This is to certify that the
dissertation entitled

Designing Optimal Item Pools for Computerized Adaptive Tests with Exposure Control

presented by

Lixiong Gu

has been accepted towards fulfillment
of the requirements for the

Ph. D. degree in Counseling, Educational Psychology, and Special Education

Mark D. Reckase
Major Professor's Signature

May 10, 2007
Date

# DESIGNING OPTIMAL ITEM POOLS FOR COMPUTERIZED ADAPTIVE TESTS WITH EXPOSURE CONTROLS

By

Lixiong Gu

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2007

ABSTRACT

DESIGNING OPTIMAL ITEM POOLS FOR
COMPUTERIZED ADAPTIVE TESTS WITH
EXPOSURE CONTROLS

By

Lixiong Gu

Computerized adaptive testing requires a well-designed item pool containing an appropriate number of items to build an individualized test that matches the examinee's ability level. An optimal item pool can be defined as a pool consisting of appropriate items for each individual test that is capable of reaching the desired level of precision. It also contains well-balanced items that will achieve optimal item usage and lower the cost of item creation. One of the methods to develop an optimal item pool is Reckase's method (2003), which is a Monte Carlo method to determine the properties of an optimal item pool. This study extends the method for designing item pools calibrated with 3PL and applies it to situations where no exposure control, Sympson-Hetter procedure, or *a*-stratified procedure is imposed to control the item exposure rate. The procedures for designing the item pool and two approaches of simulating test items are presented. The performance of each optimal item pools is evaluated along with the operational item pools.

# DEDICATION

To my wife Yanxuan, my parents, and my sister Lishu

ACKNOWLEDGEMENTS

I am deeply indebted to Professor Mark D. Reckase for his guidance in academics and the dissertation work, as well as my career paths. Without his constant support and insightful comments, this work would not have been possible.

I would also like to thank three other members of my dissertation committee: Dr. Ken Frank, Dr. Richard Houang, and Dr. Hua-Hua Chang for their helpful suggestions on this study.

I am extremely grateful to Dr. Linda Chard for her critiques on the writing and assistance in editing the dissertation.

Thanks also go to Dr. Mary Pommerich and Dr. Daniel Segall, who provided two operational item pools for this study, to Wei He who shared her MATLAB programs on designing item pools for 1PL, and to Raymond Mapuranga for his comments on the early version of my dissertation.

I am also grateful to Dr. Dianne Henderson-Montero and Dr. Venessa Lall, who supported me in balancing my time between operational work at Educational Testing Service and the work on my dissertation.

My deep gratitude goes to my wife Yanxuan for her love and support, and to my parents and my sister, for their understanding and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter I Introduction

Since its introduction in the early 1970 s, computerized adaptive testing (CAT) has been used extensively in educational and psychological assessments (Lord, 1971; Reckase, 1974; Weiss, 1976). The objective of adaptive testing is to build individualized tests by selecting items based on the examinee's current ability estimate. Test takers do not receive questions that are either too difficult or too easy, but are always challenged by appropriate items during the entire course of the testing. From the test administrator's perspective, when examinees are given questions maximizing the information about their ability levels from the item response, reduced standard errors and satisfying measurement precision can be achieved with only a handful of properly selected items. This potentially leads to more efficient item usage and accurate ability estimates (Weiss, 1976). Adaptive tests are often administered with computers, which can quickly update the estimate of the examinee's ability after each item and then select subsequent items based on the estimate. Therefore, when adaptive tests are mentioned, they are often referred to as CAT.

Figure 1.1 shows a typical CAT process. It begins with an item pool (also called item bank) that contains an adequate number of items calibrated using Item Response Theory (IRT) (Lord, 1980). The CAT algorithm is most commonly an iterative process involving a procedure for obtaining ability estimates based upon candidate item performance and an algorithm for sequencing the set of test items to be administered to candidates. A new ability estimate is computed based on the responses to all of the administered items and the "best" next item is administered. This process is repeated until it meets certain stopping criterion, such as time limit, number of items administered

(test length limit), change in ability estimate, content coverage, a precision indicator such as the standard error, or a combination of factors.

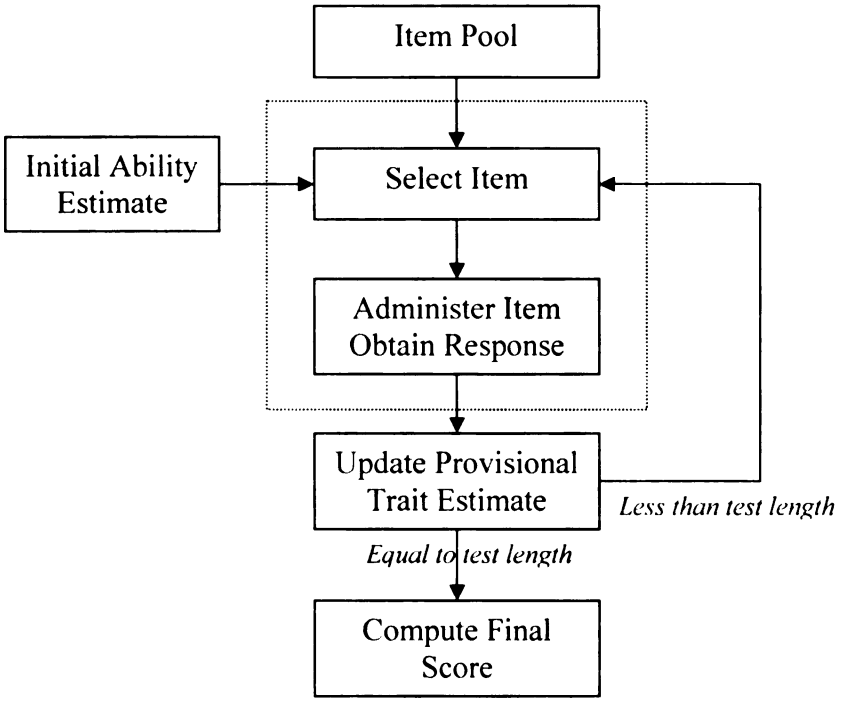


Figure 1.1 Steps of computerized adaptive testing

Computerized adaptive tests offer many advantages over paper and pencil test. Examinees get flexible testing schedules and the opportunity to obtain their scores immediately following administration of the test. More importantly, more accurate ability estimates and higher test reliability and validity are achieved. Test developers may be able to adopt new item formats in computerized administration that are not possible with paper-and-pencil tests.

CAT has also dramatically changed the way tests are administered. Traditional paper and pencil tests are usually administered in classrooms or a large auditorium where desks and chairs are the only requirement for a testing site. CAT, however, needs more expensive equipment (computer system) and a more individualized space to ensure

privacy and security. Specialized test centers with appropriate computer equipment have been setup that will accommodate only a few examinees at a time. However, tests can be offered on a nearly continuous basis—multiple administrations per day throughout the week—to compensate for the large test volumes per administration. Continuous testing gives test administrators and examinees more freedom, yet it also poses difficulties in test development and test security.

Computerized adaptive testing administers slightly different items to different examinees according to their ability levels. It thus requires a large number of diversely distributed items to measure each person with precision and efficiency (Embretson, 2001). Items are needed at the extreme levels of difficulty, although it may be arduous for even an experienced item writer to produce such items. The high costs for item development becomes one of the major impediments to cost-effective implementation of computerized adaptive testing. A quality item usually needs to go through a lengthy and costly procedure such as item writing, content and editorial review, pretest, item analysis, and final review before it can be chosen for an operational test. Particularly, item writing relies on experienced human item writers, who may rather slowly produce items of varying levels of quality. Some of the items that are produced may fail to meet screening criteria or may be unable to achieve adequate psychometric properties based on an empirical tryout.

Until the Internet became widely used, CAT was considered more secure than paper and pencil tests because different examinees receive different test items. It is difficult to artificially promote one's score by merely studying a few items (Wainer, 1990). One would have to learn a large portion of the item pool in order for pre-

knowledge to have any impact on an examinee's score, because the test was individually tailored to his or her ability level. However, the increasingly popular Internet makes it easily for examinees to share test related materials with each other. Presently, a student taking a test on a future date can obtain information from a friend who took the test on the previous day (Davis, 2002).

Item leakage and the high costs of item development have been the primary driving forces for research on designing and maintaining CAT item pools for better item usage. One of the solutions is to design high quality items more efficiently. A promising technique is item cloning, also called model based item generation. This approach starts with a formal description of a set of "parent items" along with algorithms to derive families of clones from them (Bejar, 1993, 1996; Bejar & Yocom, 1991; Hively, Patterson, & Page, 1968; Osburn, 1968). These parents are also known as "item forms," "item templates," or "item shells." More recent item generation research has tried to model the relationship between item parts and its psychometric properties, such as item difficulty and item discrimination (Bejar, Lawless, Morley, Wagner, & Bennett; 2003; Glas & van der Linden, 2003; Graf, Peterson, Steffen, & Lawless, 2005). Items with expected content coverage and psychometric properties (e.g., item difficulty and discrimination) can be produced by varying the essential parts of an item (Deane, Graf, Higgins, Futagi, & Lawless, 2006; Singley & Bennett, 2002). Comprehensive reviews of item-cloning techniques are given in Bejar (1993).

Item cloning gives some hope in producing a large amount of items in a very short time, but it can be very difficult and costly in the process of developing item models for item cloning. A more realistic solution is to optimize the blueprint for the item pool

development. In common computerized adaptive testing practice, items are selected to maximize item information at an estimated ability level so we learn the most about an examinee's ability (Wainer, 1990). It may produce variability in the frequency with which items are used because items differ in terms of the desirability of the characteristics for measuring an examinee's ability level. For example, items with average difficulty will be more often selected for administration because of the assumed normal distribution of the examinee ability level. Additionally, more discriminating items will be selected more often because they tend to be more informative. If item writers know how many items with certain psychometric properties are needed for an item pool, they will not waste resources on developing too many items that are rarely used while creating too few items that may be frequently used.

This current project investigates the applications of a simulation method developed by Reckase (2003, 2004) on designing the blueprint for a CAT item pool. Because an optimal item pool will be different under different test situations, this project also explores the item pool design when different exposure control methods are used, specifically the Sympson-Hetter and $a$-stratified methods. The first chapter briefly introduces the background under which this project is developed. The second chapter reviews the literature on computerized adaptive testing, exposure control, and various methods developed to design the blueprint to optimize the item pool construction. Chapter 3 illustrates Reckase's method in detail and provides extensions of his methods to the applications with the three-parameter logistic IRT model. Following these sections, simulation studies are conducted to demonstrate the item pool design process and

investigate the effectiveness of Reckase's method compared to a real item pool. Finally, the implications of this current study and future directions are presented.

### *1.1 Research Context*

The idea of an item pool is not new. It evolved long before computerized adaptive testing became popular. Even in conventional paper and pencil tests, a well-designed item pool provides test developers and teachers a convenient yet powerful tool to produce high quality tests. Item pools have been called by such terms as "item banks," "question banks," "item collections," "item reservoirs," and "test item libraries." Although distinctions among some of the terms can be made, they all refer to *a relatively large collection of easily accessible test questions* (Millman & Arter, 1984). In other words, an item pool has to include a number of items which exceeds by several times the number to be used in any one test. Items in the pool are indexed, structured, or otherwise assigned information that can be used to facilitate their selection for a test. Millman and Arter (1984) categorized this information into assigned item characteristics (e.g., keywords or other subject matter classifiers) and measured item characteristics (e.g., item difficulty or discrimination). The latter is also known as psychometric characteristics.

The concept of an item pool is expanded in computerized adaptive testing. Two kinds of item pools are distinguished in a typical CAT program. One is often called the master pool, which includes as many items as possibly being created for the testing use. Another kind is the operational item pool, which is a smaller subset of the master pool and by design, has to be small enough so that the computer may easily retrieve the item and minimize item exposure. Yet, it has to be large enough to provide items with the required characteristics. Due to the continuous nature with which many CATs are

administered, the useful life of an operational item or the entire operational item pool is limited. After a certain number of uses, items are retired and put back into the master pool. Some items can be reused only after a reasonably long time.

One question often asked during item pool design is how many items should be in a pool. Ideally, the more items the better, because it allows more choices in test assembly, and seldom do the same items appear in tests repeatedly. With larger pools, it is difficult for examinees to memorize answers. This is a problem in situations where learners have access to the item pool. Larger pools also mean more items that match content, item format, and statistical requirements are available (Millman & Arter, 1984). The caveats, however, are: (a) the items added to the pool should be well written, content valid, and statistically fit; and (b) the total number of items should be manageable and easily retrievable.

In paper and pencil test situations, test items are not reused as often. Millman and Arter (1984) and Prosser (1974) suggested the rule of thumb for the number of items in an item pool be 10 items for each one that could be used on a testing occasion and 50 items for each class hour of presented material. In computerized adaptive testing, Luecht (1998) suggests that between 3,800 and 21,000 items may be needed to begin a CAT program when sufficient pool size, multiple pools, and item pretesting are taken into consideration. Guidelines that have been suggested for the appropriate size of the operational item pool are 150-200 items or from six to twelve times the test length on an operational form (Luecht, 1998; Patsula & Steffan, 1997; Stocking, 1994; Weiss, 1985). However, issues of item exposure, item retirement, and pool rotation may require this number to be much larger.

An often overlooked issue in item pool design is how to construct a blueprint that outlines the optimal composition of items with desirable assigned and psychometric characteristics. The blueprint as the outcome of item pool design can tell item-writers to write items not only by format (multiple-choice or constructed-response) and content coverage, but also by the desired psychometric characteristics of the items. The blueprint is optimal in that it consists of appropriate items for each individual test that is capable of reaching the desired level of precision. An optimal blueprint also contains well-balanced items to achieve optimal item usage and lower the cost of item creation.

Optimizing an item pool may not be an important issue in paper-and-pencil tests where item exposure is not much of the concern. Usually an item pool may require only a few items in each assigned item characteristics, more moderately difficulty items, and some extremely easy and extremely difficult items. In computerized adaptive testing, items have more chances to be overexposed, and the costs of developing new items are so high that an operational item pool needs to have a more balanced item composition in order to reduce the item exposure for often used items and increase exposure for less used items.

A better way to address this problem is to design and develop item pools in a more systematic and empirical manner. The item writing process is usually guided by appropriately designed test specifications that outline the contents attributes and their distributions. Requirement for statistical attributes, such as the range of difficulty, may be provided but are often difficult to satisfy simply because the values of statistical attributes for individual items are not easily predicted. However, at the item pool level they often show persistent patterns of correlation with content attributes. These patterns

can be used to minimize the item-writing effort. Through carefully modeling of the CAT procedure, test specification for the item pool could be developed with computer simulations to forecast the number of items needed with specific attributes (van der Linden, 1999; Reckase, 2003). The methods compared here are for the design of a single item pool and can serve as tools for monitoring the item writing process.

Only a few empirical studies on optimal item pool design have been documented for computerized adaptive testing. Among them there are two bodies of research. One assumes the existence of a master item pool with research focusing on the best way to allocate items into multiple operational item pools. The other focuses on the design of the operational item pool independent of existing items. These studies assume no existing items and focus on design of a blueprint for item pool construction in order to provide precise ability estimation and minimize item exposure.

Stocking and Swanson's (1998) system of rotating item pools assumed the presence of a master item pool from which several smaller operational pools were generated. The number of operational pools each item should be included in can be manipulated so that items with higher exposure rates are assigned to a smaller number of pools and items with lower rates to a larger number of pools. By randomly rotating the operational pools during the testing, uniformly distributed exposure rates for the test items can be achieved. Ariel, Veldkamp, & van der Linden (2004) also presented a mathematical method to calculate the optimal way to allocate items from master item pool into multiple operational pools.

The effectiveness of the rotating pools method, however, relies on the quality of the master pool. Items have to be available in the master pool to be assigned into smaller

operational pools. Apparently, if the master pool contains difficult items only, even the optimally assembled rotating pools would not have the easy items. The fundamental issue in item pool design is how to design an item pool without assuming the existence of the master item pool, and to explore the ideal characteristics, such as the size and the item distribution the item pool should have to function efficiently.

Research concerning the design of the item pool from scratch focuses on developing a *blueprint* for an item pool — a document that specifies the attributes of the items needed in a new pool or an extension of an existing pool. The blueprint is designed to allow for the assembly of a prespecified number of test forms from the bank, each with its own set of specifications. The resulting item pool would allow CAT procedure to generate adequate measurement precision for a majority of the test takers even with the constraints of exposure control or content balancing. A favorable consequence is that the number of unused items in the bank is also minimized. As will become clear, the blueprint not only specifies the number of items with each content coverage, but also the number of items with certain psychometric properties, particularly the range of IRT parameters.

Boekkooi-Timminga (1991) used integer programming to calculate the number of items needed for future test forms. She used a sequential approach that maximized the test information function (TIF) under the one-parameter logistic (Rasch) model. These results were then used to improve on the composition of an existing item bank. Subsequently, several methods for the construction of rotating item pools have been demonstrated in empirical studies, some achieved the design goal with integer

programming methods (for a review of these methods, see Ariel, Veldkamp, & van der Linden, 2004).

Veldkamp & van der Linden (1999) described five steps to design an optimal blueprint for a CAT item pool with a mathematical programming method. First, a set of specifications for the CAT is analyzed and all item attributes figuring in the specifications are identified. Second, using the specifications, an integer programming model for the assembly of the shadow tests in the CAT simulation is formulated. Third, the population of examinees is identified and an estimate of its ability distribution is obtained, for example, from historical data. Fourth, the CAT simulation is carried out using the integer programming model for the shadow tests and sampling simulees from the ability distribution. Counts of the number of times items from the cells in the classification table are collected. Fifth, the blueprint is calculated from these counts, adjusting them to obtain optimal projections of the item exposure rates.

As the basic idea of mathematic programming is to optimize the resource allocation under the assumption of limited resources, in this case the optimization is constrained by the resources prespecified. It thus assumes a design space that is defined as the Cartesian product of all item attributes figuring in the specifications of the tests in the program. Each combination of the attributes represents a virtual item. For example, a combination of content coverage and *a*-, *b*-, *c*- parameters represents a virtual item available for optimal item pool. Because item parameters are real numbers, there are infinitely many combinations of item attributes. To simplify the problem, discrete values are chosen to represent the possible values of the item parameters. For example, $b$ = -3.0, -2.9, , 2.9, 3.0 represent the possible $b$-parameter values and $a$ = 0.1, 0.2, , 2.9, 3.0 represents the

possible values for $a$-parameters. Any combination of the $a$- and $b$-parameters represents a virtual item. Veldkamp & van der Linden (1999) demonstrated that modeling constraints for the CAT version of the GMAT led to a design space containing 12,096 items.

Formulating constraints is an important step in the mathematical programming method. Van der Linden (1998) laid out three kinds of constraints based on their mathematical types: categorical item attributes, quantitative item attributes, and inter-item dependencies. Categorical item attributes are attributes that characterize the content, format, or author of items. Quantitative item attributes are certain quantitative attributes items have, such as word counts, difficulty parameters, and discrimination indices. Inter-item dependencies deal with possible relations of exclusion and inclusion between the items in the pool, such as items in so-called enemy sets, in which items cannot be included in the same sets.

The advantage of the mathematic programming method is that it is able to model complicated test specifications. Once the constraints are identified and transformed to numerical constraints, special software is available to simulate the optimal item pool. However, item pool design with mathematical programming method is closely tied with shadow test procedure in item selections and requires the knowledge of special optimization software. Depending on the way item attributes are partitioned, the design space can be very large and the simulation process becomes computationally arduous.

Reckase (2003, 2004) took a slightly different approach and avoided using mathematic programming. This approach does not assume pre-existing items. Instead, items are simulated (in terms of the IRT parameters) to match the current ability

estimates to provide sufficiently optimal information. Reckase's method first partitions the target item pool into smaller ones based on different non-statistical attributes, such as content. Then the CAT process is simulated to construct the small items pools simultaneously. The simulation starts with an examinee randomly drawn from the expected examinee distribution to receive the adaptive test. Each item is simulated to be the optimal item based on the current ability estimate. The same procedure is repeated for subsequent examinees and the items needed to support a large sample of examinees is tallied and becomes the optimal item pool. Exposure control rules can be built into the simulation to decide how many times an item can be reused. This procedure has been demonstrated successfully with widely available programming software in the design of CAT item pools for TABE and NCLEX.

### *1.2 Summary*

The present study reports on the development of optimal item pools for computerized adaptive tests in the investigation of the relative merits of two different strategies. A modified version of Reckase's method is applied to designing optimal item pools calibrated with three parameter logistic models. Sympson-Hetter and *a*-stratified exposure control methods, as well as content balancing, are investigated.

For the purpose of this research, it was desired to have an operational pool of items measuring an empirically significant dimension of ability, while balancing the content areas measured. Operational item pools for two sections of the CAT-ASVAB were chosen as the design target in this study. The final item pools were designed to meet the criteria described by van der Linden (2000): (1) it would be sufficiently large to allow several thousand overlapping subtests to be drawn from its items; (2) the items would

span the entire range of item difficulty relative to the population of interest; and (3) it would consist of an appropriate mix of high and low discriminating items to lower the item creation cost while meeting the needs of test precision.

This study compared simulated optimal item pools to operational item pools on item distribution and performance for examinees randomly sampled from expected examinee distribution. The simulation study took into consideration the distribution of the examinee population, content balancing, and the expected precision of ability estimates.

The following research questions are investigated in this study:

1. What does the optimal item pool designed for a computerized adaptive test look like when the item selection procedure imposes no exposure control, when it incorporates Sympson-Hetter method, or when it incorporates the *a*-stratified method?

2. What do the optimal item pools designed for a computerized adaptive test look like when the test does not need content balancing, and when the test needs content balancing?

3. Do optimal item pools designed by a Monte Carlo simulation perform better than the real operational item pool in terms of empirical criteria?

# Chapter II Item Pool Design and Components of Computerized Adaptive Testing

An optimal item pool design is based on the indepth understanding of the mechanism of CAT. This chapter introduces computerized adaptive testing, its history, pro and cons, and the components of the CAT procedure, which includes the item pool, item selection procedure, ability estimation method, and stopping rules. Special attention will be given to exposure control procedures, which are an integral part of the item selection procedure, and how the design of the item pool should be based on the analysis of its relationship with other CAT components.

