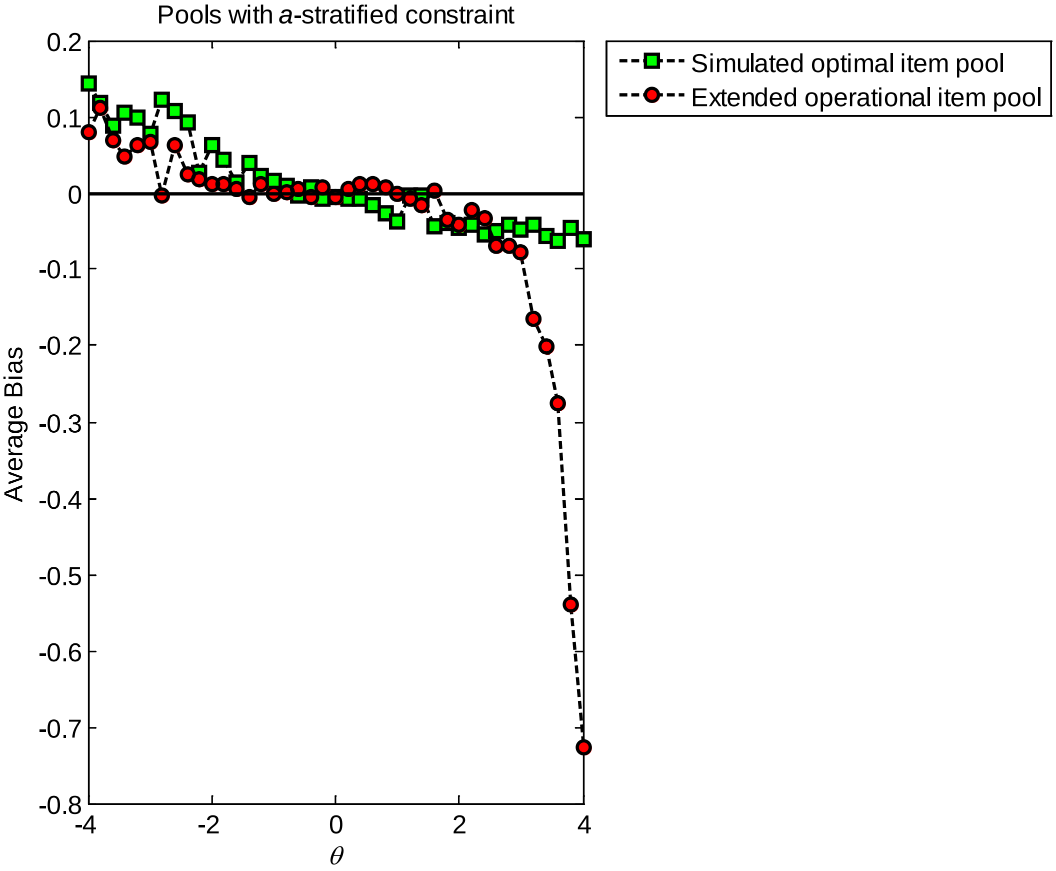


Figure 5.24 Conditional bias

Figure 5.24 illustrates that the EOP and SOP demonstrate a certain level of positive bias at the left end and negative bias at the right. As shown in Figures 5.24 and 5.25, for both pools, the magnitudes of the bias and RMSE are small except at the right end for the EOP.

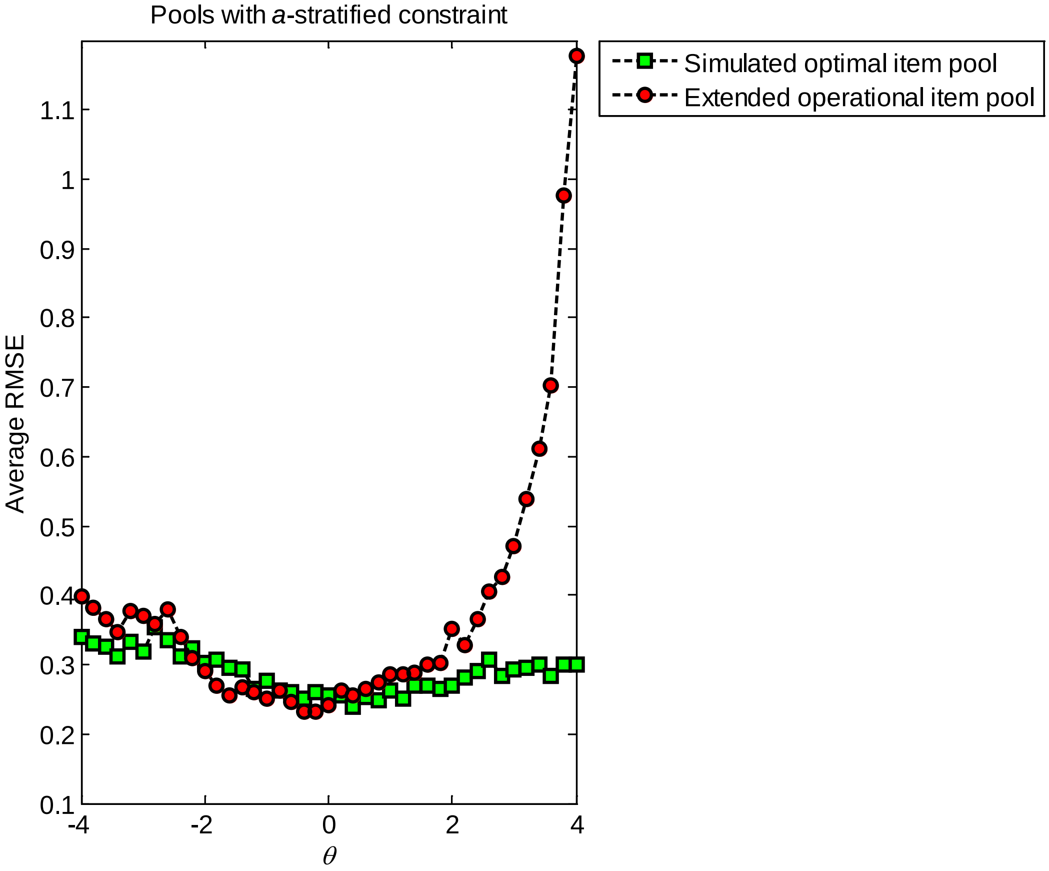


Figure 5.25 Conditional RMSE

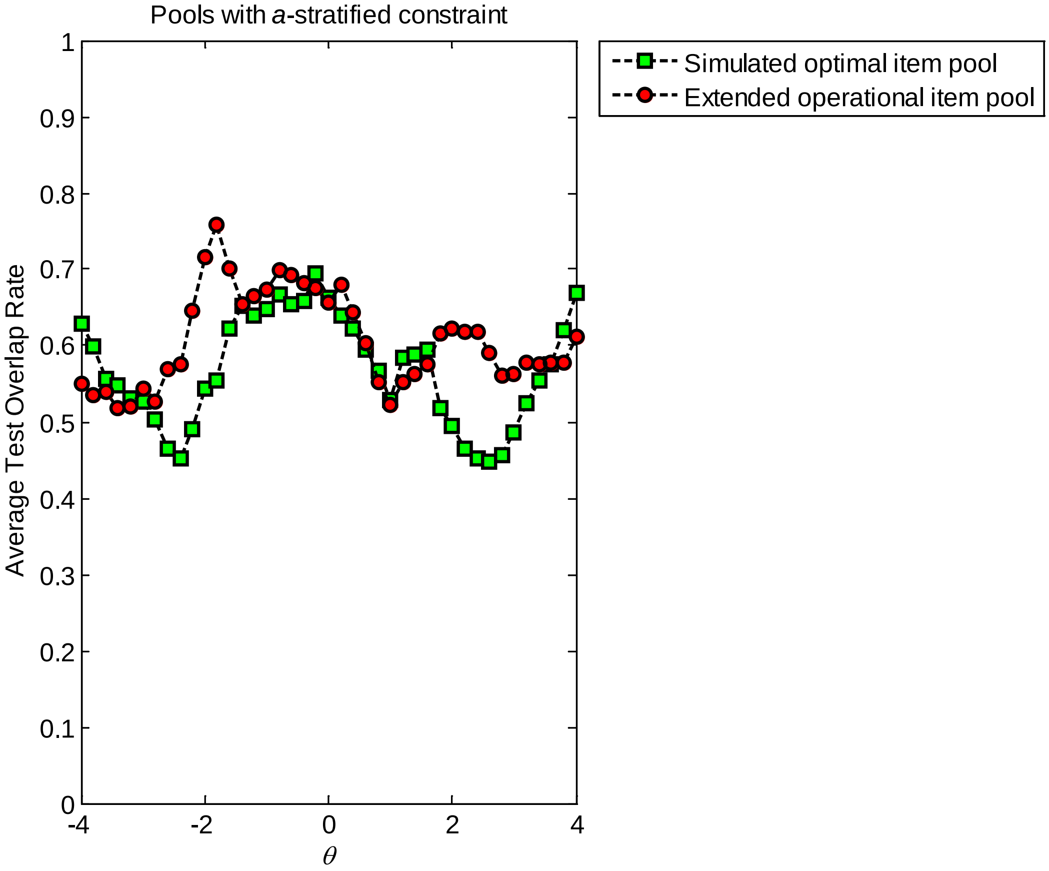


Figure 5.26 Test overlap rate conditional on θ

Figure 5.26 displays the average test overlap rate. The item exposure rate of both pools is approximately between 0.45 and 0.70. The exposure rate is comparable between the ability levels -1.4 and 1.0. The SOP results in lower exposure rates for the ability levels between -3.2 and -1.4 and between 1.0 and 3.6.