## *2.1 Brief History of the Computer Adaptive Testing*

Computerized adaptive testing (CAT), as its name suggests, is adaptive testing delivered by computers. While CAT has only recently become a major force in measurement practice, the idea of adaptive testing is not new. It has always been recognized that too easy or too difficult items contribute little to the information about an examinee's ability level. By eliminating the need to administer items of inappropriate difficulty, adaptive testing can shorten testing time, increase measurement precision, and reduce measurement error due to boredom, frustration or guessing (Wainer, 1990).

The first adaptive test is known to be Alfred Binet's (1905) intelligence test, which is still in use today in a more modern version. Since the concern was with the diagnosis of the individual child, rather than the group, Binet realized he could tailor the test to the individual by a simple strategy—first rank-ordering the items in terms of difficulty, then starting to test the child at what he deemed to be a subset of items targeted at his

approximation of the candidate's ability. If the child gave the correct answer, harder item subsets were administered until the child answered a few questions in a row incorrectly. If the child failed the initial item subset, then easier item subsets would be administered until the child succeeded frequently. From this information, the child's ability level could be estimated.

Lord's (1980) Flexilevel testing procedure and its variants, such as Henning's (1987) Step procedure and Lewis and Sheehan's (1990) Testlets, are a refinement of Binet's method. The items are stratified by difficulty level, and several subsets of items are formed at each level. The test then proceeds by administering subsets of items, and moving up or down in accord with the success rate on each subset. After the administration of several subsets, the final ability estimate is obtained. Though a crude approach, these methods can produce approximately the same results as more sophisticated CAT techniques (Yao, 1991).

The use of computers facilitates a further advance in adaptive testing with convenient administration and selection of single items. Reckase's (1974) study is an early example of this methodology of computer-adaptive testing (CAT). Since the mid 1990 s, many well known large-scale high stake testing programs, such as GRE, GMAT, NCLEX and LSAT, have switched from paper and pencil tests to computerized adaptive tests.

### *2.2 Pros and Cons of CAT*

The advantages and cautions of CAT have been well document (e.g., Rudner, 1998; Wainer, 1990, Wainer & Eignor, 2000). One advantage is that, in general, CAT greatly increases the flexibility of test management (e.g., Grist, Rudner, and Wise, 1989;

Weiss and Kingsbury, 1984). Tests are individually paced so that an examinee does not have to wait for others to finish before going on to the next section. Self-paced administration also offers extra time for examinees who need it, potentially reducing one source of test anxiety. A number of options for timing and formatting, notably interactive audio and video, can be offered. Therefore, it has the potential to accommodate a wider range of item types. Besides flexibility, CAT increases test efficiency by shortening the test length without reduction in measurement precision. A shorter test reduces examinee frustration, boredom, and fatigue, a factor that can significantly affect an examinee's test results. Since examinees see only those items appropriate for their ability level, they remain engaged and challenged. In addition, individualized test leads to easy removal of faulty items. With CAT, a poorly performing or incorrect item would only affect a segment of the test-takers, and even for those, the self-correcting nature of CAT would make it unlikely there would be any impact on the pass-fail decision.

### *2.3 Components of Computerized Adaptive Testing*

Reckase (1989) listed four major components of a computerized adaptive test: the item pool, the item selection procedure, the scoring (ability estimation) procedure, and the stopping rule. Item exposure control and content balancing have recently been extensively studied to constraint the item selection so that items are selected not only by their statistical appeals but also content specifications and security concerns. An optimal item pool should be determined by the other components of the CAT, namely test length, expected distribution of the examinee population, ability estimation and item selection procedure, and target item exposure and overlap rates (Bergstrom & Lunz, 1999).

### 2.3.1 Item Pool

The adaptive feature of CAT makes it unnecessary to use pre-designed test forms like paper and pencil test. It requires an item pool from which all tests will be drawn. In practice, there are two kinds of item pools: one is called the master pool and another is called the operational pool. The master pool is an inventory of test items maintained to supply the testing program. An operational pool is a pool of items from which individual adaptive tests are actually assembled. Typically, a master pool is much less structured than an operational pool. Its items may be in various stages of development, while the items in an operational item pool have passed all preparatory stages and the pool is ready for test assembly. The focus of this study is to investigate the methods for designing an optimal operational item pool.

Ideally, the item pool would have a sufficient number of high quality items to allow several thousand of overlapping subtests to be drawn from its items. It would have a sufficient number of items in each desired content area to meet the test specifications. It would span a wide range of item difficulties relative to the population of interest to allow the CAT to estimate ability levels for a broad range of examinees (Urry, 1977). In addition, care must be taken to ensure that the item pool would consist of appropriate items to reduce the over- and under-exposure rate while meeting the test precision requirement (Davis, 2002, Wainer, 1990). Guidelines for the appropriate size of the item pool are 150-200 items or from six to twelve times the test length on an operational form (Luecht, 1998; Patsula & Steffan, 1997; Stocking, 1994; Weiss, 1985). However, issues of item exposure, item retirement, and pool rotation may require this number to be much larger. Due to the continuous nature with which many CATs are administered, the useful

life of an item or an item pool is limited. Luecht (1998) suggests that between 3,800 and 21,000 items may be needed to begin a CAT program when sufficient pool size, multiple pools, and item pretesting are taken into consideration. Strategies to extend the life of a pool, such as drawing multiple overlapping pools from an item vat (Patsula & Steffan, 1997), have been proposed. However, the cost and effort to create and maintain a CAT item pool remains formidable and far exceeds that of paper and pencil testing, which makes optimal design of item pool more important.

An item pool is not only a reservoir of items, but also an organized list of items with clearly defined attributes attached to them. Van der Linden (2005) distinguished three types of item attributes: quantitative, categorical, and logical. Quantitative attributes are item attributes that take on numerical values. Examples of quantitative attributes are word counts, expected response times, statistics such as item p-values and IRT parameters, and frequency of previous item or stimulus usage. Categorical attributes divide or partition the item pool into subsets of items with the same attribute. Examples of categorical attributes include content category, response format of items (e.g., constructed response or multiple-choice), and use of auxiliary material (e.g., graph or table). Logical attributes differ from quantitative and categorical attributes in that they are not properties of single items or tests but of pairs, triples, and so forth. The logical attributes involve relations of exclusion and inclusion between items or tests. For example, a relation of exclusion between items exists if they cannot be selected for the same test because one has a clue to the solution of the other (so-called "enemy items"). A relation of inclusion exists if items belong to a set with a common stimulus and the selection of any item implies the selection of more than one.

2.3.2 Scoring Procedure

One of the advantages CAT has is the ability to administer items suitable to an examinee's ability level. This is achieved by repeatedly estimating the ability level after each item is administered. At the beginning of the test administration, an initial value for ability level is arbitrarily provided since no information is known about an examinee and no item is administered. This value is commonly the expected mean ability level of the testing population or a random number around the expected mean ability level. If prior information is available, it may help determine the initial value that may be closer to the examinee's real ability level, thus facilitating the proceeding estimations.

After each item is administered, an examinee's ability level is re-estimated, based on his or her responses to all previously answered items. Maximum likelihood estimation (MLE) and Bayesian estimation approaches are two commonly used ability estimation methods.

MLE determines the most likely ability level for an examinee given the response string to items with specified parameters, by multiplying together the individual probabilities of a correct or incorrect response given theta to compute a joint probability with the function,

$$L(u \mid \theta) = \prod_{i=1}^{n} P_i(u_i \mid \theta, a_i, b_i, c_i) \qquad (3)$$

where $P_i(u_i \mid \theta, a_i, b_i, c_i)$ is the probability of getting response $u_i$ on item $i$ given an examinee's true ability $\theta$ and item parameters , and $n$ is the number of items. The maximum likelihood estimate of an examinee's true ability $\theta$ is $\hat{\theta}$, the value that maximizes the likelihood function (or, equivalently, the log likelihood function).

Mathematically, this can be done by taking the derivative of the likelihood function, setting the result equal to zero and solving for $\hat{\theta}$. Iterative numerical methods such as Newton-Raphson method (Wainer, 1990) typically are used to solve this equation.

MLE ability estimates are popular in CAT contexts due in part to their desirable theoretical properties such as asymptotic consistency and asymptotic normality. Problems, however, occur in solving the likelihood equation when examinees get all items correct or all items incorrect, because such response pattern yield unbounded ability estimates. These problems are often handled in CAT by setting arbitrary minimum and maximum ability estimates for such response patterns (e.g., -4 and +4) or by using a Bayesian-based ability estimate until examinee answers at least one item correctly and one item incorrectly.

Owen's (1969) Bayesian sequential ability estimation technique was proposed as part of his adaptive testing strategy, which selects items that minimizes the expected value of the Bayesian posterior variance. This ability estimation procedure, however, has proven useful in adaptive testing strategies using other item selection criteria, as well.

Owen's Bayesian method begins with a prior distribution of ability – in effect, as an assumption that the examinee is a member of a population with a normal distribution of ability, with known mean and variance. After each test question, the mean and variance are updated using a statistical procedure that combines the information in the prior distribution with the observed score (right or wrong) on the most recent test question, and the parameters of that question's IRT model. The updated values of the ability distribution parameters specify a normal "posterior" distribution, which is used as the prior distribution for the next question. This process continues until the end of the

test. At that point, the posterior mean is used as the estimate of the examinee's ability scale location. Owen's formula for updating the prior mean is as follows:

$$\mu(\theta_i \mid u_i) = \frac{\int \theta P(u_i \mid \theta) h(\theta) d\theta}{\int P(u_i \mid \theta) h(\theta) d\theta} \qquad (4)$$

Owen (1975) showed that after each item is administered the estimates for $\hat{\theta}_{ij}$ and $\hat{\sigma}_{ij}^2$ are:

$$\hat{\theta}_i = \hat{\theta}_{i-1} - \hat{\sigma}_{i-1}^2 (a_i^{-2} + \hat{\sigma}_{i-1}^2)^{-\frac{1}{2}} [\phi(B_i) / \Phi(B_i)](1 - u_i / A_i), \qquad (5)$$

$$\hat{\sigma}_i^2 = \hat{\sigma}_{i-1}^2 - (\hat{\sigma}_{i-1}^2)^2 (a_i^{-2} + \hat{\sigma}_{i-1}^2)^{-1} \lambda_n, \qquad (6)$$

where

$u_i$ is the item response (answer is correct when $u_i = 1$ and wrong when $u_i = 0$)

$\phi(B_i)$ is the standard normal probability density function of $B_i$,

$\Phi(B_i)$ is the standard normal cumulative density function of $B_i$,

$\lambda_n = [\phi(B_i) / \Phi(B_i)](1 - u_i / A_i)[(1 - u_i / A_i)\phi(B_i) / \Phi(B_i) + B_i]$,

$B_i = (b_i - \hat{\theta}_i) / (a_i^{-2} + \hat{\sigma}_{i-1}^2)^{-\frac{1}{2}}$, and

$A_i = c_i + (1 - c_i)\Phi(-B_i)$

Adaptive test scoring using Owen's procedure takes into account just one item response at a time. All previous information is absorbed into the parameters of the prior distribution, which changes after each question. Because of the added prior information, the Bayesian procedures have the advantage of smaller standard errors than with MLE for the same number of items administered. However, use of a bad prior can result in the

need to administer more items to recover and a regression toward the mean in ability estimation tends to occur.

MLE cannot estimate ability level until one correct and one incorrect response are obtained. Thus, MLE cannot be used after the first item or in the case in which an examinee answers all items correctly or incorrectly. Therefore, with MLE, *ad hoc* procedures must be implemented to cover cases in which no finite value $\theta$ is available (Thissen & Mislevy, 2000). One soultion is to assign an examinee's interim ability estimate to be half the distance between the current ability estimate and a maximum or minimum item difficulty value depending on whether all the responses are in the upper or lower half of the response scale (Koch & Dodd, 1989). Bayesian procedures, on the other hand, can obtain an ability estimate after the first item response. This makes it a favorable estimation method in some CAT programs.

### 2.3.3 Item Selection Procedure

In CAT, new items are selected adaptively with respect to a provisional estimate of the examinee's ability level based on responses to those items already administered (Davey & Parshall, 1995). The two strategies currently most widely used for item selection in CAT are maximum information (MI) (Brown & Weiss, 1977) and maximum posterior precision (MPP) (Owen, 1975).

The MI strategy selects the item that maximizes the Fisher information value at the examinee's current ability estimate. Let $P_j(\theta)$ denote the item response function for item $j$ and $Q_j(\theta) = 1-P_j(\theta)$. Then, for a dichotomously scored item, Fisher information is (Lord, 1980):

$$I_j(\theta) = \left[\frac{\partial P_j(\theta)}{\partial \theta}\right]^2 \Big/ P_j(\theta)Q_j(\theta) = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)Q_j(\theta)} \qquad (1)$$

where $P_j(\theta)$ is the probability of a correct response, given $\theta$, and $Q_j(\theta)$ is the probability of an incorrect response. Plugging item parameters in Equation (1), it can be simplified for the dichotomous three-parameter logistic item response model (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980):

$$I_j(\theta) = \frac{D^2 a_j^2 (1 - c_j)}{(c_j + e^{DL_j})(1 + e^{-DL_j})^2} \qquad (2)$$

where $L_j = a_j(\theta - b_j)$, $D = 1.7$, $a_j$ is the item discrimination parameter, $b_j$ is the item difficulty parameter, and $c_j$ is the pseudo-chance level parameter (i.e., the probability of a very low $\theta$ examinee correctly answering the item). Equation (2) indicates that the item information increases as $b_j$ approaches $\theta$, as $a_j$ increases, and as $c_j$ approaches 0 (Hambleton et al., 1991).

Unconstrained maximum information selection chooses an item $i$ that maximizes the Fisher information evaluated at $\hat{\theta}_i$, the provisional proficiency estimate for the examinee after $n$ proceeding items. When the items that constitute CAT are selected using MI, the precision of $\hat{\theta}$ increases as each item is administered (Hambleton et al., 1991).

In practice, maximum information item selection is often based on a previously computed table in which items are sorted by the information they provide at each number of proficiency values (an "Info Table"). Item selection is equivalent for all $\theta$'s in an interval around a tabulated value. Rather than evaluating Fisher information for each item

in the pool at the current value of $\hat{\theta}$ each time it needed to select the next item, it need only be evaluated once for each item at each tabulated point. Item selection based on an Info Table is slightly less efficient but less computationally burdensome than maximum information item selection.

The MPP is also called Owen's Bayesian method, which selects an item that maximizes the expected posterior precision of the ability estimate, or, conversely, to minimize the expected posterior variance of the ability estimate. At the initial stages of a test, MPP item selection might differ from MI item selection due to its use of the posterior distribution. However, at later stages, the posterior variance approaches the reciprocal of the test information (Chang & Stout, 1993). Therefore, MPP strategy might provide results that are similar to the MI strategy. MPP is computationally easier than the MI for its quick approximation to the posterior ability distribution (Davey & Parshall, 1995). This advantage makes MPP more popular when computing power is limited. However, increases in available computing power and the fact that with the MPP strategy estimated ability level varies as a function of item order have made maximum information more widely used (Wainer, 1990). A hybrid strategy was also developed using Owen's Bayesian sequential procedure to update ability after each item response and selecting items sequentially by referring to precomputed item information lookup tables. This strategy was evaluated by several researchers and seemed to retain the advantages of MI and MPP and avoid their disadvantages (Wetzel & McBride, 1986).

These statistically motivated item selection procedures can be tempered by practical considerations such as item exposure rates and content balancing, both of which will be described in detail later in this chapter.

### 2.3.4 Stopping Rule

There are two methods to determine when to stop administering items in a CAT. One is "fixed length," which requires all examinees to take the same number of items. With the same test length for all examinees, test-taking time is similar across examinees, making test administration more predictable. However, measurement precision will differ across ability levels, which causes difficulty in reporting test reliability. The other method, the "variable length" method, requires that examinees continue to take items until reaching a pre-specified level of precision. It is, however, hard to explain to non-experts why different items are administered. In terms of the item pool use, variable length tests tend to perform better than fixed length tests as they minimize test length (Bergstrom & Lunz, 1999). However, simulation studies showed that fixed-length testing was more efficient because highly informative items were typically concentrated over a restricted range of ability examinees. In variable-length testing, examinees falling outside of a certain range of ability tend to receive long tests, with each additional item providing very little information and fatigue influencing the test precision (Segall, Moreno, & Hetter, 1997). In practice, some adjustments are made to compensate for shortcomings of either method. For example, in some certification tests, every examinee has to take a minimum number of items to cover enough contents. If the desired precision level cannot be reached for an examinee whose ability is close to the cut-score, more items are given until the precision level or a maximum test length is reached.

## *2.4 Practical constraints in Item Selection*

Without additional constraints, item selection algorithms select a single best item at each step of testing for each examinee. In practice, some constraints that play no formal

role in an IRT model are often imposed on item selection algorithms to achieve desirable patterns of item usage. For example, a given examinee should not receive a particular item twice; the number of items in each content area must not exceed certain proportions specified by test specification; and the same item should not be administered more than a certain number of times. Among these practical considerations, item exposure control and content balancing are two of the most addressed in CAT operations.

The exposure rate of an item is defined as the ratio between the number of times the item is administered and the total number of examinees. Because CAT's are administered frequently to small groups of examinees, there is a risk that items with high exposure rates might become known to examinees (Mills & Stocking, 1996). Thus, high item exposure rates can decrease test security.

Item selection based purely on a statistical model, such as maximum information or maximum posterior precision, is the primary reason that leads to high item exposure rates. For example, if no auxiliary information is used to start the CAT, every examinee sees the same item first and one of a single pair of items second. Those three items would soon become public knowledge for any widely used CAT. Research has demonstrated that typically, CAT item selection results in the fact that not only are the pool's most informative items most often administered, but also that a very small percentage of the pool's items account for a very large percentage of administered items (Wainer, 2000; Wainer & Eignor, 2000). In other words, the most popular items are popular by a large margin, resulting in an exponential decline in item usage as rank increases. Furthermore, with maximum information item selection, two examinees with the same ability estimate will likely see the same item (Hetter & Sympson, 1997).

While rarely used items seem to be a waste of resources spent on item creation, frequently exposed items may cease to be a valid measure of the ability because examinee may have prior knowledge of items either from taking a pre-test or from friends who took the test before (Parshall, Davey, & Nering, 1998). The probability the same examinee would take items he/she saw in pre-test can be minimized if items that are pretested together are put into separate item pools (Stocking & Lewis, 2000). It is hard, however, to deal with the threat from what Luech (1998) called "Examinee collaboration networks (ECNs)," which are global groups of examinees who seek to pool their resources and test experience to discover a sufficient number of test items from an item pool to artificially increase scores (Davis, 2002). This situation is exacerbated by the fact that tests are often continuously administered in CAT (Stocking & Lewis, 2000).

To lower the risk of overexposing test items, mechanisms are imposed on the item selection function to control the item exposure rate. Estimation efficiency is thus traded off against a more evenly distributed item exposure, a benefit from the point of view of test security.

Another consideration on item selection is the balance of item content. Content areas in a single test are often associated with the notion of multidimensionality. There have been extensive debates on whether or not a test composed of different content areas should be treated as unidimensional or multidimensional test. If unidimensional IRT is used in item selection and scoring procedures in CAT, one of the major assumptions is that performance on items within a given content area can be characterized by a unidimensional IRT ability or ability. Violations of the unidimensional adaptive testing model may have serious implications for validity and testing fairness (Segall, Moreno, &

Hetter, 1997). When content areas are disparate or introduce additional dimensionality, a logical option is to use a multidimensional model and estimate separate ability levels within a single test (Parshall, Davey, & Nering, 1998). Another option is to split the item pool by administering separate tests, with separate ability estimations for each content area (Segall, Moreno, & Hetter, 1997). However, it is often not practical to divide a single test into several separate ones with the goals of the testing program (Thissen & Mislevy, 2000). However, when content areas are shown to measure a single ability dimension it is possible to design the item pool with item quantities proportional to the desired content coverage for each examinee's test (Green, Bock, Humphreys, Linn, & Reckase, 1984; Segal, Moreno, & Hetter, 1997). For example, in a test of general science it might be desirable to constrain the 15-item adaptive test to include seven life science items, seven physical/earth sciences items, and a single chemistry item, all in prespecifed ordinal positions in the test.

Similar to item exposure control, programmatic restrictions to the item selection procedure would be necessary to ensure the desired content coverage during the CAT. In a fixed-length CAT, the ordinal positions for each item type or content may be specified as *a priori*, or a spiraling scheme rotating through the various kinds of items may be used. With variable length CAT, the algorithm rotates through the various types of items so that balance is approximately maintained at each possible stopping point.

Like item exposure control, one possible drawback for any content balancing method is that the most informative item in the selected content area may not be the most informative item available in the item pool. It might threaten measurement precision (in a fixed length test) and could result in longer tests (in a variable length test) due to

administration of sub-optimal items. Some alternatives were proposed to balance the content area while maintaining the test efficiency. The first is to balance fixed proportions of information based on the same fixed target content percentages of items required by the test specifications. The second is to balance the proportions of information based on target content percentages of total test information that are conditional on estimated ability (Davey & Thomas, 1996; Thomasson, 1997).

### *2.5 Exposure Control Methods*

Way (1998) classified exposure control procedures into two categories: randomization (Davis and Dodd, 2001; Kingbury and Zara, 1989; Lunz and Stahl, 1998; McBride and Martin, 1983) and conditional selection procedures. Instead of always administering the most informative item, randomization procedures select several items near the maximum information level and then randomly administer one from the selected items. Although relatively easy to implement, for randomization procedures it is hard to specify a maximum exposure rate for specific items. Conditional selection procedures juggle this problem by setting exposure control parameters to each item and adjust the administration rate accordingly. Conditional selection procedures, however, require a time-consuming iteration process to obtain exposure control parameters. If the item pool or the ability distribution of the examinee population changes, it has to go through the same process to reset the exposure control parameters. In addition to the randomization and conditional selection procedures, Chang and Ying (1996) developed the *a*-Stratified procedure in which items with low discrimination are administered first, followed by items with high discrimination as more accurate estimations of the examinees' ability levels are determined.