To sum up, the results suggest that the SOP designed with the *a*-stratified constraint performs better than the EOP in two ways: 1) it achieves measurement precision across the whole θ range as desired, and 2) it has fewer items that are never administered.

In regards to the item pool characteristics, the number of items included in each stratum is more balanced in the SOP. Also, the items in each stratum are more evenly distributed based on where the MI is located. The distribution of the ***a***-parameter is similar for the two pools. The ***b***-

parameter in the SOP shows a greater variance with the inclusion of some difficult items. Finally, the average maximum information the items in the EOP provides is larger than those in the SOP.

5.5 Item Pool Designed for Unconstrained CAT

5.5.1 Item Pool Characteristics

Figure 5.27 plots the item distributions of the SOP under Condition 4. The detailed item distributions and descriptive statistics of item parameters are presented in Appendix A.13 and A.14.

The pool designed without constraints is similar to the pool designed under Condition 2. First of all, the pool consists of predominantly highly informative items. To be specific, 88% of the items have the ***a***-parameter ranging from 0.95 to 1.10. The mean of the ***a***-parameter is 1.01 compared with 0.89 of the EOP, and the mean of the maximum information the items provides is 1.46 against 1.30 of the EOP. Secondly, the items are informative for the entire ability levels, as what the other three optimal pools demonstrate. Finally, the variance of the ***b***-parameter of the SOP is much larger than that of the EOP.

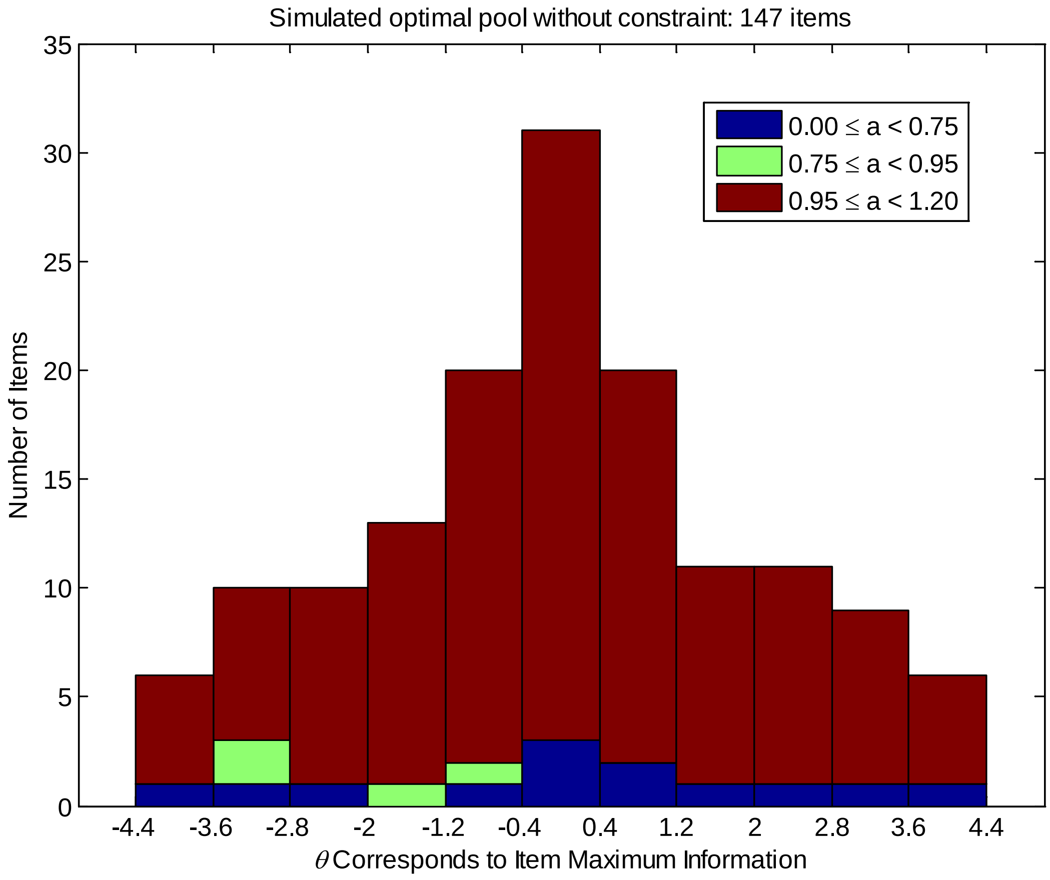


Figure 5.27 Item distributions for the SOP under Condition 4

5.5.2 Item Pool Performance

Table 5.8 and 5.9 presents the evaluation results of the two pools under Condition 4. The ability estimates from both pools exhibit slight negative bias, but the magnitude of the bias is smaller in the SOP. The SOP also results in a slightly smaller RMSE and higher correlation. Also, the average test information from the SOP is higher than that from the EOP.

The percentage of correct classification for Level 5 is higher for the SOP than for the EOP, 77 versus 74. For this category, the $\hat{\theta}$ distribution is consistent to the true θ distribution. The FP errors at Level 5 are 11% for both pools, and the FN errors are 12% and 15% for the SOP and EOP respectively.

Table 5.8 Summary Statistics of Item Pools Performance: Condition 4

Statistic	**Extended Operational Pool**	**Simulated Optimal Pool**
Bias	-0.0239	-0.0152
RMSE	0.27	0.26
Correlation	0.96	0.97
Test information	18.00	18.90
χ^2 of item exposure rate	9.00	8.13
Items with exposure rate equals 0.2	34%	25%
Items with exposure rate between .02 and 0.20	18%	31%
Items with exposure rate < .02	28%	24%
Items that are not used	19%	20%
Test overlap rate	0.18	0.17
Pool size	144	147

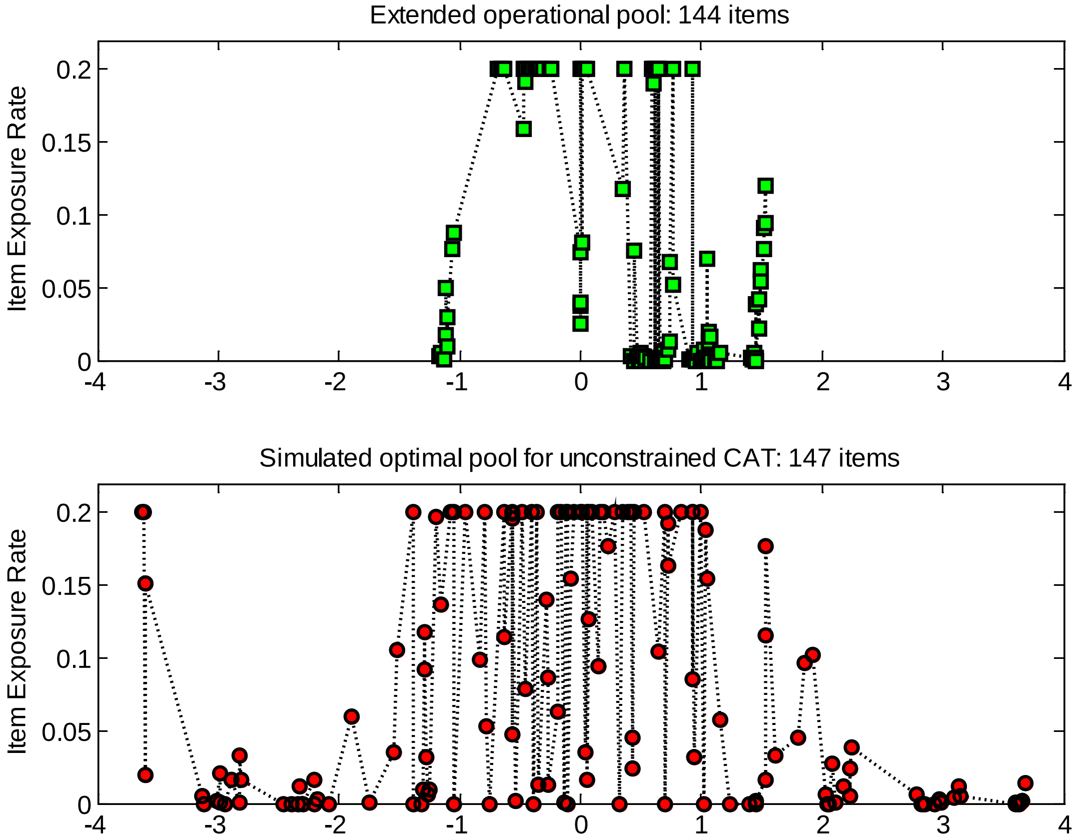
Table 5.9 Summary Statistics of Classification Results for Level 5: Condition 4

	Correct Classification	Distribution of True θ	Distribution of $\hat{\theta}$	FP Errors	Distribution of True θ	Distribution of $\hat{\theta}$	FN Errors	Distribution of True θ	Distribution of $\hat{\theta}$
EOP	74%	1.37*	1.36*	11%	0.53*	0.87*	15%	0.86*	0.51*
		(0.47)	(0.51)		(0.15)	(0.14)		(0.14)	(0.15)
SOP	77%	1.34*	1.35*	11%	0.51*	0.85*	12%	0.84*	0.53*
		(0.48)	(0.48)		(0.15)	(0.14)		(0.14)	(0.14)

Note. * The first number is the mean and in the parenthesis is the standard deviation.

As for the item pool usage, Table 5.8 demonstrates that χ^2 value of the item exposure rate of the EOP is larger than that of the SOP. In regard to item exposure rate, the EOP has a larger amount of items that are fully and rarely used. However, the SOP has a greater proportion of items that are reasonably well used, 56% versus 52%. The test overlap rate is also larger with the EOP. However, the differences in the two indices are trivial. Additionally, the EOP is better used with a larger proportion of items that are administered.

Figure 5.28 graphs the item exposure rate for the simulation with 6000 simulees. For the EOP, the items with their MI located between the ability levels -0.7 and 0.7 are well used. For the SOP, the items which are informative between the ability levels -1.5 and 1 are well used. In addition, a few items in the SOP which are informative at the left tail are fully used.


Extended operational pool: 144 items

Simulated optimal pool for unconstrained CAT: 147 items

Figure 5.28 Item exposure rate of the EOP and SOP under Condition 4

Figures 5.29 to 5.33 plot the conditional evaluation indices. For the unconstrained CAT simulation, Figure 5.29 shows that the SOP provides test information much higher than the target level of 10.0 through the ability continuum. Similarly, the CSEM is consistently below 0.28. The EOP fails to provide test information larger than 10.0 for the ability levels above 2.0. Except for the ability levels between -3.6 and -0.4, the SOP provides more accurate θ estimates.

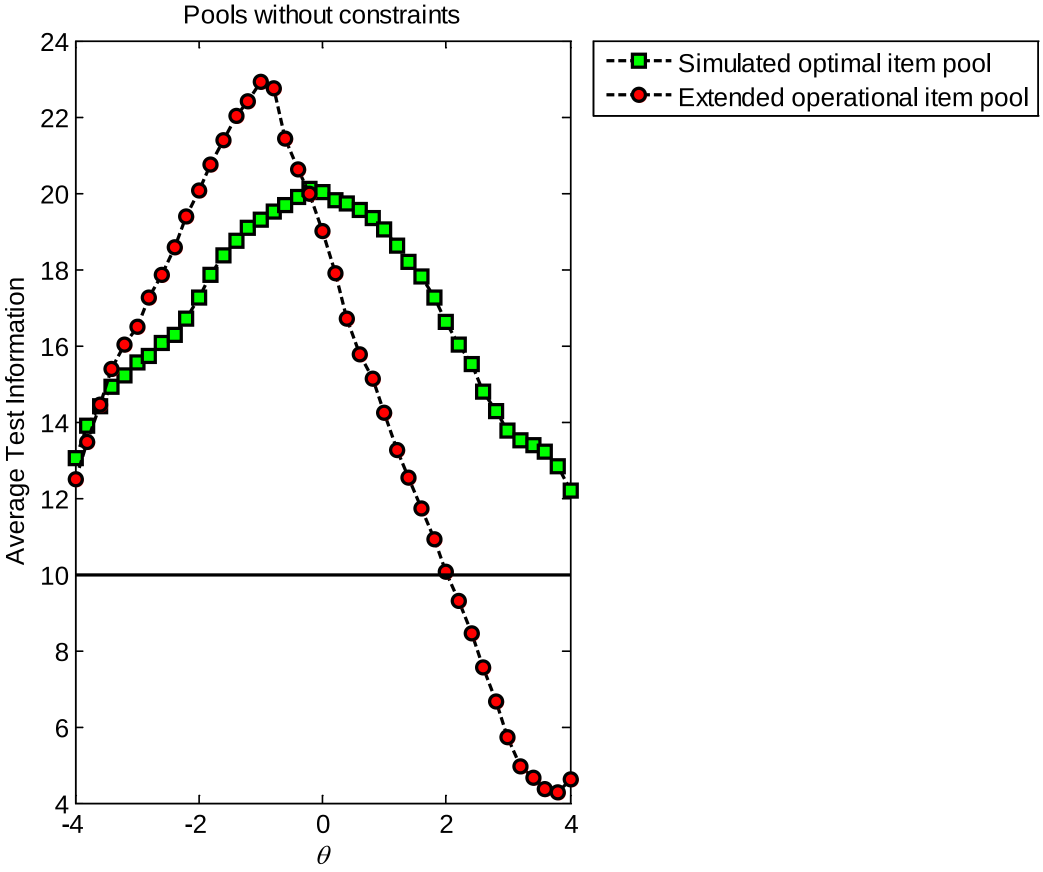


Figure 5.29 Average test information conditional on θ

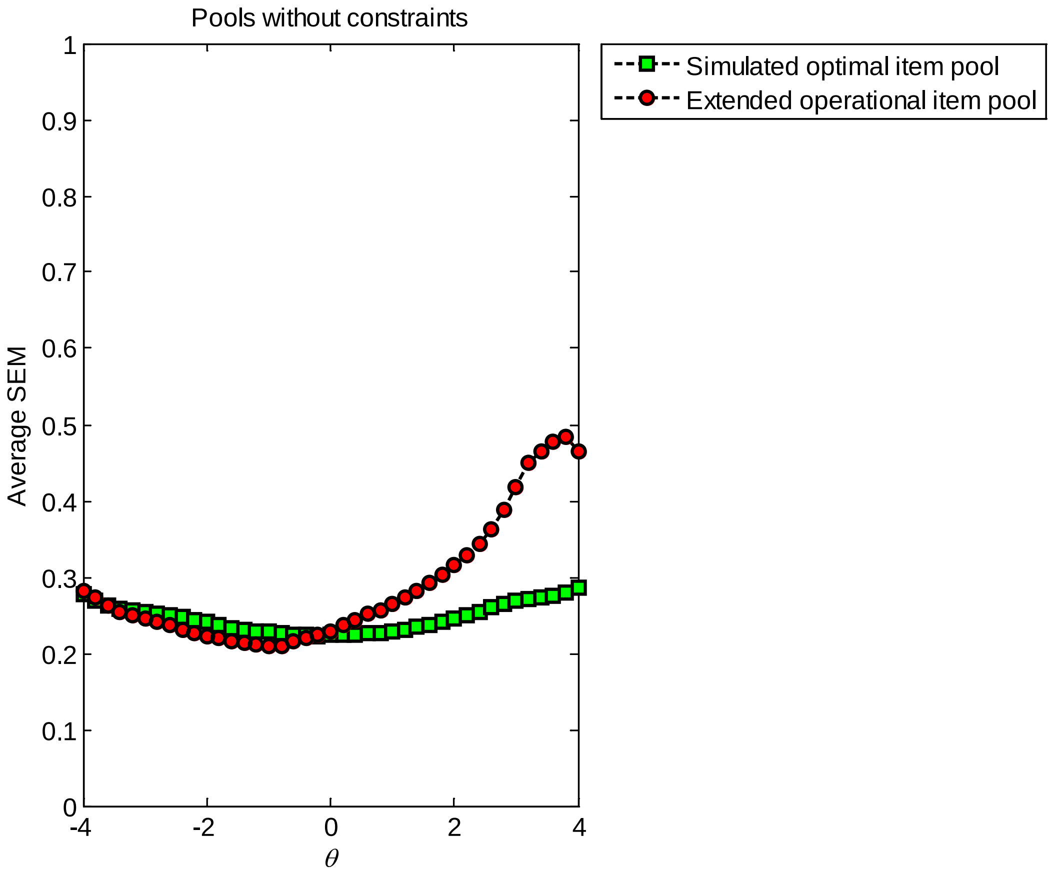


Figure 5.30 Conditional standard error of measurement

Figure 5.31 indicates that both EOP and SOP exhibit a certain level of positive bias roughly between the ability levels -4 and -2.4, and negative bias above the θ point 0.2. Figure 5.32 shows that the SOP results in a consistently smaller RMSE for the θ values greater than 0.4. The magnitudes of the bias and RMSE are small except at the right tail above 3.0 for the EOP.