2.5.1 Sympson-Hetter Exposure Control

The Sympson-Hetter exposure control procedure (Sympson & Hetter, 1985) is one of the most commonly used conditional selection procedures. This procedure assigns to each item an exposure control parameter value that is based on the frequency of item selections during an iterative CAT simulation. Items with high administration frequencies are assigned smaller exposure control parameters, which range from zero to one. During the test operations, the exposure control parameter of the selected item is compared to a random number, which also ranges from zero to one. If the exposure control parameter is larger than the random number, the item is administered. If it is smaller, the item is put back into the item pool and the same process is applied to the next best item. The item exposure control parameter is like a threshold. By controlling the thresholds, the S-H method limits the administration of frequently used items in CAT and ensures a maximum item exposure rate for less often used items.

The exposure control parameters in the S-H method are usually set by a series of iterative simulation of real CAT administrations. Simply put, it is the ratio of the target exposure rate to the probability of the item being selected in testing. How it works can be shown as follow:

Let $S_j$ denote the selection of item $j$ for a randomly sampled examinee, and let $A_j$ denote the administration of that item. The exposure rate for item $j$ can be interpreted as $P(A_j)$, the probability that item $j$ is administered to a randomly sampled examinee. The S-H method separates item administered from item selected by the probability relation $P(A_j) = P(A_j|S_j)P(S_j)$ and controls $P(A_j)$ by controlling $P(A_j|S_j)$, the proportion of selections that lead to administration. For any given exposure rate $r_j > 0$; $P(A_j) \leq r_j$ can

be achieved by setting $P(A_j|S_j) \leq r_j/P(S_j)$. If $P(S_j)$ is known or can be approximated, this method can be easily implemented by generating a uniform (0,1) random variable.

Hetter and Sympson (1997) described steps of setting exposure control parameter $K_i$ for the items. The probability of administering one item depends on the relationship between a random number $k$ and the item's $K$ value. Given that one item is selected, a random number $k$ is generated and compared to $K_i$, if $k < K_i$ the item will be administered, otherwise the item will be retained and the next highest information item will be selected and the same procedure is applied. There are five steps to set the exposure control parameters:

Step 1. Generate the first set of $K_i$ values, which are 1.0 for every item. This results in an $n$-by-one vector for $n$ items. Denote each $i^{th}$ element of it as $K_i$ associated with item $I$.

Step 2. Administer adaptive tests to a random sample of simulees. For each item, identify the most informative item $i$ available at the examinee's current ability estimate then generate a pseudo-random-number $x$ from the uniform distribution (0,1). Administer item $i$ if $x$ is less than or equal to the corresponding $K_i$. Whether or not item $i$ is administered, it is excluded from further administration for the remainder of this examinee's test. Note that for the first simulation, all the $K_i$'s are equal to 1.0 and every item is administered, if selected.

Step 3. Keep track of the number of times each item in the pool is selected ($NS$) and the number of times that it is administered ($NA$) in the total simulee sample. When the complete sample has been tested, compute $P(S)$, the probability that an item is

selected, and $P(A)$, the probability that an item is administered given that it has been selected, for each item:

$P(S) = NS/NE$

$P(A) = NA/NE$

where $NE$ = total number of examinees.

Step 4. Use the value of the expected exposure rate $r$, and the $P(S)$ values computed above to compute new $K_i$ as follows:

If $P(S) > r$, then new $K_i = r/P(S)$

If $P(S) <= r$, then new $K_i = 1.0$

Make sure that there are at least $n$ items in the item pool that have new $K_i = 1$.

Step 5. Given the new $K_i$, go back to step 2. Use the same examinees and repeat steps 2, 3, and 4 until the maximum value of $P(A)$ that is obtained in step 3 approaches a limit slightly above $r$ and then oscillates in successive simulations.

The $K_i$ obtained from the final round of computer simulations are the exposure-control parameters to be used in real testing.

The S-H method effectively limits the exposure rates of all items. However, because items that are not selected cannot be administered, items with small probabilities of being selected will still have small exposure rates; thus, the S-H method does not increase exposure rates for underexposed items. In addition, while the exposure of an item across $\theta$ levels may be controlled, the same control may not hold for examinees at a particular level of ability. For instance, even though the exposure of an item may be controlled such that it is administered to no more than 30% of the examinees overall, it may be administered to examinees of high ability 100% of the time. Furthermore,

implementation of this method requires knowledge about $P(S_j)$, which is associated with the $\theta$ distribution of the examinee population. Hence, it is necessary to specify this distribution *a priori* and then approximate the value of $P(S_j)$ using simulation.

Many variations of the S-H technique were proposed afterwards. Parshall, Davey, and Nering (1998) developed the conditional Sympson-Hetter procedure in which the exposure control parameters are determined based on ability level. Stocking and Lewis (1995) extended the technique to utilize a multinomial model and proposed another version of the technique (Stocking & Lewis, 1998) that conditions the exposure control parameter not only on the frequency with which the item is selected but also on $\theta$ level. This addition to the S-H technique (often referred to as the conditional Sympson-Hetter technique, or CSH, when a multinomial model is not used) is desirable for it overcomes the major disadvantages of SH method by establishing an exposure control parameter for each item at a number of different $\theta$ levels.

### 2.5.2 *a*-Stratified Adaptive Testing

The use of stratification testing based on item response theory is not new. Poststratification in which stratification is applied according to an examinee's test results has been widely used in assessing differential item functioning (Dorans & Kulick, 1986; Holland & Thayer, 1988; Shealy & Stout, 1993). Weiss (1973) proposed a stratified CAT design in which stratification was performed according to item difficulty. Chang and Ying's (1999) *a*-stratified adaptive testing (STR) design was proposed primarily to address the plagued concern of overdrawing items with high discriminating indices in item pools.

CAT item selection procedure based on maximum item information tends to select items with higher discrimination parameters more often in the beginning stage of the test than those with lower discrimination parameters. Because estimation of $\theta$ tends to be quite inaccurate early in the test, it seems a waste to use high discriminating items at this point (Chang et al., 2003). With the STR method, the item pool is divided into a number of strata, based on the values of the discrimination parameter of the items. During the test, item selection is always constrained to one stratum, selecting items with maximum information (van der Linden & Pashley, 2000) or the smallest distance between the value of their difficulty parameter, $b_i$ and the current estimate of $\theta$ (Chang et al.,1999). Early in the test, items are administered from the stratum with the lowest value for the discrimination parameter. However, as the test progresses, strata with higher values are used. As a consequence, *a*-stratification forces a more balanced exposure for all items, particularly if the strata in the item pool are chosen to have equal size and $nr \equiv n/R$ (Chang et al., 2003).

A simple *a*-stratified selection method can be described as follows:

1. Partition the item bank into $K$ levels according to the item $a$ values;
2. Partition the test into $K$ stages;
3. In the $k^{th}$ stage, select nk items from the $k^{th}$ level based on the similarity between $b$ and $\hat{\theta}$, then administer the items (note that $n_1 +$ $+ n_K$ equals the test length);
4. Repeat Step 3 from $k = 1, 2,$ $K$.

Note that item selection with the *a*-stratified design is based on matching *b*-parameter with $\hat{\theta}$ rather than maximizing item information (Chang & Ying, 1999). This simpler criterion is used because the $a$ values are similar within a level. Thus, for the

2PLM, maximizing item information is equivalent to matching $b$ with $\hat{\theta}$. For the more general 3PLM, matching $b$ with $\hat{\theta}$ when item $a$ values are the same very closely approximates maximizing item information (Chang & Ying, 1996). Thus, this simpler selection method should maintain higher efficiency. It should also result in more evenly distributed item exposure rates.

In practice, the strata consisting of items with high $a$ values tend to have high $b$ values. A shortage of lower $b$ items in those strata could cause low b items to be selected more frequently (Chang, Qian, & Ying, 2001; Parshall, Davey, & Nering, 1998). A refined STR selection method is called $a$-stratified with $b$ blocking (BSTR), which balances the distributions of $b$ values among all strata. In the BSTR method, the basic idea is to force each stratum to have a balanced distribution of $b$ values to ensure a good match of $\theta$ for different examinees.

In most of the stratified designs (e.g., Chang et al., 1999), four strata have been used. Hau, Wen, and Chang's (2002) simulation study shows that the ideal and optimum number of strata to be used in each specific application depends on the item pool structure, test length, and other testing conditions. There is a diminishing return in that dividing the pool into too many strata can lead to small stratum, in which there are not any items of close difficulty for each particular examinee. It is also shown in their study that when item difficulty is normally distributed in an item pool, the optimal strata are quite independent of the pool size and the correlation between item discrimination and difficulty ($r$=0.5).

### *2.6 Item Pool Design and Its Relationship with Other Components of CAT*

Parshall, Davey, and Nering (1998) discuss the three often-conflicting goals of item selection in CAT. First, item selection must maximize measurement precision by selecting the item maximizing information or posterior precision for the examinee's current ability level. Second, item selection must seek to protect the security of the item pool by limiting the degree to which items may be exposed. Third, item selection must ensure that examinees will receive a content balanced test. Stocking and Swanson (1998) add a fourth goal to this list, stating that item selection must also maximize item usage so that all items in a pool are used, thereby ensuring good economy of item development. Stocking and Lewis (2000) portray the item selection problem as a balloon — pushing in on one side will cause a bulge to appear on another.

An optimally designed item pool seeks the best compromise of the conflicting goals. To allow several thousand overlapping subtests to be drawn from its items, the item pool must have a sufficient number of high quality items. This is partly decided by the number of examinees the item pool serves and the distribution of the examinees. With item security consideration, the more examinees taking the test, the more items that should be in the item pool. The CAT item selection procedure picks items with a difficulty level approximately comparable to the ability estimates of the examinees, therefore it is expected that items in the pool have a difficulty distribution that is similar to the examinee ability distribution. It is desirable to have items in the pool to span a wide range of item difficulty relative to the population of interest to allow the CAT to estimate ability levels for a broad range of examinees (Urry, 1977).

Test length, which is closely tied to the stopping rules in CAT, also plays an important part in determining the number of items needed in an item pool. For a fixed length test, if the tests for individuals have no overlapped items, the number of items in a bank should be exactly the number of items in each form multiplied by the number of test takers. In reality, items can be used repeatedly within certain security constraints. Even with item overlap, it is expected that the more items a test requires, the more items needed in an item pool. Stocking (1994) recommended that the item pool should have a number of items that is at least 12 times the length of a test. Variable test length CAT usually reduces the items needed for individual examinees. In this case, the number of item needed for an item pool is correlated with the distribution of the test takers, i.e., number of examinees at each ability level.

With respect to the same item response patterns, different estimation methods may lead to slightly different ability estimates and, in turn, influence the choice of the best suitable item. Different item selection rules, such as picking the item that maximizes the information or minimizes posterior variance at the current ability estimates, may choose different items to be the most appropriate one for the examinee. Both situations cause different item usage and require different items in an optimal item pool.

Requirements on content balancing also require different compositions of the items in an item pool. For example, if the test blueprint for a 40-item math test requires 20 arithmetic reasoning items and 20 problem-solving items, the optimal item pool would contain a similar number of items for both contents. The goal is to have a sufficient number of items in each desired content area to assemble an individual test with the balanced content coverage required by the test design.

In addition, care must be taken to ensure that the item pool consists of the appropriate items to reduce the over- and under-exposure rate while meeting the test precision requirement. Item overuse causes security concerns because the more examinees take the same item, the more likely that item would be disclosed to the public. Item under-use potentially increases the item developing costs. It has been commonly realized that a tradeoff exists between test efficiency and item exposure control. A choice needs to be made that maximizes efficiency within the limits of security constraints, and that is essentially a matter of optimization. An optimally designed item pool should be able to compensate for the exposure control and cause very little decrease in the efficiency of ability estimation.

# Chapter III Reckase's Simulation Method and Extensions to 3PL

This chapter first introduces the key concepts of Reckase's simulation method for optimal item pool design. Then it discusses the simulation procedure when the method is applied to items calibrated with one-parameter logistic model (1PL) and the potential problems with the three-parameter logistic model (3PL). Finally, the extensions of the method and their applications in situations when exposure methods are built into item selection process are discussed in detail.

## *3.1 Basic Concepts of Reckase's Simulation Method*

An item pool could be described by a list of item parameters for the items in the pool. The basic idea of Reckase's method is to determine the item parameters with randomly sampled examinees from the expected examinee distribution. The simulated computerized adaptive tests are administered to the examinees assuming that each item administered to the examinee has the item parameters best suitable to the provisional ability estimation. After a certain number of examinees taking the test, the union of the "virtual" items is the optimal item pool for the CAT program.

Theoretically, every $\theta$ estimate is unique and the items optimally suitable for the estimate have unique item parameters. This simulation process described above would lead to as many items in the item pool as the total number of items administered to examinees, equal to the test length multiplied by the number of examinees. In practice, however, items function very similarly to those items whose parameters differ by a small amount. These items are redundant in the item pool in that any one of them could be used to estimate the ability level for a person with very small loss in precision.

The concept of "bin" is introduced to account for the redundancy of items with similar parameters. A bin is an item reservoir whose boundary is defined by numerical attributes of the items so that a number of items within a bin have similar attributes and are exchangeable in use. If items are calibrated with 1PL, the item difficulty parameter ($b$-parameter) is the main feature that controls the selection of test items. The bins are, therefore, defined as ranges on the IRT $\theta$–scale. For example, two consecutive bins with width of 0.2 on the $\theta$–scale are denoted as (0.0:0.2) and (0.2:0.4). Items with $b$-parameters 0.11 and 0.13 are considered exchangeable in CAT item selection because they all belong to the bin (0.0:0.2). The item pool, therefore, can be considered as a list of "bins" containing items with similar properties.

The bins that define an item pool should have a width that is sufficiently small so all items are considered equally good for estimating the ability level of an examinee. If the bin width is too large, items in the same bin may vary in their usefulness for estimating the ability level. The approach taken to determine bin width used here is to identify the range of $\theta$–scale for an item that includes the maximum of the item information function and the range around the maximum that is not much lower. "Not much lower" is arbitrarily defined as 98% of the maximum. Certainly, an argument could be made for using 96% or 97% as well.

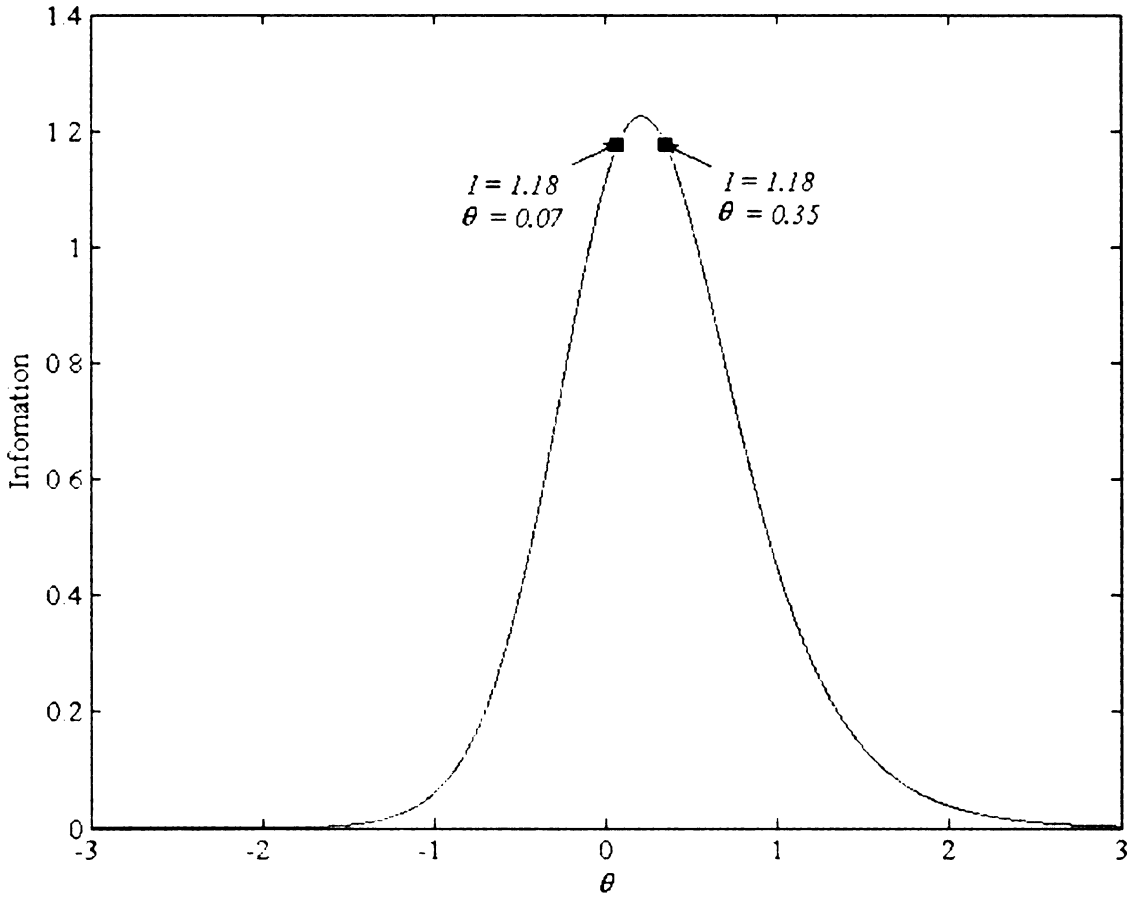


Figure 3.1 Demonstration of determining bin width

The end product of the optimal item pool design is an array of integers ($x_1$, $x_2$, , $x_B$), which tells how many items are needed in each bin to assemble all tests in a program. If no exposure control is used, the integers are bounded between zero and the test length $L$, because items in each bin can be reused and no single test requires more than $L$ items from each bin. When item exposure control is assumed, some bins may contain more items so that the shared exposure rates for items from the highly exposed bins are below the target exposure rate.

### *3.2 Reckase's Method for Optimal Item Pool Calibrated with 1PL*

When items are calibrated with 1PL, item difficulty is the only psychometric factor to decide if an item provides the most information at the $\theta$ estimate. Therefore, when designing optimal item pools that are calibrated with 1PL, Reckase's (2003) method focuses on matching the item $b$-parameters and the provisional $\theta$ estimates. Reckase's method consists of four steps:

The first step is to understand clearly the characteristics of the CAT program, because item pool design must model the test procedure as closely as possible. It is important to identify the distribution of the expected examinee population, the test length, and the type of items the test uses. For the CAT process, item exposure control, scoring algorithm, ability estimate procedure, and stopping rules should be clearly specified and strictly followed during item pool simulations.

The second step is to identify the categorical attributes required for the items, such as content area, and divide the item pool into smaller ones according to these attributes. If a test has more than one categorical attribute requirement, each separate attribute introduces a partition of the item pool. This step is to simplify the simulation procedure by focusing on determining the optimal item by the quantitative attributes such as its psychometric characteristics.

The third step of the process for determining the optimal item pool is to administer a simulated CAT to examinees randomly sampled from the expected ability distribution. If the ability follows a standard normal distribution, the initial ability level for the examinee is zero in the $\theta$ metric. The first item is the same for all examinees. It is an item with maximum information at an ability level of zero. The next optimal item will be

based on the examinee's response on previous items and the estimate of the examinee's ability level. Subsequent items are selected to have maximum information at the most recent ability estimate. If items are calibrated with 1PL, the optimal item is the one with *b*-value the same as the current theta estimates. As the test items are selected and administered, they are tallied in bins based on their *b*-values. Assuming the bin width is 0.25, the histograms in Figure 3.2 show the distribution of items across bins for two individual examinees A and B with a true ability level of -.095 and .032 taking a 15-item fixed-length test.

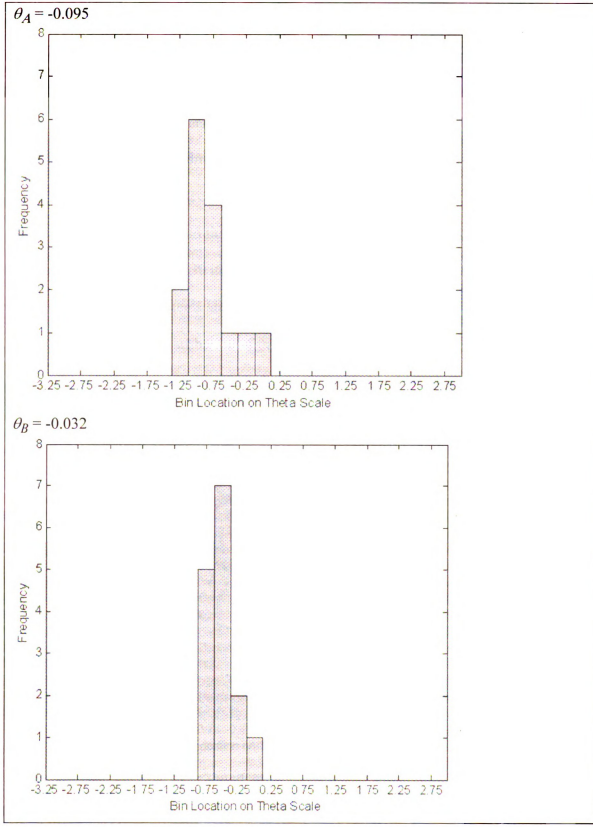


Figure 3.2 Items used for two individual examinee

As the number of test administrations increase, the number of items needed in the ideal pool will increase as well, but 15 items are not added each time because many of the needed items are already in the pool. Figure 3.2 shows that examinee A and B took some items from different bins, but also took some items in the same bins. For example, one item from bin (-0.25:0) is administered to examinee A while two items from the same bin are given to examinee B. Because the one item administered to examine A can be reused for examinee B, only one additional item in the same bin is added to the optimal item pool. Therefore, instead of 30 items being needed for an optimal item pool to support two examinees, only 23 items are needed if other constraints are not taken into account. Figure 3.3 displays the distribution of the items in the pool for two examinees.

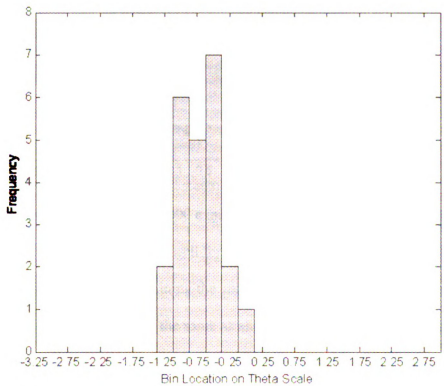


Figure 3.3 Item pool for two examinees

To determine the required size of an ideal item pool, a large number of tests are administered and the required item pool is tallied. Additional items are added to the item

pool after more examinees take the test. In the fourth step, when the expected numbers of examinees are administered the test, the union of the items forms a distribution of items that represents an optimal item pool. The sum of the number of items in the bins is the total number of items needed. Figure 3.4 shows an example of an optimal item pool for 5000 examinees taking a 15-item fixed length test.

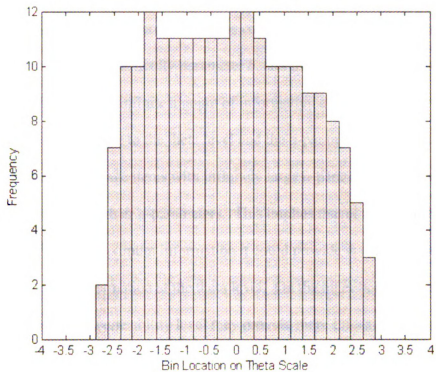


Figure 3.4 Item pool for 5000 examinees

This strategy works well for item pools calibrated with 1PL, when item difficulty is the only factor in determining the amount of information an item provides. In this case, items with *b*-parameters the same as an ability estimate will always provide maximum information at the ability estimate. Therefore, they are always the optimal items at the ability estimate compared to items with *b*-parameters different from the ability estimate. When items are calibrated with 2PL or 3PL, they may differ in the amount of information they provide even with the same *b*-parameters, simply because they have different *a*- or

*c*-parameters. Extensions to Reckase's method, therefore, are needed to account for the differences between *a*- and *c*-parameters in designing item pools calibrated with 3PL.

### *3.3 Reckase's Method Applied to 3PL*

As mentioned above, determining the optimal item pool that is calibrated with 3PL is more complicated than that with 1PL because the information an item could provide at an ability level is determined by the combination of three parameters, the discriminating parameter *a*, the difficulty parameter *b*, and the pseudo guessing parameter *c*. An item could provide an infinite amount of information at any ability level, given that the *b*-parameter is close to the theta level and the *a*-parameter is infinitely large. Although it is impossible to have items with infinitely large *a*-parameters, it is common to have items vary widely in their *a*-parameters. This implies that at a certain ability level, an item reaching the maximum information it could provide is not necessarily the item providing maximum information at the theta level. On the other hand, an item providing its highest information at one ability level may provide more information than any other items in the item pool over a range of ability levels. As demonstrated in Figure 3.5, an item with parameters $a$=1.2, $b$=0.0, and $c$=0.2 provides more information at ability level -0.28 than an item with parameters $a$ = 0.8, $b$ = -0.5, $c$ =0.2, even though the latter reaches its peak in the amount of information it can provide at this ability level.

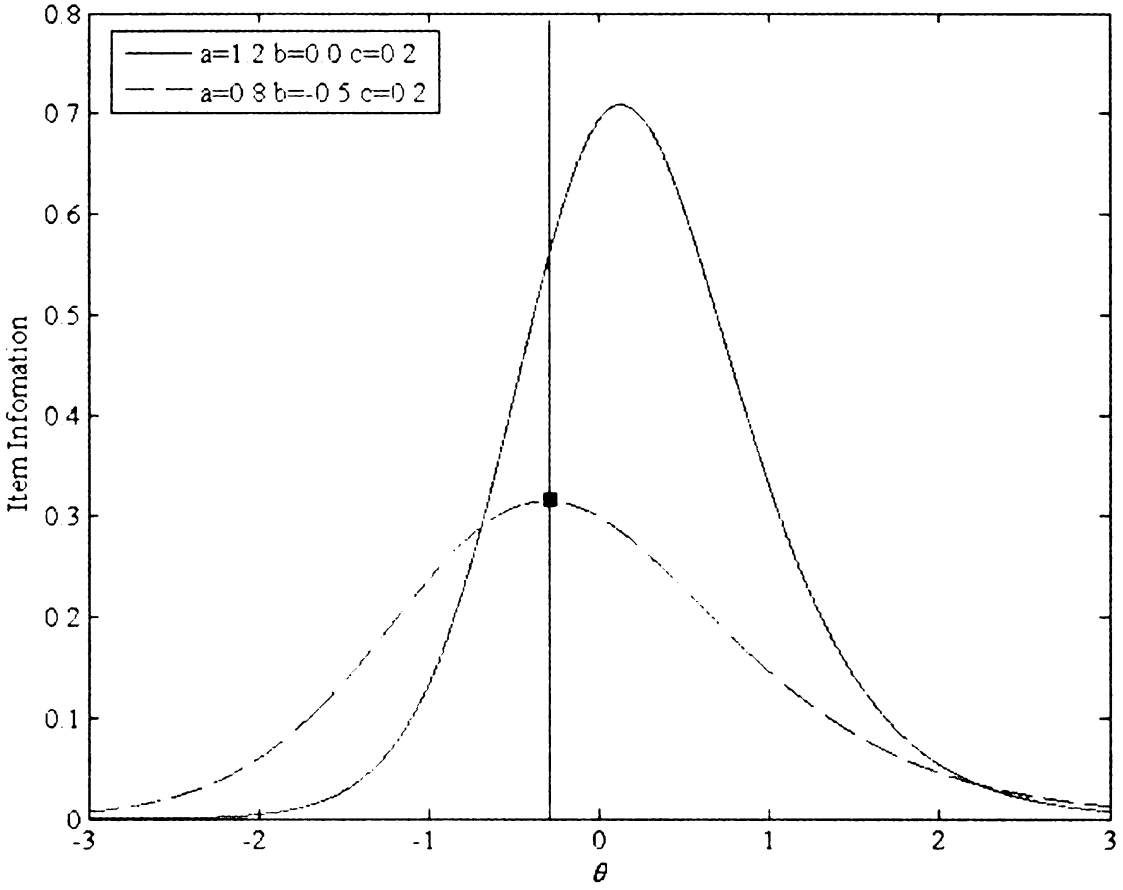


Figure 3.5 Item information provided by two different items.

Therefore, the optimal item for an ability level should not be defined as the item providing its most information at the ability level. In addition, it is unrealistic to define the optimal item pool as the one that contains items with the highest possible *a*-parameters. Instead, the optimal item pool should contain items with a range of discriminate parameters so that tests assembled from it would provide the sufficient precision the testing program requires. This study explores two strategies proposed to simulate the realistically optimal item pool. One focuses on simulating items that meet the minimum precision needed for an examinee taking the test. The other takes into consideration the relationship between the *a*-parameters and *b*-parameters in real operational items so that the simulated item parameters are within realistic boundaries.

Before introducing both strategies, it is important to extend the "bin" concept to fit the three-parameter IRT models.

### 3.2.1 Extending the "Bin" Concept

Under the framework of the three-parameter IRT model, the maximum amount of information an item could provide is decided by all three parameters. An item with high discrimination (i.e., high *a* value) generally provides more information than one with low discrimination. However, Chang and Ying (1999) demonstrated that it may provide less information at the level of theta estimate where $\hat{\theta}$ is far from the examinee's true theta. An item with smaller *c*-parameters provides more information at its maximum level, but *c*-parameters usually vary slightly across items so they have little influence on the amount of information items give. Therefore *a*- and *b*- parameters are the two primary factors to determine how much information an item is capable of providing at an ability level. Items that function similarly have similar *a*- and *b*-parameters. This leads to the extension of the "bin" concept introduced in item pool simulation with 1PL, where it is defined to be the interval of *b*-parameter values within which items provide similar amounts of information over a range of ability levels.

With 3PL, the boundary of "bin" is defined by both the *a*- and *b*- parameters. This forms a grid partitioning the plane formed by values of *a* and *b*. As illustrated graphically in Figure 3.6, each cell defined by a range of *a*- and *b*- parameters is denoted as *ab*-bin, whereas the marginal total across each row is denoted as *a*-bin and the marginal total across each column is denoted as *b*-bin. Items with parameters within the boundary of any grid defined by both *a*- and *b*-parameters provide similar information

over the entire range of ability level, and provide maximum information at the ability level around the boundary of the bin in which they are located.

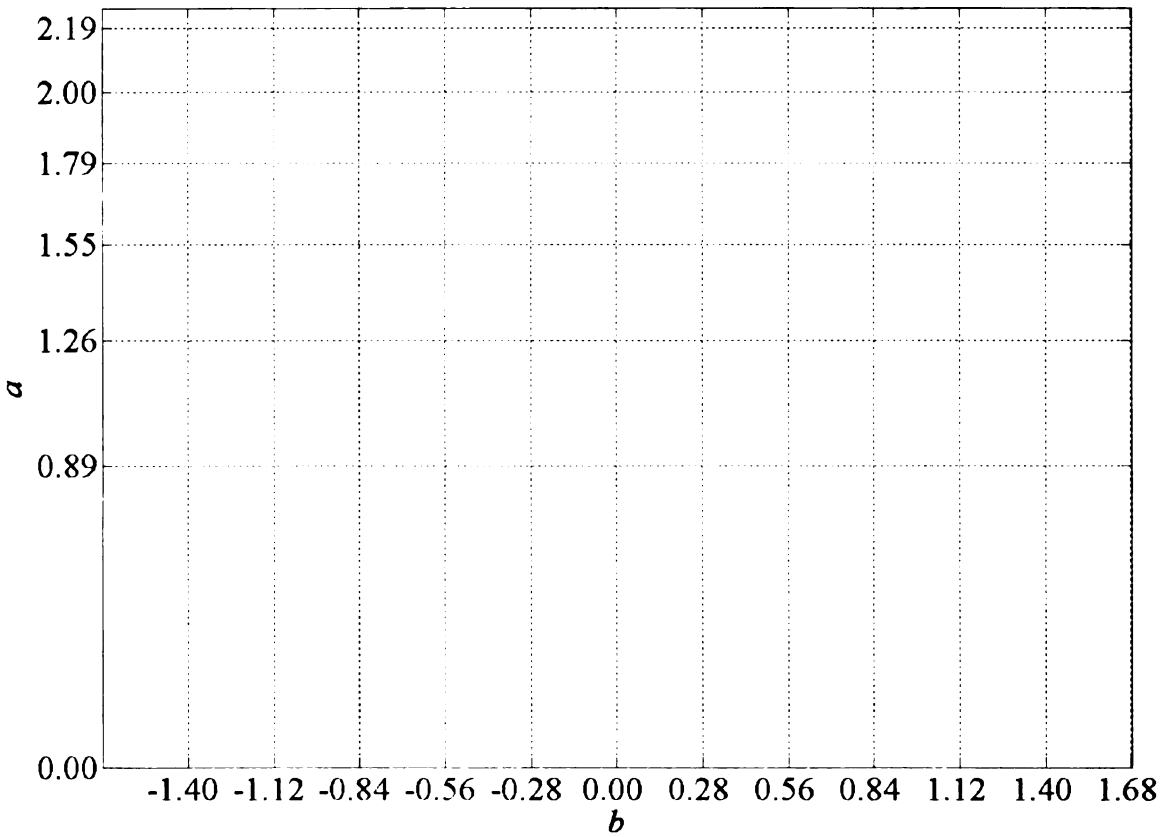


Figure 3.6 Bins defined by both *a*- and *b*- parameters.

While the boundaries of the *b*-bins are determined by dividing the $\theta$-metric (or equivilantly, metric of the *b*-parameters) into equal intervals, the width of the boundaries for the *a*-bins are set to be different, because the maximum amount of information an item can provide is proportional to the quadratic function of the *a*-parameters, assuming the *c*-parameter is constant (Lord, 1980). Equation (7) shows the relationship between the *a*-parameter and the maximum information, $M_i$, an item provides.

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1-20c_i-8c_i^2+(1+8c_i)^{3/2}] \qquad (7)$$

It can be further shown that the differences between maximum information function ( $\Delta M$ ) for items with different a parameter is

$$\Delta M = \frac{D^2[1-20c-8c^2+(1+8c)^{3/2}]}{8(1-c)^2}\Delta a^2 \qquad (8)$$

Plugging in the average $c$-parameter of the existing items, which is around 0.2, the resulted constant is 0.5, therefore

$$\Delta M = 0.5\ \Delta a^2 \qquad (9)$$

Therefore, the boundary of the $a$-bin within which the changes of the $a$-parameters cause little information change can be calculated. The grid defined by $a$-parameter intervals and $b$-parameter intervals becomes the boundary of $ab$-bins. If 0.4 is considered a small information change and 0.28 is a small b-parameter change, the bins defined by both $a$- and $b$- parameters are shown in Figure 3.7. For simplicity, an $ab$-bin is denoted by its $b$-parameter boundaries and $a$-parameter boundaries: ($b_{lower\ bound}$:$b_{upper\ bound}$, $a_{lower\ bound}$:$a_{upper\ bound}$). For example, items with $a$-parameters between 0.89 and 1.26 and $b$-parameters between 0.00 and 0.28 are in an $ab$-bin (0.00:0.28, 0.89:1.26). They are considered interchangeable in item selection.

Distinctions are made, however, to the functions of $b$-bins and $a$-bins. As mentioned above, the closeness of the $b$ parameters to ability level determines how an item would perform the best and provides the most information. On the other hand, the value of the $a$-parameter determines how much information an item can provide around the ability level where it functions the best. With the maximum information item selection approach, if an item with high information at ability level is available, it will be picked over the low information items. An optimally designed item pool, thus, should

provide sufficient items within each *b*-bin, and make sure the items with adequately high *a*-parameters are available when needed. In other words, *b*-bins tally the number of items needed that perform best over the ability levels around the *b*-bin. Within each *b*-bin, the *a*-bins record at most how many high discriminating items are needed. The item pool simulation would produce an array of integers $\vec{x} = (x_1, x_2, \quad , x_B)$, which tells how many items are needed in each *b*-bin, and a matrix $X = (X_1, X_2, \quad , X_B)$, where each element $X_B$ is a integer vector $(y_{B1}, y_{B2}, \quad , y_{BA})$ indicating at most how many items are needed in each *ab*-bin within a *b*-bin. In both cases, B is the number of *b*-bins and A is the number of *ab*-bins within each *b*-bin. The reason why they are recorded in two different matrices is that $x_B$ is usually not the same as the sum of $X_B$ in the early stage of the item pool design. After the CAT simulation, $y_B$'s from *ab*-bins with the lowest item discrimination are set to zero so that $\Sigma y_B = X_B$ and only the highest discriminating items required by the simulation are in the optimal item pool blueprint. Visual displays of the two matrices are shown in Figure 3.7, where the plot on top shows how many items in each *b*-bin are needed for the optimal pool and the plot on the bottom distinguishes different *ab*-bins with gray-scales and shows the number of items needed for each *ab*-bin within a *b*-bin.

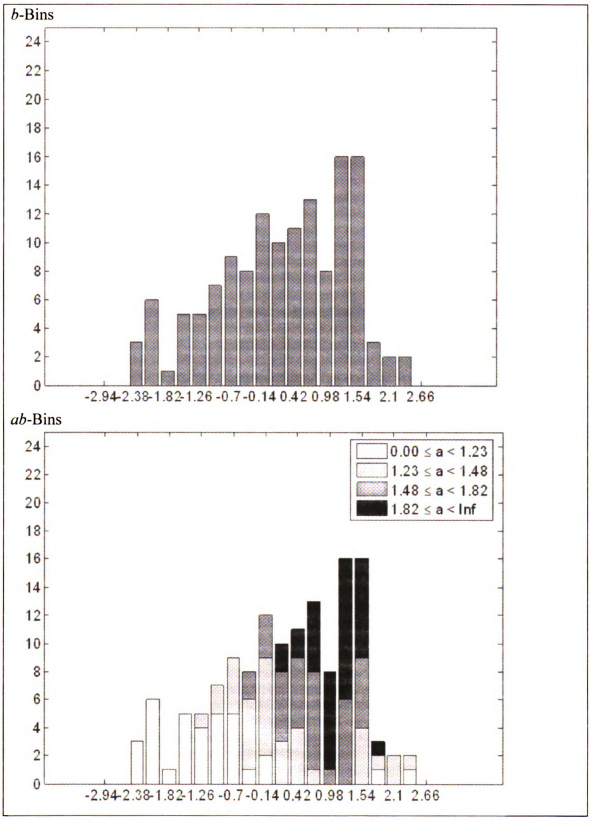


Figure 3.7 Item distribution by *b*-bins and *ab*-bins

3.2.2. Strategies to Generate Items for Item Pool Simulation with 3PL

During the item pool simulation, each item generated for the current ability estimate is assumed to provide its most information at the ability estimate. Then the *ab*-bin the simulated item belongs to is identified by its *a*- and *b*-parameters. This is similar to item generation in a 1PL situation where it is assumed the optimal item is the one with *b*-parameters close to the current ability estimate and which provides the most item information. Note that with 3PL this item is not necessarily the item that provides more information at the current ability estimate than all other items. This item simulation procedure simplifies the simulation process by not taking into account the fact that items belonging to one bin could give more information than items belonging to the other bin at the ability level close to the other bin. However, by assuming optimal items are the ones providing their most information at the ability estimates, recording the *ab*-bin items belongs to is equivalent to recording approximately how much information is needed at the ability estimate. The fact that items in one bin provide more information than items in another bin will be addressed after the item pool simulation is done with the adjustment described in section 3.2.3.

The following two strategies are proposed to generate item parameters during the item pool simulation process.

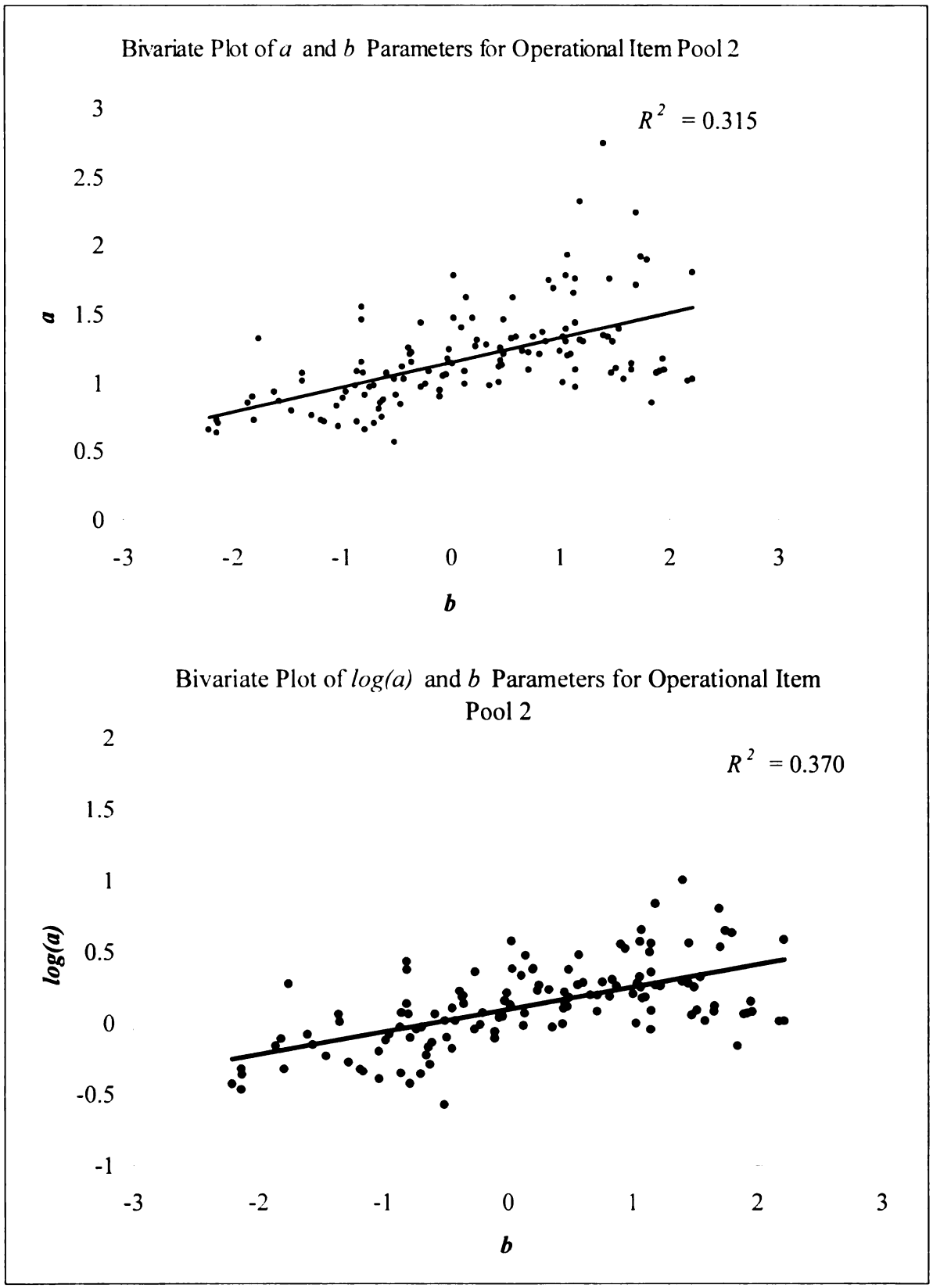


Figure 3.8 Bivariate plot of *b*-parameter and *a*-parameter for operational item pool

*3.2.2.1 Prediction Model (PM) Strategy*

The PM strategy is based on the fact that *a*-parameters and *b*-parameters are significantly correlated (Chang & van der Linden, 2003; van der Linden, Scrams, & Schnipke, 1999). As an example, Figure 3.8 shows the scatter plot of an operational item pool for ASVAB, where the *a*-parameters tend to shift with the level of *b*-parameters. In addition, the variance of the *a*-parameter increases as the *b*-parameter increases, indicating that logarithm transformations of the *a*-parameters are linearly related to *b*-parameters.

To model this relationship, the *a*-parameter for a simulated item is set equal to the regression function of the logarithm transformation of the *a*-parameter (*a*') on the *b*-parameter (Reckase, 2004).

$$a' = \log(a_i) = \beta_0 + \beta_1 b_i + \varepsilon_i \qquad (10)$$

$$a = exp(a') \qquad (11)$$

where $\varepsilon_i \sim N(0, \sigma^2)$. The variation in the *a*-parameters is included in the item pool estimation procedure by adding an error term in the regression function. Because the *c*-parameter is not significantly correlated to the *b*-parameter, it is assumed to follow a beta distribution varying around the average value.