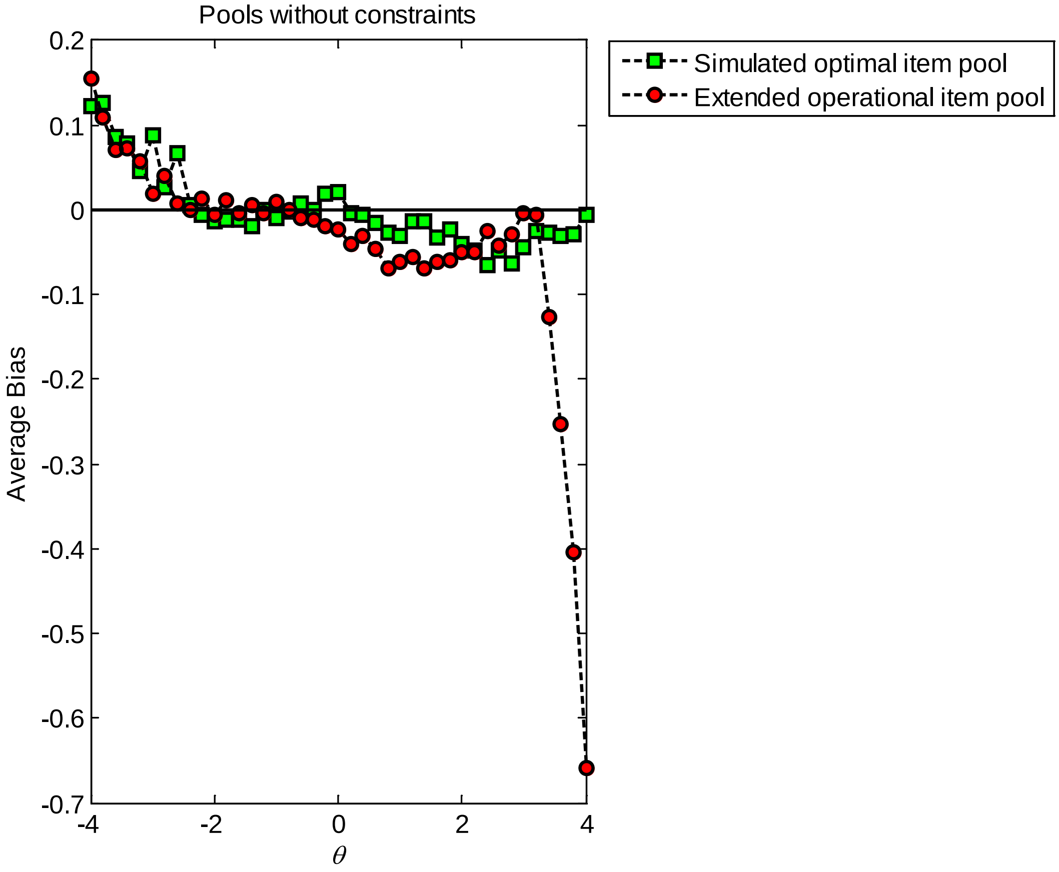


Figure 5.31 Conditional bias

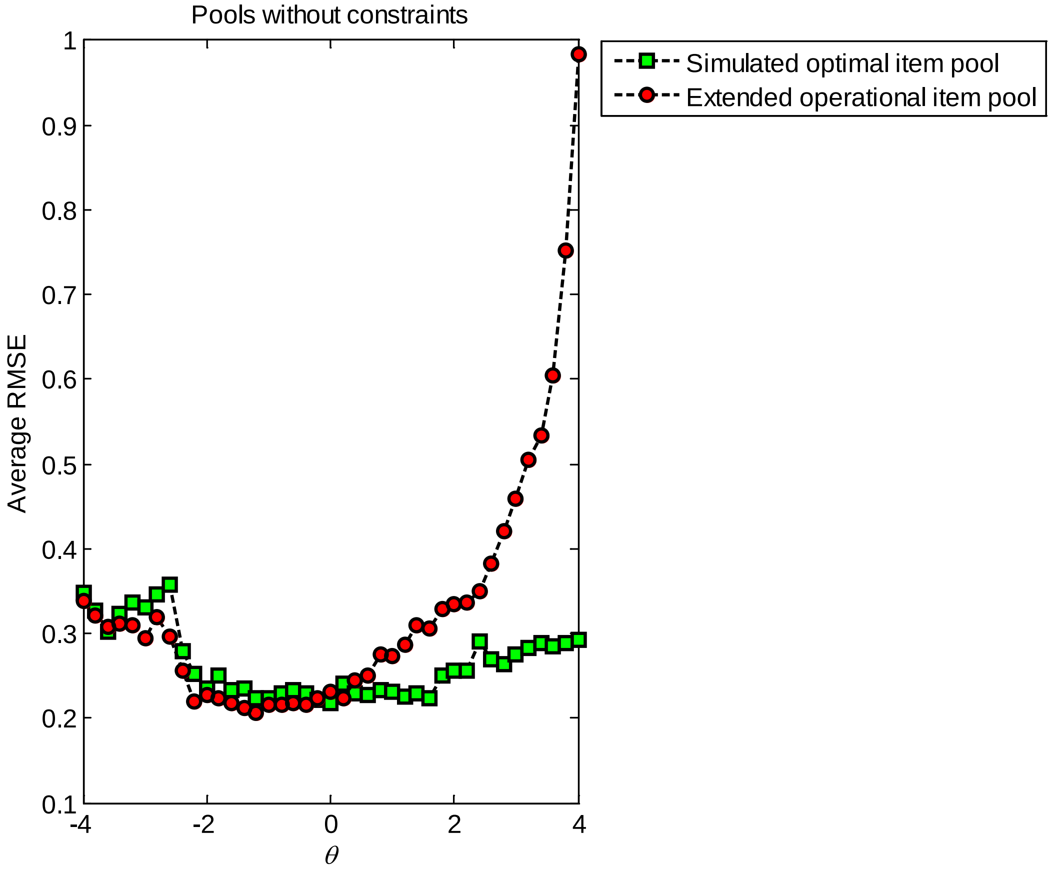


Figure 5.32 Conditional RMSE

Figure 5.32 graphs the conditional test overlap rate. Compared with the pools designed with the *a*-stratified constraint, the SOP designed with unconstrained CAT simulations results in a constantly higher test overlap rate for the SOP. Except for the ability levels between -2.4 and 0.0, the SOP has a larger test overlap rate than the EOP.

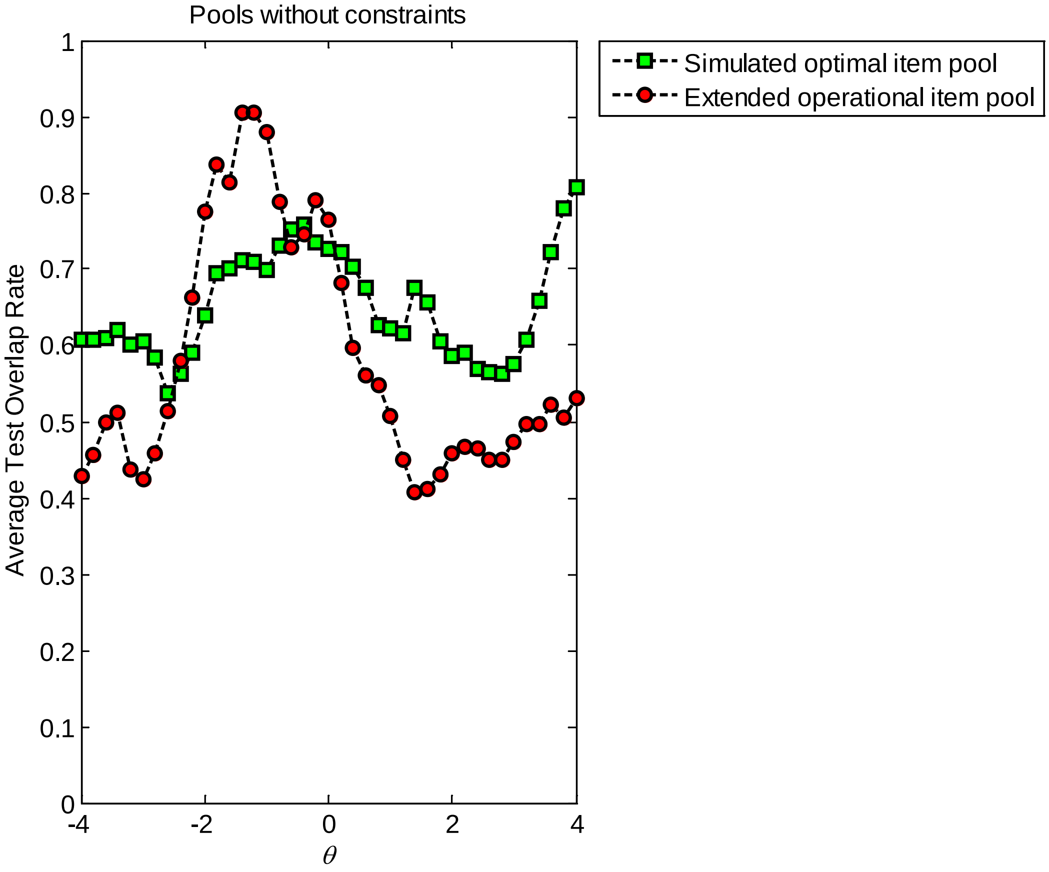


Figure 5.33 Test overlap rate conditional on θ

In summary, the results suggest that the SOP designed under Condition 4 outperforms the EOP in regards to smaller bias and RMSE on average. Also, the SOP achieves the predetermined measurement precision across the whole ability continuum.

Regarding the item pool characteristics, the pool size is almost the same for the EOP and SOP. With the combination of the MI item selection method and restricted maximum item exposure rate, the items included in the SOP are more informative: the mean of the ***a***-parameter and the MI is greater. The ***b***-parameter ranges from -1.83 to -0.50 in the EOP, and from -1.88 to 1.96 in the SOP.

CHAPTER 6
RESULTS AND DISCUSSIONS

This chapter discusses the results and their implications. It first summarizes the findings addressing the research questions; the implications for item pool development and management follow. Finally, the limitations and suggestions for future research are presented.

6.1 Item Pool Blueprint

This study aimed to examine the impact of practical constraints on designing optimal item pools and produce a blueprint for each combination of constraints. This section first reiterates how the *p*-optimality method is applied to the polytomous CAT and the process to obtain the optimal item pools. The blueprints under varying contexts are then presented.

While dichotomous items in an item pool are generally described using item parameters, the extension of the *p*-optimality method resulted in describing polytomous items based on where the items' maximum information is located. When an item's response categories are equal to or more than three, the analytic solution between the maximum information and item parameters is unavailable. Therefore, selecting items primarily based on ***b***-parameter, as in the case of dichotomous items, might be inaccurate because the location parameter under the GPCM might deviate greatly from an item's maximum information point. This approach of describing items is consistently used in this study throughout.

Furthermore, without the analytic solution, generating a most informative item based on updated $\hat{\theta}$s during a CAT simulation is unattainable. Instead, a bootstrapping method was used with a simulated master pool that mirrors an operational achievement test. CAT simulations with 6000 simulees were conducted with five replications based on the predetermined CAT characteristics and the constraints as summarized in Table 4.3. The resulting item pools were averaged and then trimmed to meet the *p*-optimality criterion. The remaining items define the final optimal pools and the blueprints. Findings concerning the research questions are summarized below.

How practical constraints such as item exposure control and content balancing affect the optimal item pool design and its performance?

The SOPs resulting from various constraints contain 144, 147, 150, and 151 items respectively. Namely, the practical constraints of the *a*-stratified exposure control and content balancing do not affect pool size to a large extent. However, the maximum item exposure rate of 0.20 was applied for all the conditions, which controlled the pool size to a certain degree.

Regarding the optimal item pool design, the *a*-stratified exposure control affects the pool characteristics greatly. First of all, the items included in the SOPs without the constraint have larger ***a***-parameter than those in the SOPs with the constraint, as can be seen in Figures 5.3, 5.11, 5.19, and 5.27. Correspondingly, the average MI that the items provide is greater. Secondly, the pool information of the SOPs without the constraint is much larger than the SOPs with the constraint across the entire θ range, as depicted in Figure 5.1.

In addition, the content balancing has little impact on the pool design. For instance, when the conditions differ only by the content balancing, such as Conditions 1 versus 3, and Conditions 2 versus 4, the distribution of the ***a***-parameter is quite similar with respect of the magnitudes and deviations. Also, when the *a*-stratified exposure control applies, the pool information curve without the content balancing is slightly larger than the one with it. For Conditions 2 and 4, the difference is negligible between the pool information curves.

The evaluation results of the pool performance are closely related to the pool characteristics. Under Conditions 1 and 3, the average test information is approximately 15.0, but it reaches nearly 19.0 under Conditions 2 and 4. The difference in the average test information does not imply meaningful difference in the average SEM for the conditions with and without the *a*-stratified procedure. While the average SEM under both conditions is smaller than the desired level of 0.30, a decrease of as small as 0.03 using a larger number of highly discriminative items might not be practical. This is consistent with what previous research has found: the *a*-stratified control resulted in small decreases in measurement precision in polytomous CAT (Davis, 2004; Pastor et al., 2002). Consistently, pools designed with the *a*-stratified exposure control yielded larger RMSE and smaller correlation coefficients. As for the classification results, the percentage of correct classification for Level 5 is higher under Conditions 2 and 4 than that under Conditions 1 and 3.

With regard to the item pool usage, the percentage of items that are fully used, well used, rarely used, and never used are quite comparable for the three pools designed under Conditions 1 to 3. In addition, the number of items that are not administered does not differ much under all conditions: 15%, 17%, 17%, and 20%. However, compared with the EOP, when the *a*-stratified method applied, the conditional test overlap rate of the SOP is consistently lower. Without the constraint, the conditional test overlap rate is higher for the whole θ range except for the θ values between -2 and -0.6.

For each combination of the constraints, what does the blueprint display with regard to the characteristics and distribution of the items, item pool information distribution, and pool size for a modeled CAT procedure?

When the *a*-stratified method is applied, the item distribution in three strata is similar for the resulting pools. In general, while the number of items in the second and third stratum is nearly identical, there are fewer items in the first stratum. Particularly, 28% of the items fall into the first stratum, and 36% in the higher strata. It should be noted that the distribution is quite different from the EOP, in which there are 15%, 50%, and 35% in each stratum respectively.

With the content balancing control, the number of items for each content area differs as well. For both pools with the content balancing control, there are 47% and 53% in each content area correspondingly.

Without any constraints, the SOP consists of predominantly highly discriminative items. While the SOP yielded more precise measurement, the item pool usage is not as good as the one with the constraints.

Because the normal ability distribution is assumed and the maximum item exposure rate is restricted for all simulations, the SOPs demonstrated that more items are needed in the middle of the θ scale. As the optimal pool design included the informative items at both ends, the number of items is somewhat symmetrical for the ability levels below -2.0 and above 2.0. Nevertheless, these items are seldom used in comparison to the items that are informative in the middle of the θ scale.

Concerning item characteristics, following properties are observed from the SOPs:

1) The distributions of the ***a***- and ***b***-parameter are similar for the SOPs with and without the *a*-stratified constraint. For the SOPs with the constraint, the mean of the ***a***-parameter is approximately 0.90 with the lowest discrimination of 0.65 and the highest of 1.10. For the SOPs without the constraint, the average ***a***-parameter is around 1.0. The variation of the ***a***-parameter in all the SOPs is small. The mean of the ***b***-parameter is about -0.30 with the lowest value of -1.90 and the highest of 1.90. Again, the distribution of the ***b***-parameter in all the SOPs is similar. The distribution of the ***b***-parameter differs from the EOP with a smaller mean and larger deviation. This is due to the absence of the items that are informative at the right end in the EOP.

2) The threshold parameters, d_j, are orderly distributed from the easiest to the hardest. This is the property demonstrated in the operational items and embedded in item generation. The range of the five threshold parameters is approximately between 2.8 and -2.8. Because the sum of the threshold parameters is constrained to zero and there are five of them in total, the threshold parameters are fairly symmetrical around zero.

For the information distribution at the item and item pool level, the SOPs consist of items whose maximum information is located across the θ scale, especially at both tails. For the item pools, the pool information curves of the SOPs are smooth and somewhat bell shaped compared with the EOP. Finally, the pool information curves without the item exposure control are larger than those with it.

6.2 Discussions and Implications

6.2.1 Discussions

The pools produced from an optimal item pool design is a union of items that meet all of the predetermined statistical and psychometric specifications and are informative for a given population of examinees (Veldkamp & van der Linden, 2000). Therefore, they are not absolutely optimal, rather optimal under many specific constraints. In general, the optimal pool is one characterized by several criteria: 1) it has a sufficient number of items for overlapping test assemblies; 2) it consists of items with evenly distributed difficulty; and 3) it includes items with both high and low discrimination (Flaugher, 2000; Veldkamp & van der Linden, 2000).