The regression function and the variation in the *c*-parameters are estimated with the item parameters obtained from the operational items, which give realistic estimates of the item parameters for a specific testing program. During item pool simulation, items are generated in three steps:

Step 1) After each response, obtain the estimate of ability level $\hat{\theta}$ and use it as the approximation of the *b*-parameter for the next optimal item.

Step 2) Predict *a'* from the *b*-parameter with a regression function estimated from the operational items. To account for the variation in the *a*-parameter, a random number simulated from $N(0, \sigma^2)$ is added to the predicted value and then the natural exponential function of *a'* is the simulated *a*-parameter.

Step 3) Generate the *c*-parameter from the beta distribution. Re-compute the *b*-parameter so that the item provides maximum information at $\hat{\theta}$:

$$b_i = \hat{\theta}_{ij} - \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \qquad (12)$$

With the PM strategy, the *a*-parameters usually fall in a range similar to the operational items. The resulted blueprint for the optimal item pool would most likely contain items developers could easily produce. It, however, focus primarily on matching the *b*-parameters with the examinee's ability level estimates. The *a*-parameters are generated randomly, albeit within a practical range. It does not take into account the amount of information an item could provide, nor does it consider the information a simulated test can possibly provide for the examinee.

*3.2.2.2 Minimum Test Information (MTI) Strategy*

The MTI strategy posits that the item pool is optimal when computerized adaptive tests assembled from it can provide just sufficient information to examinees. The more information a test can provide, the more precise the test can estimate an examinee's ability level. However, more test information needs more highly discriminating items, which are usually expensive or hard to create, especially for easy items. The MTI strategy makes sure that tests provide sufficient precision, but do not contain overly abundant high discriminate items.

The MTI strategy sets a target information value over a range of $\theta$-scales. The target test information value is broken down for each item administered to the examinee. With $c$-parameters (which can be generated from beta distribution) and $b$-parameters (which is estimated by current ability estimate) both known, $a$-parameters can be calculated.

The MTI strategy generates items in three steps:

Step 1) Determine how much information a test needs to achieve acceptable ability estimate precision for individual examinees. Break down the target information $I$ for each item $i$ according to the following formula:

$$I_i = \frac{I_{target} - I_{adm}}{I_{test} - I_{adm}} \qquad (13)$$

To mimic the way CAT selects items, which picks the items providing the largest information at current ability estimates, the target information could be manipulated so that target information starts with a reasonably large number and decreases with the test going forward, and stays at the value of the expected target information for the last few items. While simulating the $a$-stratified exposure control method, the target information is set at a lower level and then increased to the expected value because the $a$-stratified method uses low $a$-parameter items first.

Step 2) Generate the $c$-parameter from the beta distribution. According to Lord (1980), the relationship between $\hat{\theta}$ and the parameters of the item providing its maximum information at $\hat{\theta}$ can be depicted as

$$\hat{\theta}_i = b_i + \frac{1}{Da_i} \ln\left(\frac{1+\sqrt{1+8c_i}}{2}\right), \qquad (14)$$

where $D$ is a scaling factor and is equal to 1.7. The most information a logistic item with specific parameters $a_i$ and $c_i$ can provide at $\hat{\theta}$ is

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1-20c_i-8c_i^2+(1+8c_i)^{3/2}] \qquad (15)$$

By rearranging the equation and plug in $I_i$ for $M_i$, it can be shown that

$$a_i = \sqrt{\frac{8(1-c_i)^2 I_i}{D^2[1-20c_i-8c_i^2+(1+8c_i)^{3/2}]}} \qquad (16)$$

Given that $I_i$ and $c_i$ are known, an optimal $a$-parameter can be found with equation (16) so that the item provides a minimum amount of information at the current ability estimate.

Step 3) Calculate the $b$-parameter with $a$- and $c$- parameters from equation (12) so that the generated item provides its most information at $\hat{\theta}$.

3.2.3 Post-simulation Adjustment

The results from the item pool simulation, the vector $\bar{x}$ and the matrix $X$ showing the number of items needed from each bin defined by both $a$- and $b$-parameters, essentially show how many items providing a certain amount of information are needed within the interval of $b$-parameters. This is because bins defined by both $a$- and $b$-parameter cluster items by the point where they provide the most information. As mentioned above, item simulation does not take into account the fact that items belonging to one bin could give more information than items belonging to the other bin, even at the ability level within the other bin. For example, Figure 3.9 shows that items from *ab*-bin A (-0.84:-0.56, 1.26:1.55) may provide more information at $\theta$s between -1.12 and -0.84

than items from bin B (-1.12:-0.84, 0.00:0.89), although these items in bin B may provide their most information over the same $\theta$ range. In other words, item selection procedure would choose the item in bin A over the item in bin B for ability estimates around -1.12 to -0.84.

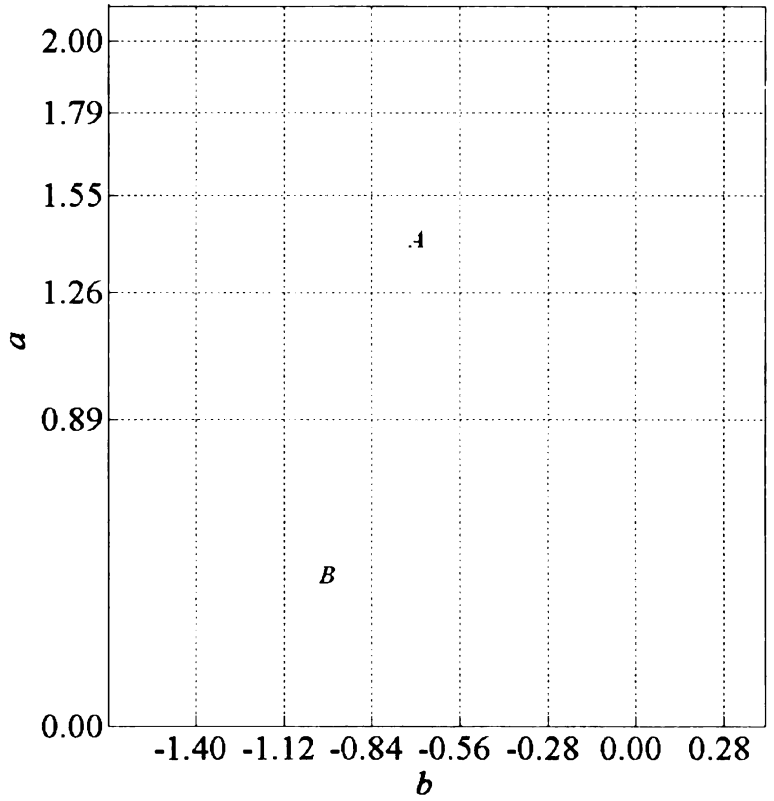


Figure 3.9 Demonstration of items in one bin offering more information than items in another bin.

Therefore the optimal item pool actually requires a sufficient number of items providing large enough information at each *b*-bin, regardless of the bin to which the items belong. An item pool constructed strictly following the number of items required by the blueprint resulted from item pool simulation will have redundant items.

These items are trimmed by using an info table that is created to select the highest information items from the bins identified by the simulation procedure. This will assist in forming the final blueprint for the item pool. To get the highest information item for

each *b*-bin, the mid-points of the *b*-bins are treated as the anchor ability level, and the mid-points of both the *a*-bin and *b*-bin as *a*-parameter and *b*-parameter, respectively, representing the bin the item comes from. For example, the ability levels needed to form the information table are (-3.90, -3.70  3.70, 3.90); an item with parameters $a = 1.08$, $b = 0.10$, $c = 0.187$ represents an item from bin (0.00:0.20, 0.89: 1.26). If three items are needed from this bin, three items with the same item parameters are entered into the info table. Sufficient items are drawn in this way to represent the number of items needed in each bin.

As shown in Figure 3.10, each column represents a *b*-bin and the number of items needed in that bin is shown in the second row. The rows below are the item IDs with each number representing an item. Items are rank ordered by the information it provides within the boundary of the *b*-bin. Items closer to the top are the items providing the most information. In practice, items ranked higher will be selected first, regardless of the bins they are from. Therefore, even though each *b*-bin still requires a certain number of items, they can be from the other bins. The graphical way to select needed items is to highlight the exact number of items needed for each *b*-bin from the item providing the most information. The unique items for all highlighted items are the items needed for the optimal item pool.

| -2.00 | -1.75 | -1.50 | -1.25 | -1.00 | -0.75 | -0.50 | -0.25 | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 12 | 11 | 11 | 11 | 11 | 11 | 11 | 12 | 12 | 11 | 10 | 10 | 10 | 9 | 9 | 8 |
| 35 | 35 | 35 | 61 | 67 | 67 | 90 | 90 | 109 | 112 | 139 | 139 | 152 | 171 | 174 | 182 | 182 |
| 40 | 40 | 45 | 45 | 56 | 82 | 85 | 109 | 112 | 127 | 137 | 152 | 158 | 172 | 171 | 174 | 193 |
| 19 | 30 | 40 | 56 | 61 | 85 | 67 | 97 | 111 | 111 | 127 | 137 | 157 | 157 | 172 | 187 | 187 |
| 24 | 45 | 30 | 73 | 82 | 90 | 82 | 104 | 115 | 109 | 112 | 151 | 151 | 153 | 175 | 189 | 194 |
| 28 | 24 | 61 | 67 | 73 | 83 | 97 | 85 | 116 | 123 | 111 | 143 | 171 | 152 | 180 | 184 | 184 |
| 14 | 31 | 56 | 35 | 83 | 84 | 83 | 98 | 127 | 137 | 123 | 150 | 156 | 166 | 173 | 188 | 195 |
| 31 | 28 | 31 | 57 | 45 | 79 | 79 | 99 | 104 | 110 | 130 | 141 | 153 | 158 | 189 | 183 | 188 |
| 12 | 41 | 52 | 30 | 57 | 56 | 96 | 115 | 110 | 118 | 141 | 158 | 143 | 165 | 165 | 180 | 189 |
| 30 | 47 | 47 | 40 | 64 | 73 | 84 | 116 | 118 | 115 | 126 | 156 | 139 | 156 | 176 | 173 | 191 |
| 3 | 19 | 57 | 66 | 66 | 81 | 99 | 96 | 98 | 130 | 131 | 142 | 150 | 154 | 183 | 176 | 202 |
| 4 | 50 | 41 | 52 | 84 | 64 | 104 | 101 | 97 | 126 | 121 | 153 | 154 | 155 | 163 | 175 | 197 |
| 13 | 61 | 50 | 64 | 71 | 96 | 101 | 79 | 90 | 139 | 142 | 130 | 137 | 159 | 164 | 193 | 204 |
| 29 | 14 | 46 | 53 | 81 | 71 | 95 | 84 | 99 | 116 | 151 | 144 | 141 | 164 | 157 | 172 | 183 |
| 41 | 52 | 55 | 60 | 74 | 61 | 98 | 93 | 123 | 131 | 150 | 121 | 166 | 175 | 166 | 171 | 196 |
| 47 | 46 | 24 | 55 | 79 | 66 | 93 | 110 | 101 | 104 | 143 | 157 | 144 | 163 | 153 | 190 | 174 |

Figure 3.10 Items in the order of information provided most in each *b*-bin.

One caveat of this procedure is that because items provide more information over a range of ability levels, they may be administered more times than the test administrators want. Within a *b*-bin, the expected number of times an item is administered depends on the rank order of the information it provides. During the item pool simulation it can be estimated by recording the number of times items from each bin is simulated and administered. If an item is to be selected more than the target exposure rate, a new item from the same *ab*-bin is added to the final item pool. For example, Figure 3.11 shows the expected item usage within each *b*-bin for 8000 examinees ordering by the information it provides. Item 109 in Figure 3.8 is expected to be selected 11,800 times (8000 + 2471 + 1329), which is 0.48 times more than an item can be selected. The optimal item pool will need one more item in the same *ab*-bin as item 38. If the target exposure rate is 0.33, then it is 3.43 times more than the target exposure times. Four more items from the same *ab*-bin, therefore, need to be added.

| -2 | -1.75 | -1.5 | -1.25 | -1 | -0.75 | -0.5 | -0.25 | 0 | 0.25 | 0.5 | 0.75 | 1 | 1.25 | 1.5 | 1.75 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 261 | 599 | 1043 | 1858 | 2386 | 3916 | 3590 | 3832 | 8000 | 5167 | 3041 | 3078 | 2120 | 1376 | 899 | 545 | 309 |
| 203 | 460 | 809 | 1293 | 1642 | 2366 | 2358 | 2471 | 3362 | 2644 | 1798 | 1542 | 1131 | 777 | 519 | 355 | 214 |
| 162 | 349 | 604 | 961 | 1225 | 1729 | 1754 | 1864 | 2244 | 1852 | 1289 | 1116 | 819 | 578 | 376 | 247 | 144 |
| 117 | 272 | 463 | 734 | 918 | 1271 | 1314 | 1378 | 1648 | 1329 | 946 | 790 | 580 | 384 | 255 | 164 | 97 |
| 79 | 201 | 351 | 550 | 703 | 939 | 986 | 1011 | 1238 | 927 | 676 | 576 | 427 | 268 | 182 | 97 | 58 |
| 54 | 135 | 267 | 380 | 484 | 683 | 661 | 699 | 877 | 645 | 444 | 390 | 273 | 161 | 105 | 57 | 25 |
| 27 | 86 | 164 | 253 | 318 | 443 | 414 | 433 | 597 | 409 | 263 | 232 | 163 | 84 | 40 | 21 | 11 |
| 9 | 46 | 96 | 149 | 180 | 267 | 217 | 230 | 383 | 233 | 139 | 118 | 67 | 28 | 8 | 4 | 3 |
| 4 | 22 | 54 | 72 | 71 | 132 | 105 | 92 | 198 | 110 | 54 | 46 | 18 | 7 | 1 | 1 | 0 |
| 1 | 5 | 23 | 31 | 17 | 46 | 25 | 24 | 76 | 30 | 8 | 9 | 3 | 1 | 0 | 0 | 0 |
| 0 | 2 | 4 | 6 | 2 | 13 | 5 | 4 | 13 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 3.11 Item usage in the order of information provided most in each *b*-bin.

Figure 3.12 and 3.13 demonstrate an item pool blue print before and after post-simulation adjustment.

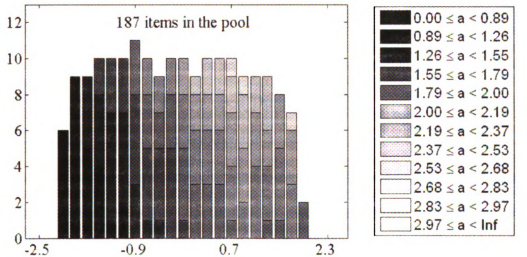


Figure 3.12 Item distribution for optimal item pool before adjustment

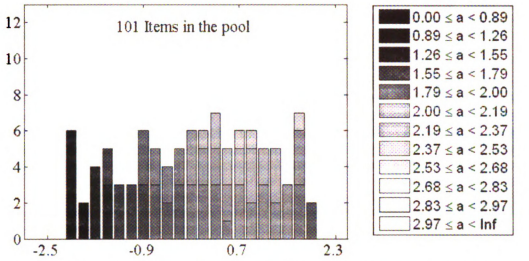


Figure 3.13 Item distribution for the optimal item pool after post-simulation adjustment

*3.4 Design Adjustments to Different Exposure Control Methods*

3.4.1 Item Pool Design without Exposure Control

The item exposure rate for an item is the number of times an item is administered divided by the number of examinees. During the optimal item pool design process, the real items are not available, but the simulated items representing items from the bins they belong to are. The number of times items from each bin are administered is the estimate of the marginal exposure count of the items in this bin, which is shared by the number of items needed in this bin. If the adaptive test has no item-exposure control, the design blueprint for the item pool after post-simulation adjustment follows directly from these exposure rates.

3.4.2 Item Pool Design with Sympson-Hetter Exposure Control

The Sympson-Hetter method controls item exposure with a probabilistic index $k$ to adjust the number of times an individual item is administered. An item that potentially

has a high exposure rate is assigned a small $k$ value so that the probability of administering the item is brought down below the maximum exposure rate. When a selected item is not administered because of the exposure control, the next most informative item will be selected. The goal of the optimal item pool design with Sympson-Hetter exposure control is that in addition to being optimal to accommodate test length, content balancing, and other aspects of the test, it makes sure the exposure control only slightly reduces the test precision. The goal is achieved by making sure that there are sufficient items in the bins where items are selected more often.

The same method used in the post-simulation adjustment introduced previously is used to retain sufficient items in each bin. Because the number of times an item is used can be recorded during item pool design process, if it reaches $rN$, another item from the same bin is retained so that the share of the total exposure for each of the items within the bin is not larger than $r$.

### 3.1.3 Item Pool Design with *a*-Stratified Exposure Control

The $a$-stratified exposure control requires that the item pool be partitioned into an equal number of items in each stratum. However, before the item pool design simulation, the number of items in a pool has is unknown. It is impossible to decide the number of items needed in each stratum, and the boundary (defined by $a$-parameters because it is stratified by $a$) of each stratum. However, by modeling the way the $a$-stratified method selects items, it is possible to approximate the psychometric properties of the items needed in each stage of the test administration. The $a$-stratified method splits a fixed length test into a number of stages that is equal to the number of strata. Each stage has an equal number of items, and they are all selected from the same stratum of the item pool.

The important rule is that in the first stage, items are selected from the strata with the lowest *a*-parameters and, therefore, they provide the least information at the current ability estimates. The next stage items are selected from strata with the next lowest *a*-parameters, and in the last stage, items are selected from strata with the largest *a*-parameters. Based on the characteristics of the *a*-stratified method, each test administration is divided into stages with an equal number of items in each stage. Within each stage, adjustments to the item simulation strategy are made to approximate the *a*-stratified exposure control.

For PM item simulation, a random term of the prediction model is partitioned into four equally likely intervals ordered by its possible values. During each stage of the test, items are simulated from the corresponding interval. For example, with four strata and a 16-item test, the first four items are simulated so that their *a*-parameters are from the lowest quarter of possible range, i.e., after $a$ is predicted from $b$, random error $\varepsilon$ is drawn from $N(\Phi^{-1}(0.125),(\sigma/4)^2)$, the next four item are drawn from $N(\Phi^{-1}(0.375),(\sigma/4)^2)$, the next four items are $N(\Phi^{-1}(0.625),(\sigma/4)^2)$, and the last four items are from $N(\Phi^{-1}(0.875),(\sigma/4)^2)$.

For MTI simulation, the target information for one test is set to be different for each stratum of items. For example, assume the target information for the whole test is 10.0. During the simulation, the target test information is set to 7.0 for the first four items. Because the target information is for the whole test, the average target information for each item is 7.0/16 = 0.438 at the current $\theta$ estimate. Then, the target test information increases to 8.0 for the next four items, 9.0 for the next four items, and 10.0 for the last

four items. Test information of 10.0 for the last four items makes sure the simulated test achieve the minimum information needed.

# Chapter IV Methods

This study is composed of two closely related parts. In the first part, a Monte Carlo simulation is used to design the optimal item pools for two CAT-ASVAB sections. The second part evaluates the optimal item pools with empirical criteria and compares them to the operational item pools on performances. This chapter briefly introduces the characteristics of the operational item pool and the psychometric procedures CAT-ASVAB used in test administration. It then describes the simulation procedure and the independent variables on which this study focuses. Finally, a definition and brief description of the item pool evaluation criteria are provided.

## *4.1 Operational Item Pools*

Operational item pools for two sections of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), Arithmetic Reasoning and General Science, were used as the target and benchmark of this study. Originally designed to predict future academic and occupational success in military occupations, ASVAB is a nationally normed, multi-aptitude test battery that assesses academic ability and predicts success in a wide variety of occupations. It tests a student's knowledge in eight areas including general science, mathematics, word knowledge, paragraph comprehension, electronics information, auto and shop information, and mechanical comprehension. Its Computerized Adaptive Testing version, CAT-ASVAB, began in 1997 and was the first large-scale adaptive battery to be administered in a high-stakes setting, influencing the qualification status of applicants for the U.S. Armed Forces. It is also one of the most thoroughly researched tests of human proficiencies in modern history.

The Arithmetic Reasoning (AR) section of the CAT-ASVAB is a 15-item test measuring the ability to solve basic arithmetic word problems. The General Science (GS) section has 15 items and measures knowledge of life science, earth and space science, and physical science. The General Science test is content-balanced among three content areas (Life Science, Physical Science, and Chemistry). Items are administrated in roughly the same proportion and order for each content area, as found in the reference P&P form:

*L, P, L, P, L, P, L, P, L, P, L, P, L, C,*

Where L = Life Science, P = Physical Science, C = Chemistry.

CAT-ASVAB test items are selected using maximum information. To save computation time, an information "look-up" table is used. The Sympson-Hetter method is incorporated to reduce overexposure of certain highly informative items. Owen's approximation to the posterior mean (Owen, 1975) was used to update the ability estimate during test administration. For each test, the prior distribution had a mean of 0.0 and a standard deviation of 1.0. The estimate after the last item was used as the score for the test. Each test was terminated after a fixed number of items.

### *4.2 Simulation Procedure*

Programs were developed with MATLAB® Student Version R14 (2004) to simulate item pool design and evaluate both simulated and operational item pools. Item pool design simulation was conducted in the following steps:

Step 1: Modeling CAT Procedures

Because the purpose of this study is to investigate the optimal item pool design for two sections of the CAT-ASVAB testing program, the simulation procedure follows closely the psychometric procedure used in CAT-ASVAB operations.

Test length was the same as the operational ASVAB sections, both of which are 15-item fixed length sections. Content balance was not considered in AR item pools. For GS item pools, the item pool was divided into three small pools, each containing items for one content. Items were administered in the same order as described in Chapter 4.1, where item 1 through 15 are identified as L, P, L, P, L, P, L, P, L, P, L, P, L, and C.

Maximum information and the info table were used to select items. Owen's approximation to the posterior mean (Owen, 1975) was used to update the ability estimate during test administration. For each test, the prior distribution of $\theta$ had a mean of 0.0 and a standard deviation of 1.0.

Step 2: Generating Examinee Population

CAT-ASVAB item pools are designed to serve examinees whose ability distribution is assumed normal with a mean of 0.0 and variance of 1.0. The item pool simulation followed the same assumption, and examinees were randomly sampled from $N(0,1)$.