Based on the criteria above, while the size of the pools that are compared in this study is quite similar, the SOPs show advantages in two ways: including items with the maximum information spread across the entire ability continuum and the ***a***-parameter spans evenly between

0.60 and 1.10 when the *a*-stratified method applies. The first feature suggests that, corresponding to the normal ability distribution of a population of examinees, the SOPs are informative at a wide range of ability levels. The second feature, the inclusion of the less discriminating items, is practical for item writing effort and for reducing the item writing cost when the item exposure control applies.

The evaluation results are mixed. The SOPs performs well consistently in that they achieved the target level of measurement precision across the entire ability continuum, whereas the EOP failed to meet the criterion for the ability levels above 2.0. Because the operational test modeled is a criterion-referenced test, even though the measurement precision is unequal at different θ levels, a conditional reliability consistently larger than 0.90 suggests the SOPs support the CAT administration with the desired precision.

On the other hand, the EOP contains many highly informative items, especially at the ability levels between -2.2 and 0.2. Consequently, the percentage of the items that are fully used is larger than the SOPs. Except for the unconstrained situation, the percentage of the items that are never used is much smaller in the SOPs. The presence of fewer unused items is preferable for reducing the cost to produce an item pool.

Finally, the evaluation criteria such as bias, RMSE, and correlation show that the performance of the EOP is comparable to that of the optimal pool. This suggests that the operational pool provides accurate measurement over a specific range of ability levels. What is concerned is that it demonstrates problems in measuring high-achieving students accurately. Also, because the items shrink into a certain range of ability levels, many items that are less informative are not administered.

From the discussion above, it appears that the significance of the study is two-fold. First, extending the *p*-optimality method to design an optimal item pool for polytomous CAT is feasible. The resulting optimal pool performs well to fulfill the purposes of evaluating students' ability at a predetermined precision level and maintaining item pool usage and security at a desired level, especially when the item exposure control applies. Describing items based on where the maximum information is located captures items' characteristics well, and combining items in the same $a\theta$-bin does not reduce the measurement precision at the individual/test level. Second, the pool blueprints can be used to help transforming a P&P test pool into a pool for adaptive testing purposes as needed. Specifically, items that are informative for θ levels above 2.0 are needed to supplement the existing item pool for the P&P test. Also, the current item pool for the P&P test contains a large number of items that are discriminating at the center of the θ scale and many of them, even the ones with high ***a***-parameter, are not administered. While the blueprints of the SOP indicate that more items are required for the middle ability levels, an effective approach might be to divide these items into subpools based on what the blueprints delineate. Additionally, as the results show, the inclusion of the items with low ***a***-parameter in the SOPs has little impact on measurement precision. Item pool development in the future can include more items with low ***a***-parameter to ease item writing efforts.

A final remark about implementing the blueprint: constructing an item pool blueprint is a continuous process (Veldkamp & van der Linden, 2000). Namely, after items' statistical and psychometric attributes become available through field testing, a blueprint can be updated with the same CAT simulation process. This dynamic process can help adapt the item writing effort based on what is already included in the pool to achieve the goal of constructing optimal pools. In addition, if a CAT program varies from the current context of simulation with respect to factors such as ability distribution of the population or practical constraints, a blueprint can be created by modifying the existing program.

6.2.2 Implications

Because the blueprint obtained from an optimal item pool design represents the maximal combination satisfying the preset specifications, it is an instructive to guide item writing process, develop and manage an item pool.

First of all, because the statistical attributes of the items included in the optimal pool are derived from the operational test, the best way to start item writing effort is to build up from the current operational pool. For instance, if an existing pool needs to be supplemented informative items at the high θ levels, the new items to be written can follow the categorical and statistical attributes of the similar operational items. While it is impossible in practice to write items with the preferred quantitative attributes exactly, identifying an optimal goal may still benefit the initial item-writing efforts. Furthermore, a blueprint serves as a target for item writing, as items are written and field tested, repeated application of the method is necessary to update the blueprint to direct item writing in a sequential manner.

Secondly, when there is more than one content area to be tested, item pool development needs to take into account the order that they are administered. As the blueprint shows, the first content area requires fewer items in total, but more items that are informative in the middle of the θ scale. For the second content area, the items included are more evenly distributed across the ability levels. Although it is common in practice that not all domains are balanced, such as algebra and geometry in mathematics, the first content area should be a one that can potentially produce many items of medium difficulty.

When multiple content areas are included in a test, dimensionality should always be examined. If unidimensionality is theoretically and psychometrically solid, one final θ score can be derived as did in this study. When the test displays multidimensionality, domain and overall scores can be obtained through different methods using either unidimensional or multidimensional IRT models (Reckase, 2009; Yao, 2010).

Finally, managing an item pool is a continuous process. For many reasons, items need to be removed from or supplemented in an item pool. As discussed above, exact replication of items, especially the quantitative attributes, is difficult to realize. Therefore, recalculating a blueprint when pool composition changes is necessary to manage an item pool in order to fulfill the CAT purposes.

6.3 Limitations and Future Studies

The results of this study demonstrate the advantages of using an optimal item pool in ability estimation and pool security as shown in the CAT simulations. This conclusion, however, is restricted by the fact that an operational pool of the similar size as the generated optimal item pool is not available. The extended operational item pool reduplicated from a limited number of operational items may not represent an operational CAT item pool. Furthermore, the item pools in previous studies consist of predominantly dichotomous items and a small number of polytomous items, it is therefore difficult to evaluate how the EOP is typical in practice.

Another factor that will affect the results in this study is the width of the θ-bin. Because the items from a same θ-bin are supposed to be equally informative and the final pool is the union of the items from all θ-bins, the width determines the amount of items that will be trimmed out. It is thus worthwhile to investigate to what extent pool size can be decreased without losing measurement precision by manipulating the width of θ-bin based on the operational item characteristics. In addition, as the results indicate that the informative items at both tails are rarely used, a combination of narrow widths in the center and wide ones at the end might be a reasonable approach to approximate optimal pools of smaller sizes.

Finally, because polytomous items are mostly performance-oriented, they usually require more time to complete. While it is unreasonable to assume a test consisting of polytomous items only, there is an increasing use of the mixed-format test design. It is of further interest to examine how the current findings could be incorporated with optimal item pool design research using dichotomous IRT models to design item pools for mixed-format CAT implementation.

APPENDIX

Table A.1 Item Distribution for Extend Operational Item Pool

θ \ a		0.00 0.75	0.75 0.95	0.95 1.20	Total
-5.0	-3.6	0	0	0	0
-3.6	-2.8	0	0	0	0
-2.8	-2.0	0	0	0	0
-2.0	-1.2	0	0	0	0
-1.2	-0.4	0	9	18	27
-0.4	0.4	2	4	22	28
0.4	1.2	20	42	10	72
1.2	2.0	0	17	0	17
2.0	2.8	0	0	0	0
2.8	3.6	0	0	0	0
3.6	5.0	0	0	0	0
Total		22	72	50	144

Table A.2 Descriptive Statistics of Item Parameters for Extended Operational Pool

	Parameter	N	Mean	SD	Minimum	Maximum
Extended Operational Pool		144				
	a		0.89	0.14	0.64	1.14
	b		-1.08	0.40	-1.83	-0.50
	d_1		2.87	0.56	2.10	4.27
	d_2		1.26	0.24	0.85	1.54
	d_3		-0.11	0.09	-0.31	0.02
	d_4		-1.47	0.27	-2.05	-1.03
	d_5		-2.55	0.48	-3.36	-1.73
1st Stratum		22				
	a		0.65	0.01	0.64	0.65
	b		-0.59	0.07	-0.69	-0.50
	d_1		3.21	0.01	3.19	3.22
	d_2		1.38	0.16	1.22	1.54
	d_3		-0.08	0.05	-0.13	-0.02
	d_4		-1.61	0.08	-1.71	-1.52
	d_5		-2.90	0.12	-3.03	-2.76
2nd Stratum		72				
	a		0.87	0.06	0.77	0.95
	b		-0.97	0.32	-1.50	-0.56
	d_1		3.11	0.56	2.23	4.27
	d_2		1.33	0.19	0.95	1.54
	d_3		-0.12	0.10	-0.31	0.02
	d_4		-1.59	0.26	-2.05	-1.15
	d_5		-2.74	0.41	-3.36	-1.84
3rd Stratum		50				
	a		1.03	0.07	0.95	1.14
	b		-1.45	0.23	-1.83	-1.18
	d_1		2.37	0.26	2.10	2.73
	d_2		1.09	0.23	0.85	1.50
	d_3		-0.11	0.09	-0.23	0.01
	d_4		-1.23	0.17	-1.51	-1.03
	d_5		-2.11	0.33	-2.68	-1.73

Table A.3 Item Distribution for Simulated Optimal Pool under Condition 1[*]: Content 1

	a	0.00	0.75	0.95	
θ		0.75	0.95	1.20	Total
-5.0	-3.6	0	2	2	4
-3.6	-2.8	1	2	2	5
-2.8	-2.0	1	2	2	5
-2.0	-1.2	1	2	2	5
-1.2	-0.4	2	2	3	7
-0.4	0.4	6	4	4	14
0.4	1.2	2	3	3	8
1.2	2.0	1	2	2	5
2.0	2.8	1	2	2	5
2.8	3.6	0	2	2	4
3.6	5.0	1	2	2	5
Total		16	25	26	67

Note. [*]Condition 1: Content balancing and *a*-stratified constraints.

Table A.4 Item Distribution for Simulated Optimal Pool under Condition 1[*]: Content 2

	a	0.00	0.75	0.95	
θ		0.75	0.95	1.20	Total
-5.0	-3.6	2	2	2	6
-3.6	-2.8	2	2	2	6
-2.8	-2.0	2	2	2	6
-2.0	-1.2	2	2	2	6
-1.2	-0.4	3	3	3	9
-0.4	0.4	3	4	4	11
0.4	1.2	3	3	3	9
1.2	2.0	2	2	2	6
2.0	2.8	2	2	2	6
2.8	3.6	2	2	2	6
3.6	5.0	2	2	2	6
Total		25	26	26	77

Note. [*]Condition 1: Content balancing and *a*-stratified constraints.

Table A.5 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 1*: Content Area 1

Parameter	N	Mean	SD	Minimum	Maximum
SOP: Content 1	67				
a		0.92	0.14	0.67	1.10
b		-0.24	1.02	-1.91	1.96
d_1		2.78	0.19	2.40	3.23
d_2		1.34	0.21	0.99	1.77
d_3		-0.05	0.11	-0.25	0.35
d_4		-1.35	0.19	-1.87	-0.63
d_5		-2.72	0.30	-3.76	-2.18
1st Stratum	16				
a		0.72	0.02	0.67	0.75
b		-0.23	0.85	-1.23	1.96
d_1		2.80	0.20	2.46	3.23
d_2		1.41	0.24	1.07	1.76
d_3		-0.07	0.10	-0.25	0.11
d_4		-1.34	0.29	-1.87	-0.63
d_5		-2.81	0.29	-3.63	-2.44
2nd Stratum	25				
a		0.90	0.04	0.80	0.95
b		-0.20	1.05	-1.91	1.91
d_1		2.90	0.17	2.57	3.21
d_2		1.39	0.19	1.01	1.77
d_3		-0.02	0.13	-0.23	0.35
d_4		-1.41	0.15	-1.63	-1.17
d_5		-2.86	0.31	-3.76	-2.35
3rd Stratum	26				
a		1.07	0.02	1.02	1.10
b		-0.30	1.11	-1.75	1.96
d_1		2.66	0.12	2.40	2.87
d_2		1.25	0.19	0.99	1.57
d_3		-0.08	0.09	-0.24	0.11
d_4		-1.30	0.14	-1.49	-1.06
d_5		-2.54	0.19	-2.88	-2.18

Note. *Condition 1: Content balancing and *a*-stratified constraints.

Table A.6 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 1*: Content Area 2

	Parameter	N	Mean	SD	Minimum	Maximum
Item Pool (Content 2)		77				
	a		0.89	0.14	0.68	1.10
	b		-0.26	1.08	-1.90	1.97
	d_1		2.81	0.21	2.15	3.26
	d_2		1.39	0.20	0.89	1.88
	d_3		-0.05	0.13	-0.24	0.38
	d_4		-1.39	0.22	-1.74	-0.57
	d_5		-2.76	0.30	-3.83	-2.14
1st Stratum		25				
	a		0.72	0.02	0.68	0.75
	b		-0.13	1.04	-1.62	1.97
	d_1		2.82	0.26	2.15	3.24
	d_2		1.44	0.20	1.13	1.88
	d_3		-0.01	0.16	-0.24	0.38
	d_4		-1.44	0.31	-1.74	-0.57
	d_5		-2.83	0.35	-3.83	-2.15
2nd Stratum		**26**				
	a		0.90	0.03	0.83	0.94
	b		-0.35	1.12	-1.90	1.88
	d_1		2.90	0.19	2.55	3.26
	d_2		1.44	0.16	1.19	1.74
	d_3		-0.07	0.12	-0.22	0.17
	d_4		-1.43	0.16	-1.67	-1.16
	d_5		-2.84	0.25	-3.39	-2.54
3rd Stratum		**26**				
	a		1.04	0.04	0.95	1.10
	b		-0.29	1.11	-1.88	1.93
	d_1		2.72	0.15	2.53	3.11
	d_2		1.28	0.18	0.89	1.52
	d_3		-0.07	0.11	-0.23	0.19
	d_4		-1.33	0.13	-1.56	-1.12
	d_5		-2.60	0.24	-3.04	-2.14

Note. * Condition 1: Content balancing and *a*-stratified constraints.

Table A.7 Item Distribution for Simulated Optimal Pool under Condition 2[*]: Content 1

a		0.00	0.75	0.95	
θ		0.75	0.95	1.20	Total
-5.0	-3.6	0	0	4	4
-3.6	-2.8	0	1	4	5
-2.8	-2.0	0	0	5	5
-2.0	-1.2	0	0	6	6
-1.2	-0.4	0	0	8	8
-0.4	0.4	0	0	17	17
0.4	1.2	0	0	10	10
1.2	2.0	0	0	5	5
2.0	2.8	0	0	5	5
2.8	3.6	0	0	3	3
3.6	5.0	0	0	3	3
Total		0	1	70	71

Note. [*]Condition 2: Content balancing control.

Table A.8 Item Distribution for Simulated Optimal Pool under Condition 2[*]: Content 2

a		0.00	0.75	0.95	
θ		0.75	0.95	1.20	Total
-5.0	-3.6	0	1	2	3
-3.6	-2.8	1	1	4	6
-2.8	-2.0	0	1	5	6
-2.0	-1.2	0	0	5	5
-1.2	-0.4	2	0	9	11
-0.4	0.4	3	0	10	13
0.4	1.2	3	0	9	12
1.2	2.0	1	0	5	6
2.0	2.8	1	0	5	6
2.8	3.6	1	0	5	6
3.6	5.0	1	1	3	5
Total		13	4	62	79

Note. [*]Condition 2: Content balancing control.

Table A.9 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 2*: Content Area 1

Parameter	N	Mean	SD	Minimum	Maximum
SOP: Content 1	71				
a		1.05	0.03	0.94	1.10
b		-0.47	0.93	-1.80	1.96
d_1		2.69	0.16	2.40	3.36
d_2		1.25	0.18	0.95	1.72
d_3		-0.08	0.10	-0.27	0.35
d_4		-1.29	0.14	-1.58	-0.98
d_5		-2.57	0.26	-3.77	-2.18

Note. *Condition 2: Content balancing control.

Table A.10 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 2*: Content Area 2

Parameter	N	Mean	SD	Minimum	Maximum
SOP: Content 2	79				
a		0.98	0.15	0.56	1.10
b		-0.38	1.02	-1.88	2.00
d_1		2.78	0.21	2.45	3.44
d_2		1.31	0.25	0.89	1.88
d_3		-0.07	0.11	-0.30	0.31
d_4		-1.36	0.18	-1.88	-1.02
d_5		-2.65	0.32	-3.68	-2.14

Note. *Condition 2: Content balancing control.

Table A.11 Item Distribution for Simulated Optimal Pool under Condition 3*

θ \ a		0.00 0.75	0.75 0.95	0.95 1.20	Total
-5.0	-3.6	2	2	4	8
-3.6	-2.8	3	4	4	11
-2.8	-2.0	3	4	3	10
-2.0	-1.2	3	4	4	11
-1.2	-0.4	5	9	8	22
-0.4	0.4	11	10	8	29
0.4	1.2	5	7	7	19
1.2	2.0	3	4	4	11
2.0	2.8	3	3	4	10
2.8	3.6	2	4	4	10
3.6	5.0	2	4	4	10
Total		42	55	54	151

Note. * Condition 3: *a*-Stratified constraint.