Step 3: Generating Item Parameters

For each test, the first item was generated to be optimal for an ability level of 0 in the $\theta$-metric. After each response, optimal items were generated for the current $\theta$ estimate. It is assumed that items are calibrated by the three-parameter logistic model.

Therefore, *a*-, *b*-, and *c*- parameters were generated by one of the two methods (PM and MTI) described in Chapter 3.2. With either method, the *c*-parameter was generated from a beta distribution with mean and variance equal to the mean and variance of the operational items. The *a*-parameters were generated depending on the current ability estimate and method used (PM or MTI). The *b*-parameters were generated so that the item provide its most information at $\tilde{\theta}$.

Step 4: Generating Response Data

The examinee response was generated following each item generation according to a three-parameter logistic model, where the probability of examinee *i* correctly answering item *j* is expressed as:

$$P_i(\theta_j) \equiv c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta_j - b_i)}} \qquad (17)$$

$P_i(\theta_j)$ is the probability that a person $j = 1 \quad J$ with an ability parameter $\theta_j$ gives a correct response to an item $i = 1 \quad I$; $a_i$ is the value for the discrimination parameter, $b_i$ for the difficulty parameter, and $c_i$ for the guessing parameter of item *I*.

Because the examinee's true $\theta$ was known in the simulation, $P_{ij}$ was computed after each item administered to the examinee was simulated. Then a random number $m_{ij}$ was drawn from a uniform distribution $U(0,1)$ and compared to $P_{ij}$. If $m_{ij}$ was equal to or less than $P_{ij}$ then it was assigned a 1 as the response, otherwise 0 was assigned.

Step 4: Post-Simulation Adjustment

Five replications were conducted for each combination of methods and control variables so that a relatively stable approximate of the optimal item pool could be obtained. The blueprints and the item exposure counts from the five replications were averaged before a post-simulation adjustment was done.

### *4.3 Control Variables*

Two independent variables were controlled for in all item pool designs: design method and exposure control method. Both AR and GS have the same target exposure control rates (1/3 for Sympson-Hetter method), which is the same as the operational procedure does. For simplicity, four strata were assumed for the *a*-stratified method.

Each of the two item pools, AR and GS, had its unique control variables. Specifically, (1) there was no content balancing for AR while three contents were administered in a fixed order for GS; (2) each item pool defined the bin width differently in the $\theta$-metric, depending on the characteristics of the operational item pools.

The simulation design is illustrated in Table 4.1.

Table 4.1 Simulation Design

| | | |
|---|---|---|
| CAT Item Pool Design | Item Pools | Arithmetic Reasoning (AR) |
| | | General Science (GS) |
| | Test length | 15 |
| | Examinee distribution | $N(0,1)$ |
| | Exposure control | No exposure control |
| | | Sympson-Hetter (target exposure rate is 1/3) |
| | | $a$-stratified (4 strata) |
| | Design Method | Prediction Model |
| | | Minimum Test Information |
| | Bin width | $b$-bin: 0.20 for AR and 0.26 for GS<br>$a$-bin: $\Delta a^2 = 2\Delta I_{Maximum} = 0.8$ |
| | Content Balancing | Single content for AR<br>Three contents for GS |

*4.4 Evaluating Simulated and Operational Item Pools*

Two types of distribution were considered in the item pool evaluation: (a) 6,000 $\theta$'s were simulated from $N(0,1)$, and these values were treated as the true abilities for the examinees, and (b) 65 fixed values ranging from -4 to 4 with an interval of 0.125 were selected (i.e., $\theta = -4.0, -3.875, \ldots, 3.875, 4.0$). Five hundred examinees were set to have an identical latent ability at each $\theta$ level. The former is to evaluate general performances, and the latter is to compute statistics conditional on $\theta$.

The item pool evaluation criteria used by Chang & Ying (1999) and Reckase (2005) were adopted for this study. Precision of proficiency estimation include average test

information at each theta level, bias, Mean Square Error (MSE), and correlation coefficients between estimated and true person parameters. Test security indicators include skewness of item exposure rate distribution, percentage of overexposed items, item overlap rate, and percentage of underexposed items.

***Conditional Test Information***

Test information is the sum of all the Fisher item information in the test. In a fixed-length CAT test, it can be taken as the index of test efficiency. The larger the amount of information a test provides, the more efficient the test is.

***Conditional Standard Error of Measurement (CSEM)***

At each fixed $\theta$ point, the standard error of measurement (SEM) was calculated by the formula:

$$SEM(\theta_i) = \sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(\hat{\theta}_{ij} - \bar{\theta}_i)} \qquad (18)$$

Where $N_i = 500$ is the number of replications (i.e., the number of adaptive tests administered) at each fixed $\theta$ point, and $\bar{\theta}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\hat{\theta}_i$ is the mean of the ability estimates over the $N_i$ replications at $\theta_i$.

***Bias and Mean Square Error (MSE)***

These quantities are defined as follows:

$$Bias = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j) \qquad (19)$$

and

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2 \ , \qquad (20)$$

where $N$ is the number of simulees, and $\hat{\theta}_j$ is the estimator of the $j$th simulee with ability level $\theta_j$.

***Conditional Bias and Conditional Mean Square Error (CMSE)***

These quantities are defined as

$$Conditional\ Bias = \frac{1}{N_i}\sum_{j=1}^{N_i}(\hat{\theta}_{j_i} - \theta_i) \ , \qquad (21)$$

and

$$CMSE = \frac{1}{N_i}\sum_{j=1}^{N_i}(\hat{\theta}_{j_i} - \theta_i)^2 \ , \qquad (22)$$

where $\theta_i$ = -4.0, -3.875, , 3.875, 4.0, for $i$ = 1, 2, , 65, respectively, and $\hat{\theta}_{j_i}$ ($j$ = 1, 2, , 500) is the corresponding estimator of $\theta_i$. These values are estimated as the conditional averages of errors and squared errors in the final estimates of $\theta_i$ in simulations. As additional overall measures of the quality of the final estimates of $\theta$, the estimates of the bias and MSE functions in (19) and (20) were averaged over all simulated values of $\theta$ in the study. They give a picture of item pool performance for individual ability level.

***Skewness of Item Exposure Rate Distribution***

A $\chi^2$ statistic proposed by Chang and Ying (1999) is used to measure skewness of item exposure rate distribution. It is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(r_i - L/n)^2}{L/n}, \qquad (23)$$

where

$r_i$ is the observed exposure rate for the $i^{th}$ item,

$L$ is the test length, and

$n$ is the number of items in the item pool.

Equation (23) captures the discrepancy between the observed and the ideal item exposure rates, and it quantifies the efficiency of item pool usage. A low $\chi^2$ value implies that most of the items are fully used.

The ratio of $\chi^2$ measures follows an $F$ distribution and can be used to compare the exposure rates of two methods.

$$F_{\text{method1, method2}} = \chi^2_{\text{method1}} / \chi^2_{\text{method2}} \qquad (24)$$

If $F < 1$, then method 1 is regarded as superior to method 2 in terms of the overall balance of item exposure rates.

***Percentage of Overexposed Items***

The exposure rate of an item can be defined as the ratio of the observed number of item administrations to the total number of the examinees. A moderate level of item exposure rate is generally desired. A high exposure rate of an item means an increased risk of the items being known by the prospective examinees. If so, both the test security and validity are threatened by the high item exposure rate. Therefore, the percentage of overexposed items is taken as an important criterion to evaluate the success of a CAT program. The expected exposure rate was set equal to 1/3 for both tests in ASVAB (Segall, Moreno, & Hetter, 1997).

***Percentage of Underexposed Items***

Low item exposure rate means that the item is rarely used. An item pool with too many items with too low an exposure rate is a sign of the underutilization of the pool. Both the cost-effectiveness of developing the items and the appropriateness of the item selection method are challenged by the low item exposure rate. In this study, an item with an exposure rate lower than .02 is considered as underexposed.

***Test Overlap Rate and Conditional Test Overlap Rate***

Test overlap rate is the expected number of common items encountered by two randomly selected examinees divided by the expected test length. Ideally, the number of common items between any two randomly sampled examinees should be minimized.

Test overlap rate can be calculated by (1) counting the number of common items for each of the $N(N-1)/2$ pairs of examinees, (2) summing all the $N(N-1)/2$ counts, and (3) dividing the total counts by $LN(N-1)/2$ (Chang & Ying, 1999). The following equation summarizes the calculation (Chen, Ankenmann, & Spray, 1999):

$$\bar{T} = \frac{\sum_{i=1}^{n}\binom{m_i}{2}}{L\binom{N}{2}} = \frac{\sum_{i=1}^{n} m_i(m_i-1)}{LN(N-1)} \qquad (25)$$

where $N$ denotes the number of fixed-length CATs administered, $L$ is the number of items in each of the CATs, $n$ is the number of items in the pool, and $m_i$ is the number of times item $i$ was administered across all $N$ CATs.

Conditional test overlap rate computes the test overlap rates for 500 tests administered to each of the sixty-five fixed $\theta$s. The same procedure Chang and Ying (1999) described in equation (25) can be used to compute test overlap rate for the tests

administered to each fixed $\theta$. In this case, $N$ is 500 and $m_i$ is the number of times item $i$ was administered across all 500 CATs. The conditional test overlap rate gives a more accurate picture of test overlap at a particular ability level, instead of the average across all ability levels.

# Chapter V The Performance of the Item Pools without Exposure Control

## *5.1 Item Pools for Test without Content Balance*

Figure 5.1 compares the distributions of the operational item pool and two optimal item pools designed by MTI and PM, assuming no exposure control. Table 5.1 presents the sizes and the summary statistics of the item parameters for the three pools. The optimal item pools consist of the fewest items. This is not surprising, partly because both assume no exposure control, while the operational pool is designed for tests with the Sympson-Hetter exposure control.

Table 5.1 indicates that all item pools have items that span a wide range of difficulty levels, roughly from -2.5 to 2.5. However, the items in the optimal item pools have slight smaller ranges. The operational pool has a large number of items with *b*-parameters between 0.0 and 1.5, while the optimal pools display a more even distribution across *b*-bins. The MTI pool consists of the fewest items and their *a*-parameters are more concentrated, ranging from 1.275 to 1.781. The PM pool shows the characteristics of the item parameters similar to the operational pool, in which difficult items tend to have high *a*-parameter and easy items tend to have moderate to low *a*-parameters.

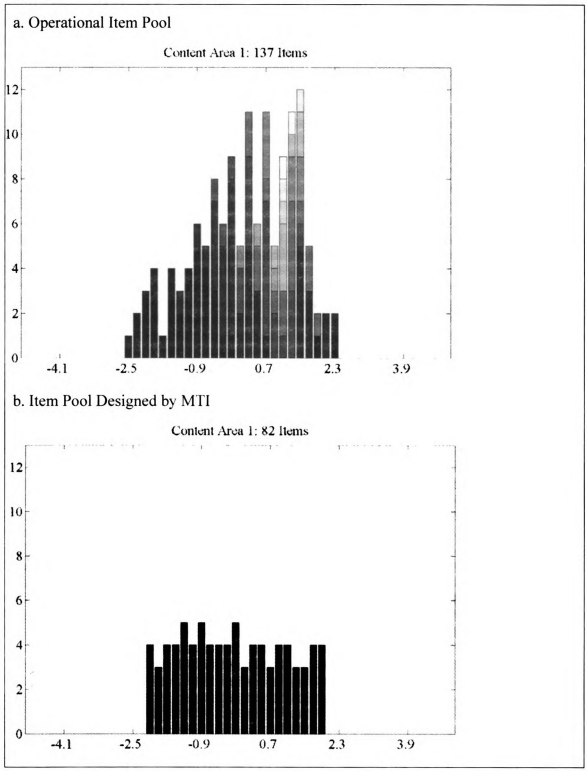


Figure 5.1 Item distribution for item pools without exposure control

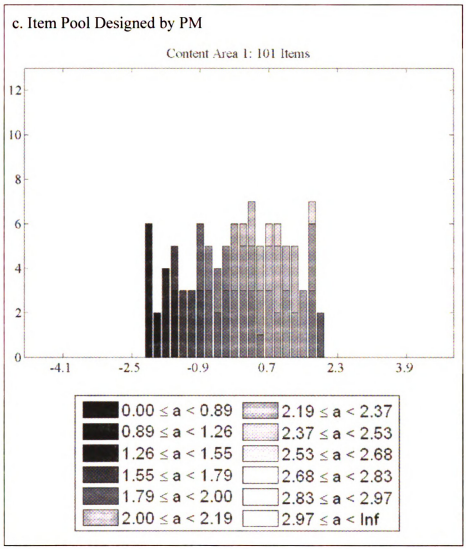


Figure 5.1 (Cont'd) Item distribution for item pools without exposure control

The overview of the evaluation results for these item pools are presented in Table 5.2. The ability estimates from all pools exhibit a certain level of positive bias; however, the magnitudes of the bias are negligible. MSEs from optimal item pools are smaller than that from operational pool. The MTI pool and PM pool resulted in a higher correlation coefficient than the operational pool.

Table 5.1 Item Pool Size and Item Parameter Statistics for Arithmetic Reasoning without Exposure Control

| Pool | Pool Size | *a* | | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. |
| OP | 137 | 1.556 | 0.487 | 3.141 | 0.746 | 0.115 | 1.170 | 2.343 | -2.625 | 0.186 | 0.063 | 0.328 | 0.038 |
| MTI | 82 | 1.601 | 0.105 | 1.781 | 1.275 | -0.159 | 1.194 | 1.942 | -2.186 | 0.218 | 0.075 | 0.423 | 0.069 |
| PM | 101 | 2.000 | 0.427 | 2.638 | 0.932 | -0.031 | 1.143 | 1.943 | -2.143 | 0.177 | 0.059 | 0.398 | 0.063 |

Table 5.2 Summary Statistics of the Performance of the Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0025 | 0.0022 | 0.0114 |
| MSE | 0.0857 | 0.0739 | 0.0576 |
| Correlation | 0.9563 | 0.9636 | 0.9703 |
| Skewness of item exposure rate | 31.3822 | 12.0199 | 15.0003 |
| Item overlap rate | 0.3385 | 0.3294 | 0.2969 |
| Pct of items with item exposure Rate> 1/3 | 8.76% | 14.63% | 8.91% |
| Pct of items with item exposure Rate<.02 | 44.53% | 7.32% | 13.86% |
| Pool Size | 137 | 82 | 101 |

Table 5.2 also shows that optimal item pools have a smaller test-retest overlap rate despite having fewer items. It indicates that the magnitude of the item overlap rate may not be related to the pool size with the optimal combinations of the items in the pool. The plots for the conditional test-retest overlap rate in Figure 5.2 reveal that optimal item pools have higher overlap rates for $\theta$ levels below approximately -2.00 and above 2.00. However, in practice, there are very few examinees at these ability levels.

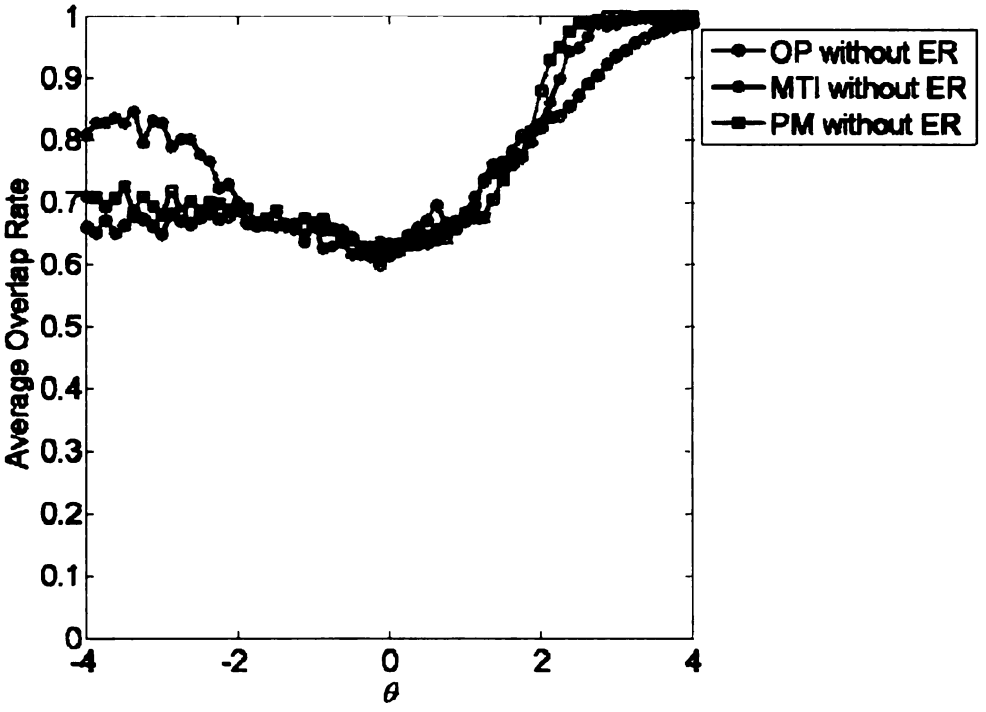


Figure 5.2 Test-retest overlap rate conditional on $\theta$

Both optimal item pools have significantly smaller percentages of under-exposed items. Although the MTI pool has a higher percentage of over-exposed items, it is reasonable given that it is the smallest pool and no exposure control was imposed. Increasing the pool size reduced the item overlap rate.

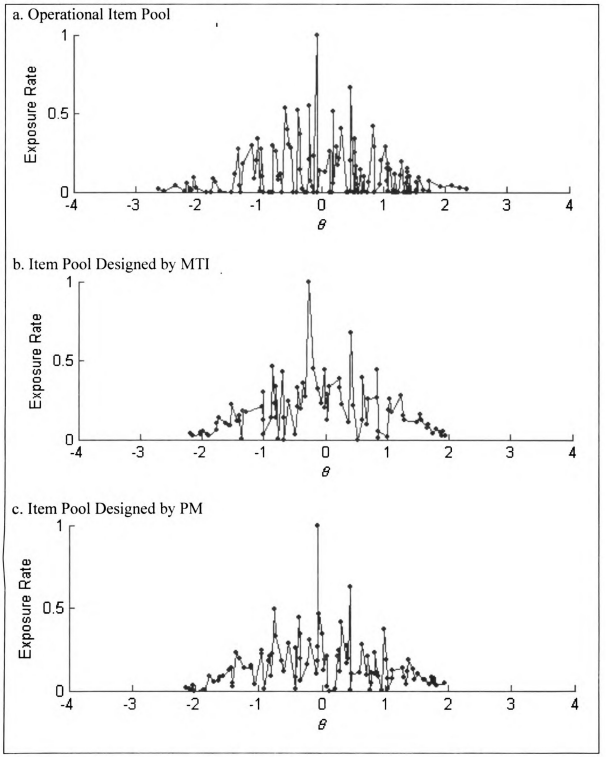


Figure 5.3 Item exposure rate by difficulty level

Figure 5.3 plots the item exposure rate for individual items in the order of their difficulty levels. Extremely easy and extremely difficult items tend to have smaller

exposure rates, but under-exposed items are across all difficulty levels, especially those in the operational item pool. Table 5.2 indicates that the MTI pool has the fewest under-exposed items and Figure 5.3a shows that items with extreme difficulty levels are utilized more often in MTI pools.

As shown in Figure 5.4, the three item pools resulted in quite different average test information plots at various fixed $\theta$ levels. The plot for the PM item pool looks similar to the one for the operational pool, but provides more information over most ability levels. The MTI item pool provides significantly less information over the ability level between approximately -1.5 and 2.0, but the amount of information it provides over a long range of ability levels exceeds the target information, which is 10.0 between ability levels ±2.0, and 8.0 beyond ability levels ±2.0.

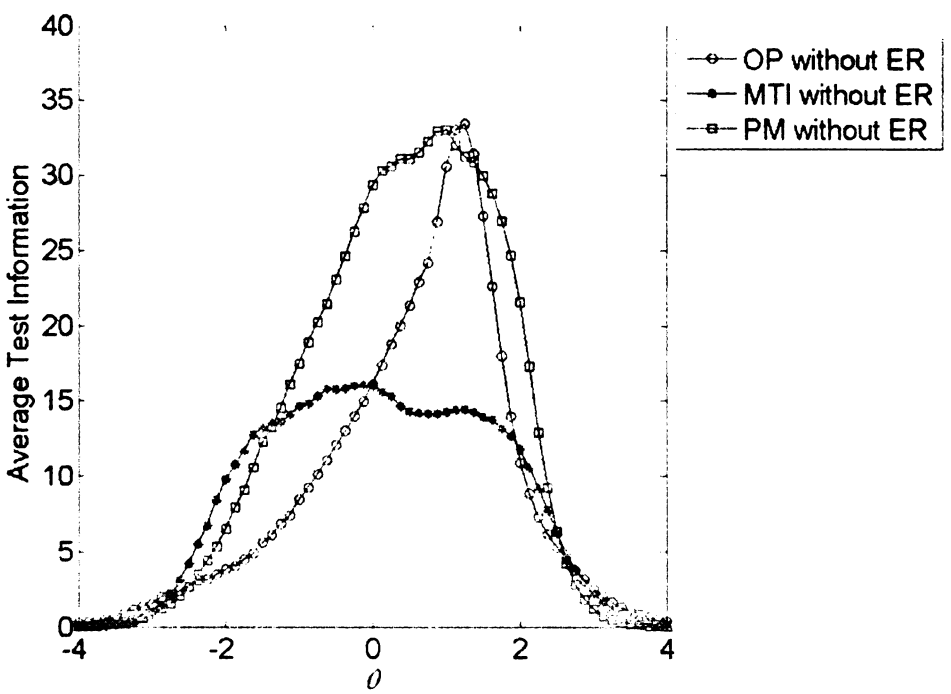


Figure 5.4 Average test information conditional on true $\theta$

Figure 5.5 to 5.7 present the CSEM, conditional bias, and CMSE for the three item pools. Figure 5.6 shows a significant increase in the biase of ability estimation, which is positive for the ability levels below around -2.0 and negative for the ability levels above around 2.0. It is not surprising because of the short test length and the Baysian estimation method. The charts show that MTI perform better for ability level below -2.0 and PM performs better for ability level over 2.5.

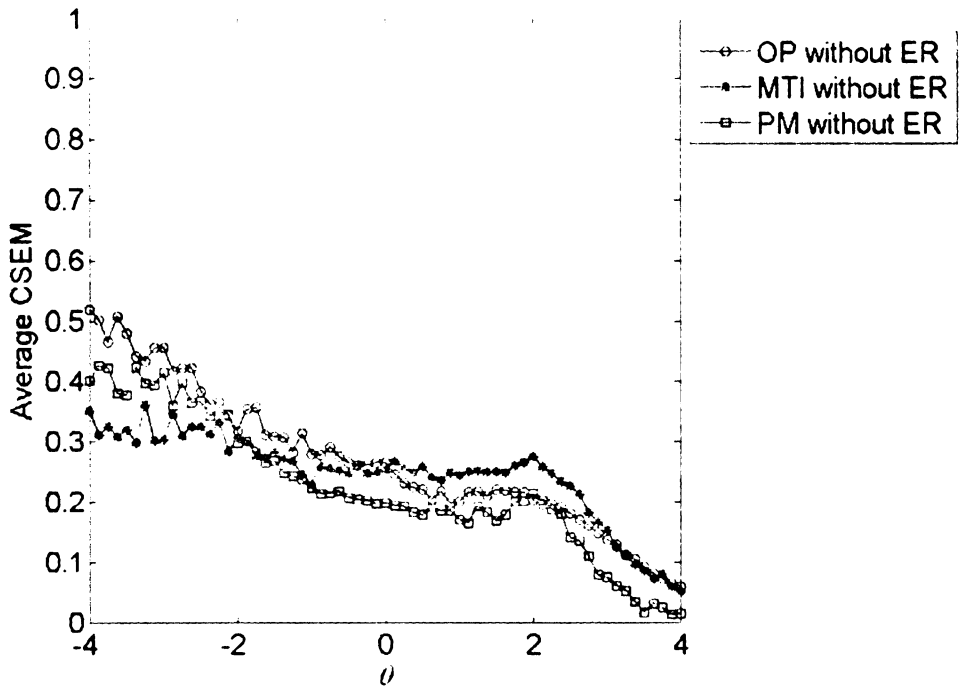


Figure 5.5 Conditional standard error of measurement (CSEM)

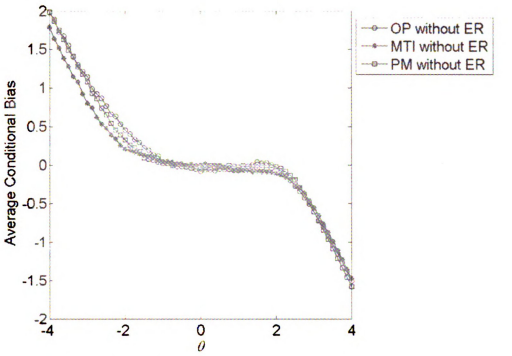


Figure 5.6 Conditional bias

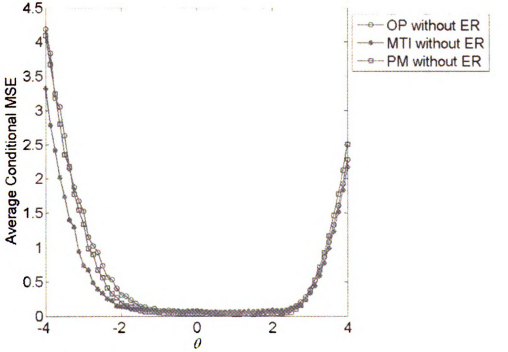


Figure 5.7 Conditional mean square error (CMSE)

*5.2 Item Pools for Tests with Content Balance*

As can be seen from Figure 5.8, the distribution of the item pools with content balancing shares a very similar pattern with the distribution of the item pools without content balancing. Both optimal pools show a more even distribution in difficulty levels, but the PM pool has more highly discriminating items in difficult items and very few for easy items. The MTI pool has mostly moderately large *a*-parameters, regardless of the difficulty levels. Table 5.3 lists the item pool size and the summary statistics of the items in the pool. Compared to the operational pool, on average the MTI pool and the PM pool have slightly higher *a*-parameters and lower *b*-parameters but the range of the item parameters are smaller. Both optimal pools consist of fewer items in two larger contents

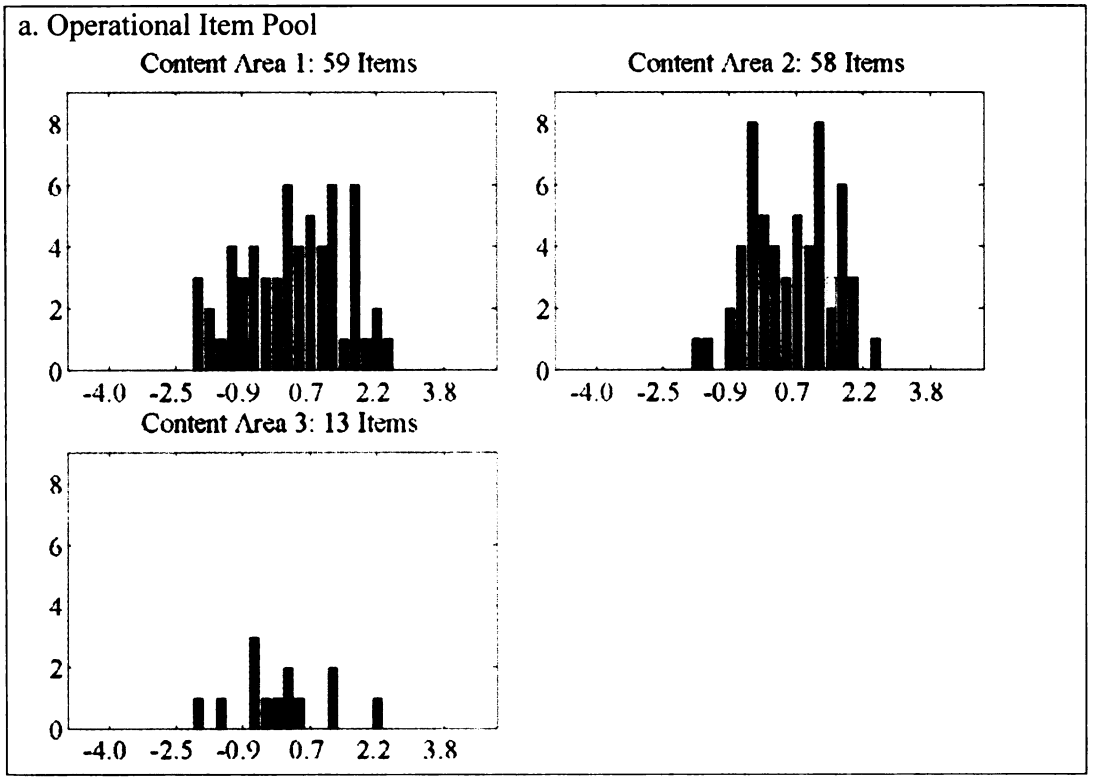


Figure 5.8 Item distribution for item pools with content balancing and without exposure control

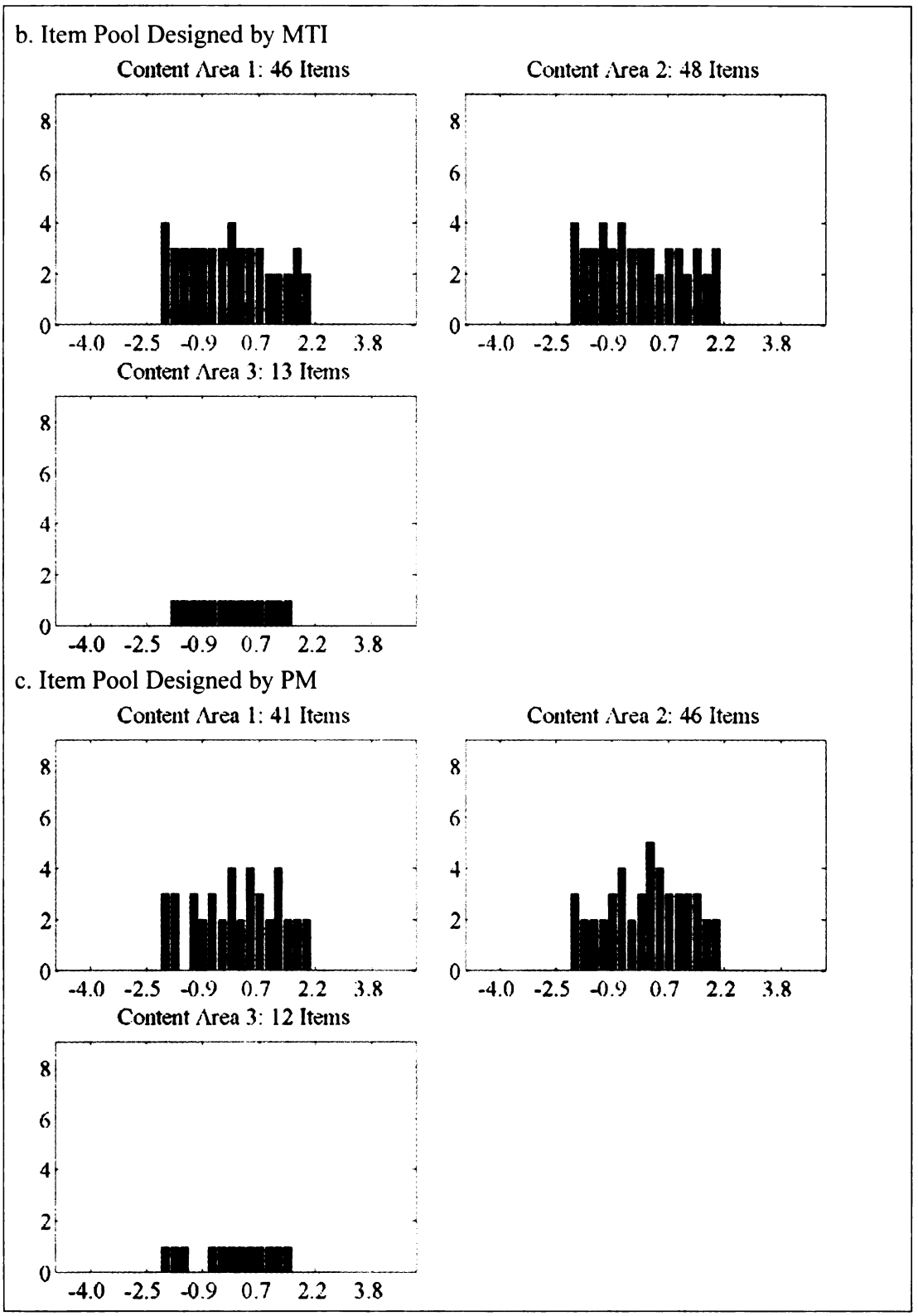


Figure 5.8 (Conti'd) Item distribution for item pools with content balancing and without exposure control

Table 5.3 Item Pool Size and Item Parameter Statistics for General Science without Exposure Control

| Pool | Pool Size | *a* | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. |
| Content 1 | | | | | | | | | | | | | |
| OP | 59 | 1.167 | 0.354 | 2.314 | 0.628 | 0.099 | 1.189 | 2.211 | -2.211 | 0.224 | 0.067 | 0.396 | 0.091 |
| MTI | 46 | 1.610 | 0.116 | 1.783 | 1.268 | -0.191 | 1.186 | 2.010 | -2.069 | 0.276 | 0.082 | 0.436 | 0.130 |
| PM | 41 | 1.558 | 0.274 | 1.894 | 0.923 | 0.031 | 1.156 | 2.015 | -2.065 | 0.234 | 0.054 | 0.360 | 0.115 |
| Content 2 | | | | | | | | | | | | | |
| OP | 58 | 1.215 | 0.377 | 2.732 | 0.564 | 0.350 | 0.997 | 2.215 | -1.863 | 0.218 | 0.069 | 0.459 | 0.078 |
| MTI | 48 | 1.593 | 0.100 | 1.767 | 1.295 | -0.127 | 1.214 | 2.066 | -2.054 | 0.267 | 0.086 | 0.463 | 0.104 |
| PM | 46 | 1.582 | 0.285 | 1.977 | 0.897 | 0.030 | 1.119 | 2.053 | -1.945 | 0.223 | 0.070 | 0.401 | 0.094 |
| Content 3 | | | | | | | | | | | | | |
| OP | 13 | 1.087 | 0.282 | 1.776 | 0.726 | -0.195 | 1.128 | 1.917 | -2.137 | 0.220 | 0.058 | 0.305 | 0.115 |
| MTI | 13 | 1.625 | 0.098 | 1.758 | 1.398 | -0.121 | 1.007 | 1.343 | -1.742 | 0.220 | 0.050 | 0.327 | 0.158 |
| PM | 12 | 1.285 | 0.223 | 1.525 | 0.722 | -0.106 | 1.092 | 1.368 | -1.894 | 0.202 | 0.062 | 0.336 | 0.084 |

Table 5.4 Summary Statistics of the Performance of the Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0080 | 0.0080 | 0.0135 |
| MSE | 0.1204 | 0.0915 | 0.0860 |
| Correlation | 0.9365 | 0.9536 | 0.9569 |
| Skewness of item exposure rate | 13.6889 | 8.8820 | 8.0765 |
| Item overlap rate | 0.4190 | 0.3325 | 0.3268 |
| Pct of items with item exposure Rate> 1/3 | 11.54% | 12.15% | 13.13% |
| Pct of items with item exposure Rate<.02 | 51.54% | 21.50% | 19.19% |
| Pool Size | 130 | 107 | 99 |



Table 5.5 Number of Over- and Under-Exposed Items by Content

| | Content 1 | | | Content 2 | | | Content 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Statistic | OP | MTI | PM | OP | MTI | PM | OP | MTI | PM |
| Pct of items with item exposure Rate> 1/3 | 11.86% | 13.04% | 14.63% | 12.07% | 14.58% | 13.04% | 7.69% | 0.00% | 8.33% |
| Pct of items with item exposure Rate<.02 | 55.93% | 26.09% | 12.20% | 46.55% | 16.67% | 17.39% | 53.85% | 23.08% | 50.00% |
| Pool Size | 59 | 46 | 41 | 58 | 48 | 46 | 13 | 13 | 12 |

but have similar number of items in the content with only one item in the test. The MTI pool has the fewest items.

The overviews of the evaluation results for these item pools are presented in Table 5.4 and 5.5. The ability estimates from all pools exhibit slightly positive bias. The MSE from the optimal item pools are smaller than that from the operational pool. In addition, both optimal item pools result in a higher correlation coefficient.

The results show that optimal item pools have smaller test-retest overlap rate despite having fewer items. Figure 5.9 also shows that the MTI item pools have the smallest test-retest overlap rates at most ability levels.

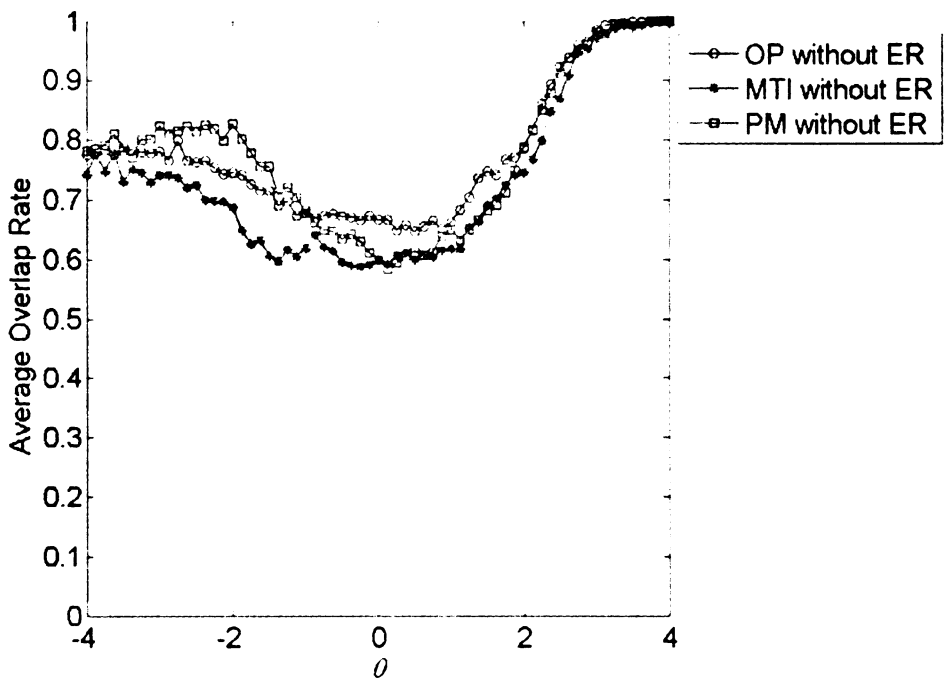


Figure 5.9 Test-retest overlap rate conditional on $\theta$

Figure 5.10 plots the item exposure rate for individual items in each item pool by content level. Regardless of the content, both optimal item pools have significantly smaller numbers of under-exposed items. On the other hand, Table 5.4 shows that the

percentages of over-exposed items are similar for all item pools, although optimal item pools have fewer items.

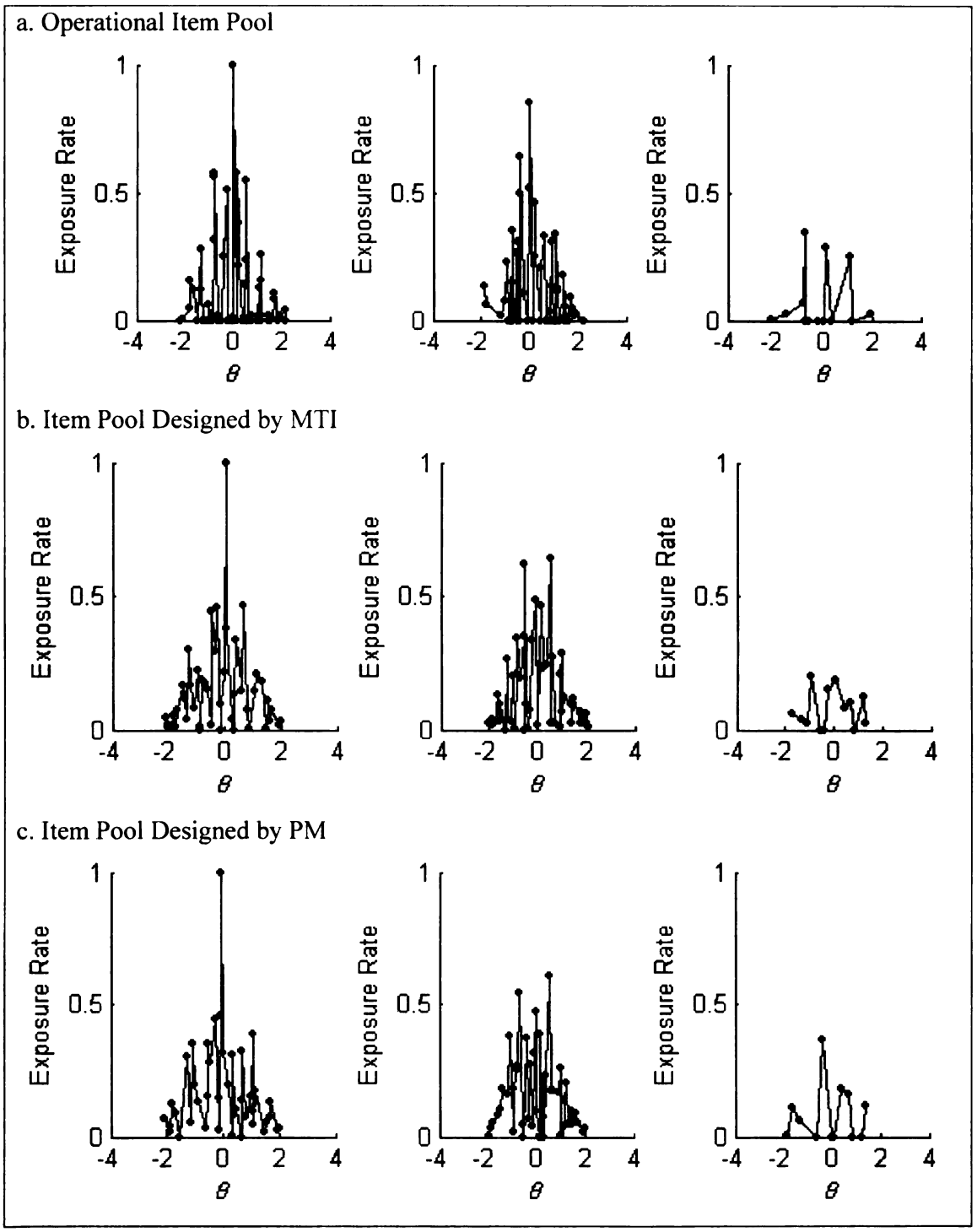


Figure 5.10 Item exposure rate by difficulty level

Figure 5.11 displays the plot of conditional test information for the item pools. It can be seen that the PM pool results in a plot similar to the operational pool but provides more information at most ability levels. The MTI item pool provides similar amount of information, which exceeds the target information, over the ability levels between approximately -2.0 and 2.0.

The CSEM, conditional bias, and CMSE for three item pools are presented in Figure 5.12 to 5.14 present. The charts show that at most ability levels, the three item pools perform very closely. The MTI item pool results in the smallest bias and CMSE for an ability level below approximately -1.5.