Table A.12 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 3*

	Parameter	N	Mean	SD	Minimum	Maximum
SOP		151				
	a		0.90	0.14	0.59	1.10
	b		-0.30	1.02	-1.90	1.98
	d_1		2.79	0.20	2.15	3.26
	d_2		1.35	0.21	0.89	1.83
	d_3		-0.05	0.12	-0.36	0.38
	d_4		-1.35	0.21	-1.87	-0.57
	d_5		-2.74	0.31	-3.83	-2.14
1st Stratum		42				
	a		0.71	0.03	0.59	0.75
	b		-0.30	0.93	-1.67	1.98
	d_1		2.79	0.23	2.15	3.24
	d_2		1.38	0.22	1.05	1.83
	d_3		-0.04	0.14	-0.36	0.38
	d_4		-1.32	0.30	-1.87	-0.57
	d_5		-2.82	0.33	-3.83	-2.15
2nd Stratum		**55**				
	a		0.90	0.03	0.82	0.95
	b		-0.34	1.06	-1.90	1.98
	d_1		2.90	0.17	2.52	3.26
	d_2		1.41	0.18	1.01	1.77
	d_3		-0.04	0.12	-0.23	0.35
	d_4		-1.42	0.16	-1.75	-1.13
	d_5		-2.85	0.28	-3.76	-2.35
3rd Stratum		**54**				
	a		1.06	0.04	0.96	1.10
	b		-0.27	1.06	-1.80	1.96
	d_1		2.67	0.13	2.52	3.02
	d_2		1.26	0.19	0.89	1.57
	d_3		-0.08	0.09	-0.23	0.11
	d_4		-1.30	0.14	-1.60	-1.02
	d_5		-2.56	0.22	-3.02	-2.14

Note. *Condition 3: *a*-Stratified constraint.

Table A.13 Item Distribution for Simulated Optimal Pool under Condition 4*

θ \ *a*		0.00 0.75	0.75 0.95	0.95 1.20	Total
-5.0	-3.6	1	0	5	6
-3.6	-2.8	1	2	7	10
-2.8	-2.0	1	0	9	10
-2.0	-1.2	0	1	12	13
-1.2	-0.4	1	1	18	20
-0.4	0.4	3	0	28	31
0.4	1.2	2	0	18	20
1.2	2.0	1	0	10	11
2.0	2.8	1	0	10	11
2.8	3.6	1	0	8	9
3.6	5.0	1	0	5	6
Total		13	4	130	147

Note. *Condition 4: Unconstrained.

Table A.14 Descriptive Statistics of Item Parameters for Simulated Optimal Pool under Condition 4*

Parameter	**N**	**Mean**	**SD**	**Minimum**	**Maximum**
SOP	147				
a		1.01	0.13	0.55	1.10
b		-0.43	0.95	-1.88	1.96
d_1		2.73	0.20	2.40	3.60
d_2		1.27	0.21	0.89	1.91
d_3		-0.08	0.11	-0.30	0.35
d_4		-1.31	0.17	-1.86	-0.63
d_5		-2.61	0.31	-3.76	-2.14

Note. *Condition 4: Unconstrained.

REFERENCES

REFERENCES

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62(4), 569 - 78.*

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bennett, R. E., Morley M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement*, *24*(4), 294-309.

Bennett, R.E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive test. *Journal of Educational Measurement, 34*(2), 162-176.

Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more normal categories. *Psychometrika, 37,* 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46,* 443-459.

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229–255). New York: Routledge.

Burt, W., Kim, S., Davis, L., & Dodd, B. G. (2003). *Three exposure control techniques in CAT using the generalized partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized daptive testing. *Applied Psychological Measurement*, *20*(3), 213-229.

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

Chen, S., Ankenmann, R. D., & Spray, J. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40(2)*, 129-145.

Chen, S., & Cook, K. F. (2009). SIMPOLYCAT: A SAS program for conducting CAT simulation based on polytomou IRT models. *Behavioral Research Methods, 41*(2), 499-506.

Chen, S., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, *58*(4), 569-595.

Chen, S., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement*, *57*(3), 422-439.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, *33*(6), 419-440.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28*(3), 165-185.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335-356.

Davis, L. L., Pastor, D. A., Dodd, B.G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, *4*(1), 24-42.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations initem step values on item and test information in the Partial Credit Model. *Applied Psychological Measurement*, *11*(4), 371-384.

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

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*(1), 5-22.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31(4),* 295-311.

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.) *Computerized adaptive testing: A primer (2nd ed).* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Green, B. F. (1983). The promise of tailored tests, In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp.69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gu, L. (2007). Designing optimal item pools for computerized adaptive tests with exposure controls. Unpublished doctoral dissertation. Michigan State University.

Hau, K.T., Wen, J.B., & Chang, H.H. (2002, April). *Optimum number of strata in the astratified computerized adaptive testing design.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

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

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19* (3), 221-239.

He, W. (2010). Optimal item pool design for a highly constrained computerized adaptive test. Unpublished doctoral dissertation. Michigan State University.

Ho, T. (2010). A Comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the Generalized Partial Credit Model. Unpublished doctoral dissertation. University of Texas at Austin.

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

Johnson, M. A. (2007). An investigation of stratification exposure control procedure in CATs using the Generalized Partial Credit model. Unpublished doctoral dissertation. University of Texas at Austin.

Kim, J. (2010). The comparison of computer-based classification testing approaches using mixed-format tests with the Generalized Partial Credit Model. Unpublished doctoral dissertation. University of Texas at Austin.

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

Koch, W. R., & Dodd, B. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2(4)*, 335- 357.

Leung, C.K., Chang, H.H., & Hau, K.T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63*(2), 257-270.

Lima Passos, V. L., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement, 31(3)*, 213-232.

Lima Passos, V. (2005). Optimal test and sampling designs for polytomous item response theory models. Unpublished doctoral dissertation. Maastricht University, the Netherlands.

Liu, O. L., Lee, H. S., Hofstetter, C.,& Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment, 13*(1), 1-23.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, andof their parallel-forms reliability. *Psychometrika, 48*(2), 233-45.

Morris, S. B., Fortmann, K. A., & Oshima, T. C. (2007, April). An evaluation of the item parameter replication method for DFIT analysis of polytomous items. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muraki, E. & Bock, D. (1999). *PARSCALE: Parameter scaling of rating data (Version 4.1)* [Computer software]. Lincolnwood IL: Scientific Software International.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Parshall, C. G., Davey, T., & Pashley, P. J. (2002). Innovating item types for computerized testing. In W. J. van der Linden and C. A. W. Glas (Eds.*), Computerized Adaptive Testing: Theory and Practice* (pp.129-148). Boston: Kluwer Academic.

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Pastor, D. A., Dodd, B. G., & Chang, H.H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement, 26(2),* 147-163.

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1-20.

Raju, N. S., Fortmann,K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33(2),* 133-147.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practices, 8*(3), 11-15.

Reckase, M. D. (2007). The design of *p*-optimal item bank for computerized adaptive tests. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Reckase, M. D. (2009). Multidimensional Item Response Theory. London: Springer.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph*, No.17.

Samejima, F. (1998). *Expansion of Warm's weighted likelihood estimator of ability for the three-parameter logistic model to general discrete responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego CA.

Segall, D. O., Moreno, K. E., & Hetter, D. H. (1997). Item pool development and evaluation. In W. A.Sands, B. K.Waters,&J.R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington DC: American Psychological Association.

Smarter Balanced Assessment Consortium: Technology-enhanced items guidelines. (2012). Retrieved from www.smarterbalanced.org/

Snyman, J. A. (2005). *Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithm.* New York, NY: Springer Science+Business Media, Inc.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C.A.W. Glass (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-82). Boston: Kluwer Academic.

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

Stocking, M. L., & Swanson, L. (1998). Optimal design of item pools for computerized adaptive tests. *Applied Psychological Measurement, 22,* 271-279.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of Military Testing Association, San Diego, CA.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50(4),* 411–420.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Urry, V. W., (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement, 34,* 253-269.

van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26(4),* 393-411.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201-216.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer-Verlag.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer Academic.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized* *adaptive testing: Theory and practice* (pp. 1-25). Boston: Kluwer Academic.

van der Linden, W. J., Veldkamp, B. P., and Reese, L. M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement, 24(2),* 139-150.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y Kano, & J. J. Meulman (Eds.). New developments in psychometrics (pp. 207-214). Tokyo, Japan: Springer-Verlag.

Veldkamp, B. P. & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*. Dordrecht, the Netherlands: Kluwer.

Wainer, H. (2000). *Computerized adaptive testing: A primer.* Mahwah, N. J.: Lawrence Erlbaum Associates.

Wang, T., Hanson, B. A. & Lau, C. M. (1999). Reducing bias in computerized adaptive testing trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23(3),* 263-78.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, *25*(4), 317-331.

Wang, S., & Wang, T. (2002). *Relative precision of ability estimation in polytomous CAT: A comparison under the generalized partial credit model and graded response model.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weiss, D. J., & Schleisman, J. L. (1999). Adaptive testing. In G. N. Masters, & J. P. Keeves (Eds.). Advances in measurement in educational research and assessment (pp.129-137). Kidlington, UK: Elsevier Science Ltd.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*(4), 299-300.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47* (3), 339-360.

Yi, Q., & Chang, H.H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56* (2), 359-378.

Yi, Q., Wang, T., & Wang, S. (2003). Implementing the *a*-Stratified method with *b* blocking in computerized adaptive testing with the Generalized Partial Credit Model. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.