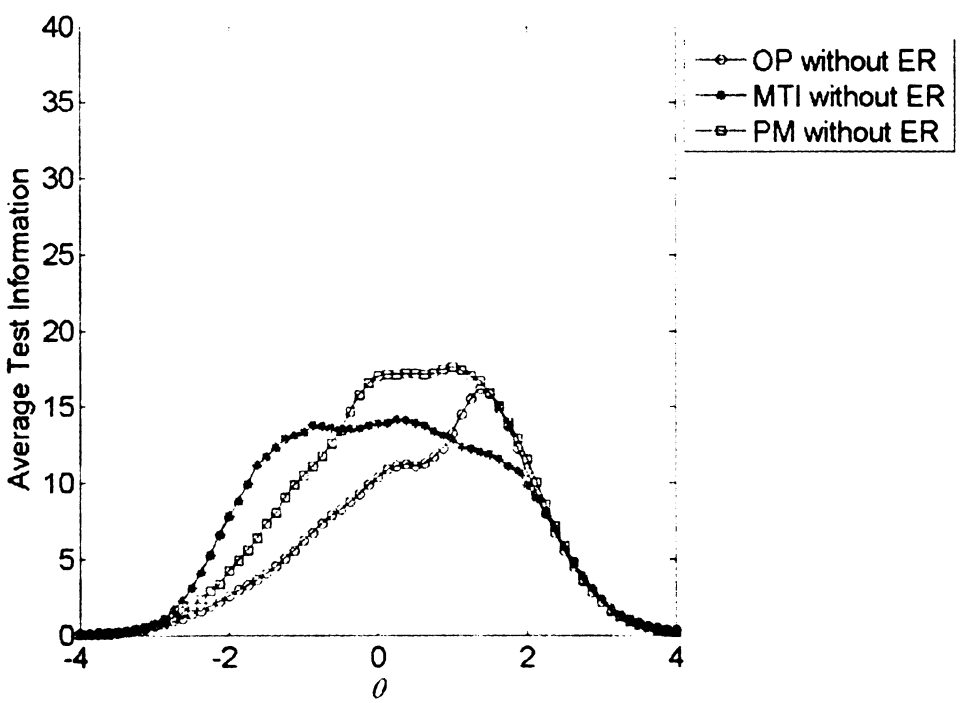


Figure 5.11 Average test information conditional on true $\theta$

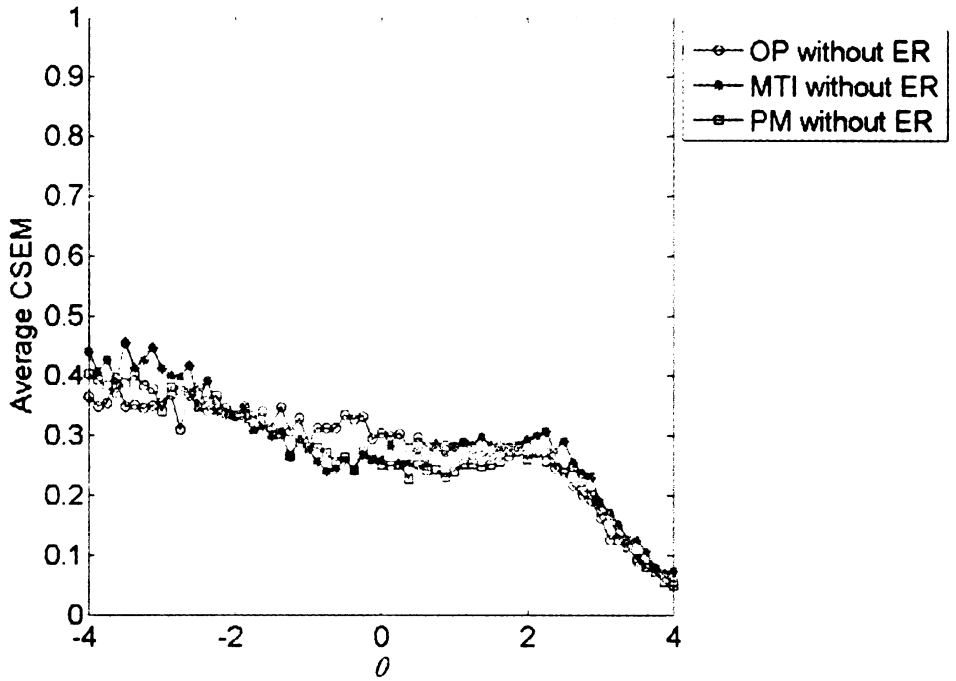


Figure 5.12 Conditional standard error of measurement (CSEM)

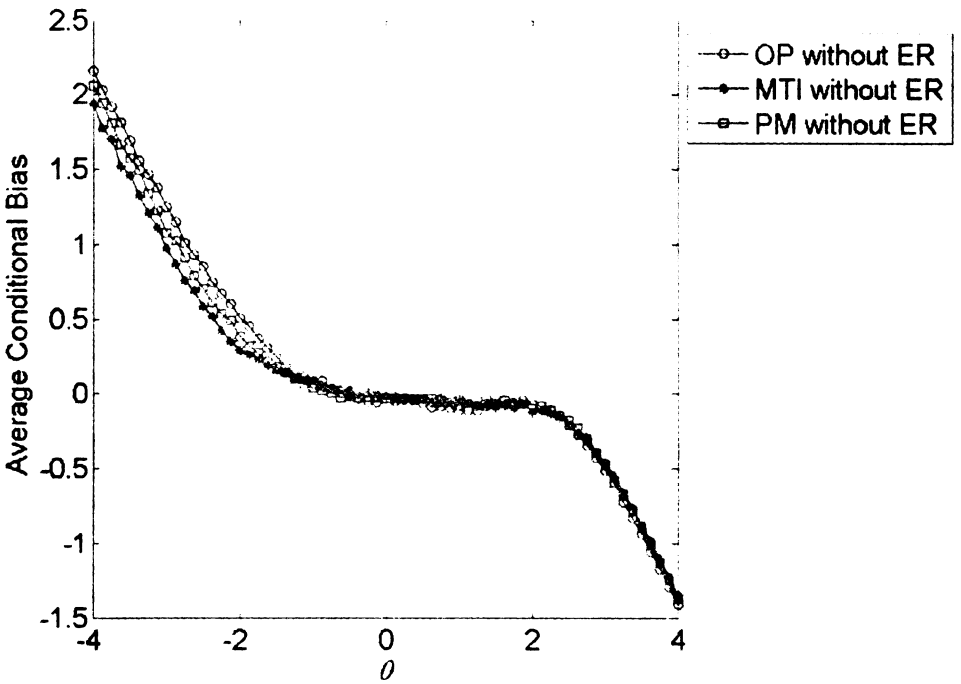


Figure 5.13 Conditional bias

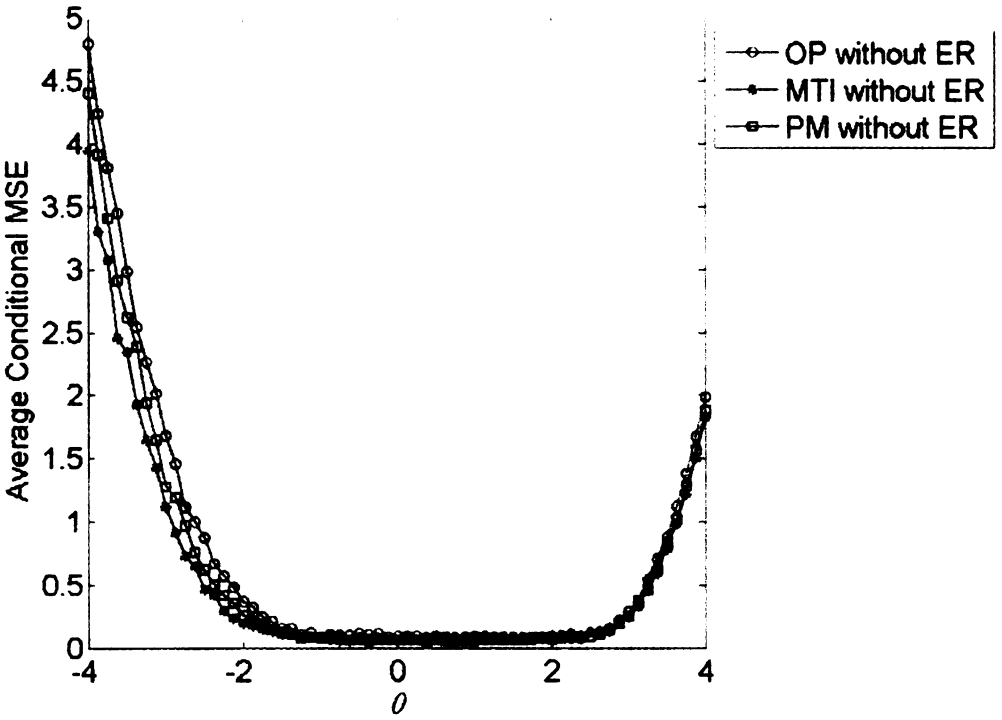


Figure 5.14 Conditional mean square error (CMSE)

### *5.3 Summary*

The results suggest that regardless of the constraints of content balancing, the optimal item pools perform better than the operational item pool based on pool size, test security and measurement accuracy, although each design method has its preferable features. The operational item pool performs better over a given range of ability levels because a large number of items, including very discriminating items, are clustered around these levels. Optimal item pools, especially ones designed with MTI, provide information more evenly over most ability levels and provide sufficient measurement precision with a minimum number of items. All optimal pools, compared to operational pools, save about 20 or more items and yield better correlations. In addition, optimal pools have a significantly lower percentage of items with exposure rate below 0.02. With

or without content balancing, PM item pools resulted in the highest correlation and the lowest item overlap rate.

Overall, it seems that an item pool designed with the MTI method performs the best, which indicates that the optimal item pool needs the fewest items to achieve desirable precision if all the items have moderate item discrimination and distribute roughly uniformly over a wide range of difficulty levels.

Chapter VI The Performance of the Item Pool with Sympson-Hetter Exposure Control

### *6.1 Item Pools for Tests without Content Balance*

The blueprint of the optimal item pools with SH exposure control is based on the blueprint of those without exposure control. Specifically, more items are added in the bins where items tend to be selected more often than the desired exposure rates. This relationship is reflected in the item distribution illustrated in Figure 6.1, where compared to the optimal pools without exposure control, there are noticeably more items with *b*-parameters -1.0 to 1.0 in optimal pools designed by both MTI and PM.

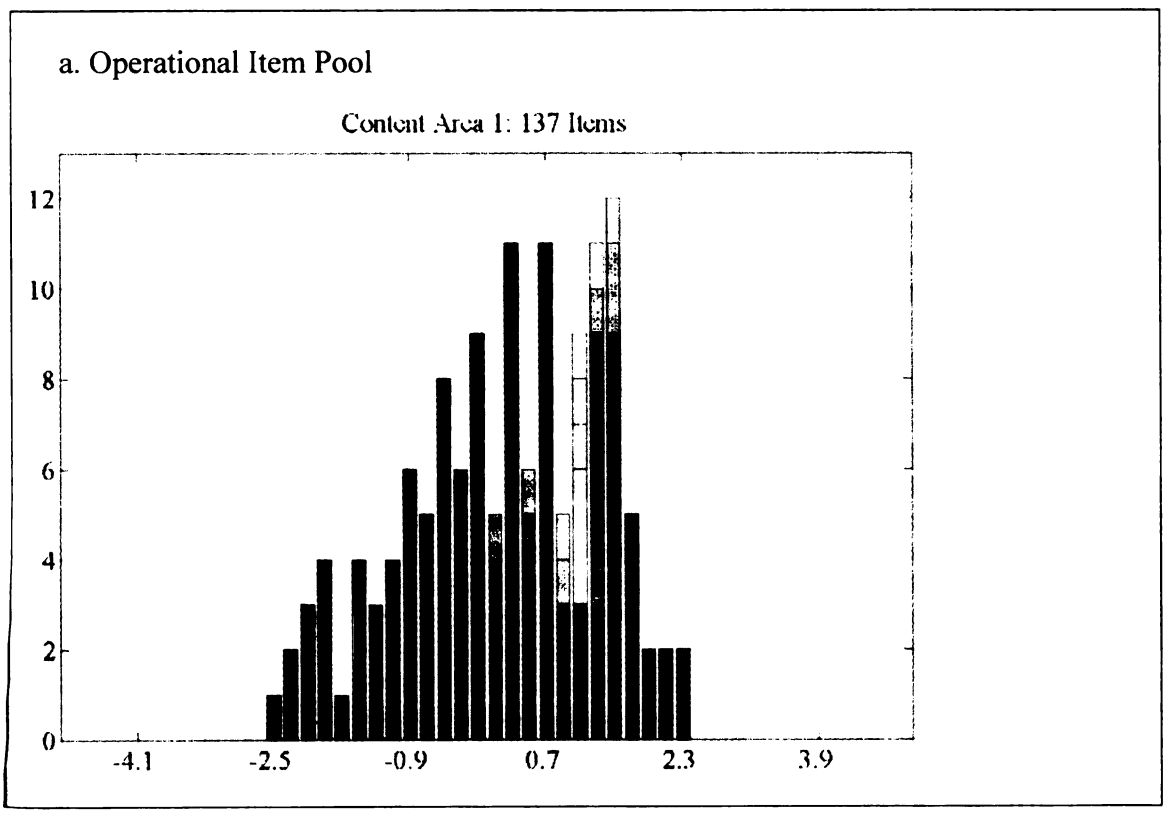


Figure 6.1 Item Distributions for item pools with Sympson-Hetter exposure control

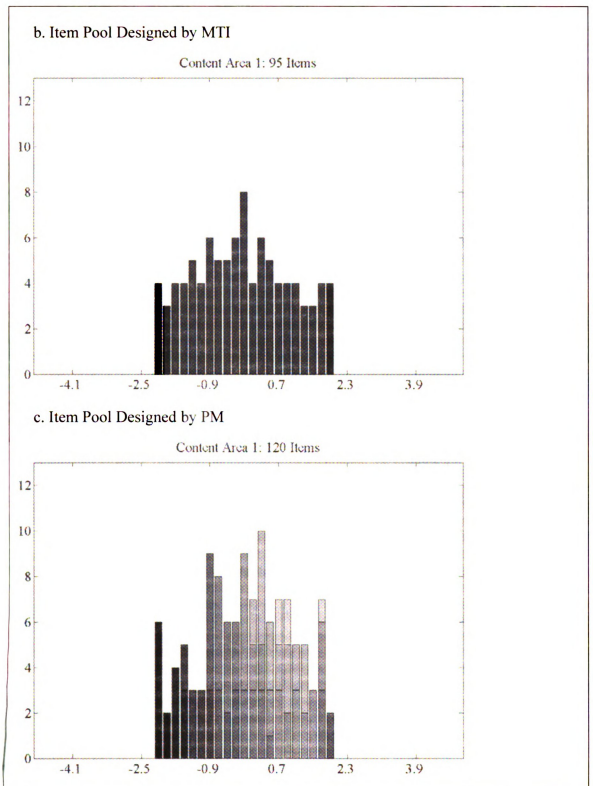


Figure 6.1 (Cont'd) Item Distributions for item pools with Sympson-Hetter exposure control

Table 6.1 shows the item pool size and the summary statistics of the item parameters within each pool. The MTI pool consists of the fewest items and their *a*-parameters vary within 1.307 and 1.777, a smaller range than the other two pools. The optimal item pool designed by PM has more high *a*-parameter items. However, items in the operational pool have the maximum *a*-parameter value in 3.141, compared to 1.777 for items in the MTI pool and 2.633 for items in the PM pool.

The MTI pool has 13 more items and the PM pool has 19 more items than the item pools without exposure control, but the size of either pool is still smaller than that of the operational pool. The added items are mostly highly discriminating items because they tend to have higher exposure rates. This leads to a slightly higher average *a*-parameter for the optimal item pools.

Table 6.2 lists the performance overview of the item pools. On average, all three pools yielded slightly positive bias for ability estimates. The operational pool displayed the smallest bias, but the difference from the optimal pools is negligible. Both optimal pools exhibited better performance on all other criteria. The PM pool resulted in the highest correlation coefficient and the lowest mean square error. The MTI pool, however, consists of the least items, which is 42 items less than the operational pool and 25 less than the PM pool.

Table 6.1 Item Pool Size and Item Parameter Statistics for Arithmetic Reasoning with Sympson-Hetter Exposure Control

| Pool | Pool Size | *a* | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. |
| OP | 137 | 1.556 | 0.487 | 3.141 | 0.746 | 0.115 | 1.170 | 2.343 | -2.625 | 0.186 | 0.063 | 0.328 | 0.038 |
| MTI | 95 | 1.616 | 0.092 | 1.777 | 1.307 | -0.141 | 1.130 | 1.935 | -2.172 | 0.228 | 0.076 | 0.498 | 0.082 |
| PM | 120 | 2.027 | 0.410 | 2.633 | 0.922 | -0.055 | 1.083 | 1.922 | -2.182 | 0.180 | 0.059 | 0.337 | 0.054 |

Table 6.2 Summary Statistics of the Performance of the Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0073 | 0.0104 | 0.0105 |
| MSE | 0.0929 | 0.0823 | 0.0564 |
| Correlation | 0.9525 | 0.9593 | 0.9728 |
| Skewness of item exposure rate | 18.9078 | 8.5813 | 10.7972 |
| Item overlap rate | 0.2474 | 0.2481 | 0.2149 |
| Pct of items with item exposure Rate> 1/3 | 5.11% | 11.58% | 4.17% |
| Pct of items with item exposure Rate<.02 | 39.42% | 11.58% | 17.50% |
| Pool Size | 137 | 95 | 120 |

As shown in Figure 6.2, all three pools exhibit smaller test-retest overlap rates compared to the item pools without exposure control. This is anticipated because item selection with the exposure control method tends to utilize more items. Of the item pools with exposure control, the operational pool clearly has the smallest test-retest overlap rate at the ability levels larger than 2.0. However, Table 6.2 shows that on average, optimal item pools have slightly smaller test-retest overlap rate despite having fewer items.

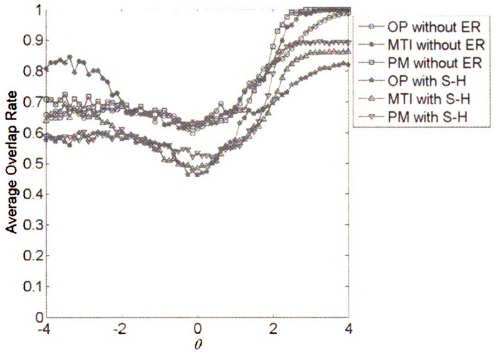


Figure 6.2 Test-retest overlap rate conditional on $\theta$

Figure 6.3 shows the item exposure rate for individual items in each pool in the order of the item difficulty. It can be seen that the exposure control mechanic works very well, with the exposure rates for all individual items around or below the target exposure rate. The MTI pool seems to utilize items more evenly and have the fewest under-exposed items. The operational item pool seems to have large numbers of difficult items underexposed.

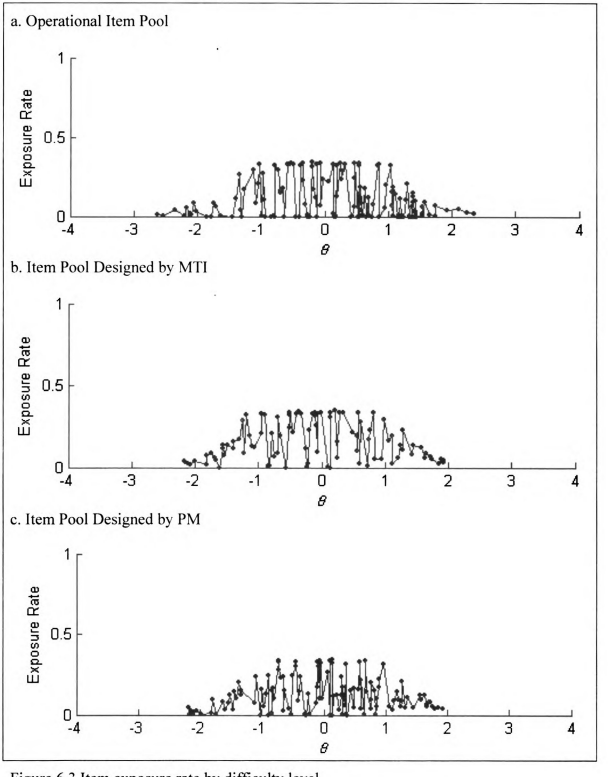


Figure 6.3 Item exposure rate by difficulty level

A closer look at the measurement precision at the individual ability level is displayed with the conditional test information plots in Figure 6.4. The plots for item

pools with exposure control look very similar to those without exposure control. Because of added items, optimal pools with SH exposure control yield more information at some ability levels and closely match the information provided at other levels with the optimal pool without exposure control. The operational pool, on the other hand, produces less information at the ability levels between -0.5 to 0.75 when SH exposure control is used.

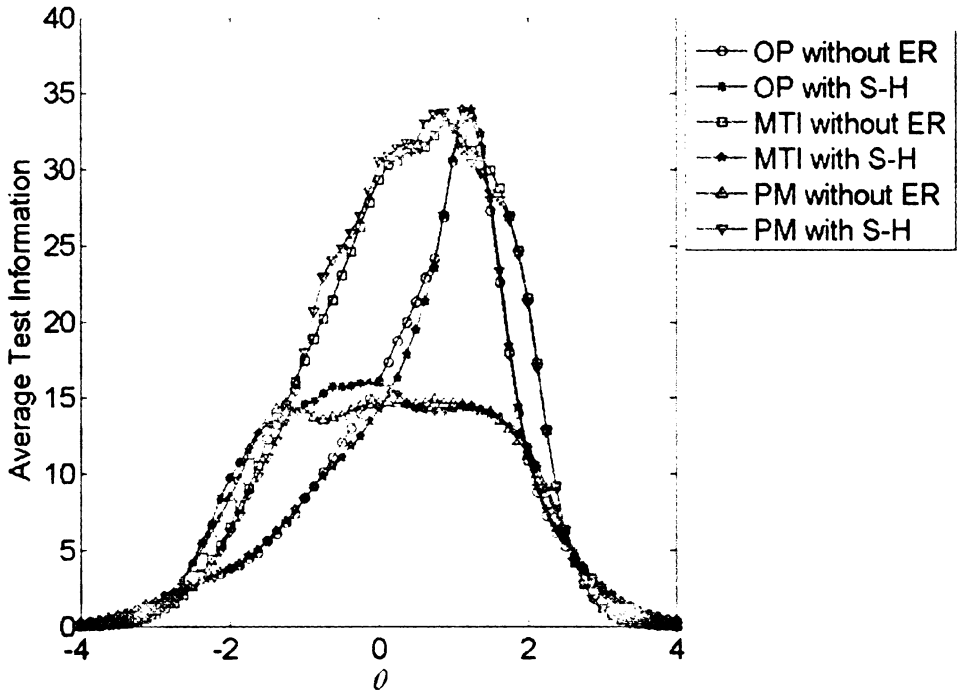


Figure 6.4 Average test information conditional on true $\theta$

The plots for conditional SEM, conditional bias, and conditional MSE are presented in Figure 6.5 to 6.7. Smaller values indicate better accuracy in ability estimate. The plots indicate that all item pools yield similar performance at ability levels between -2.0 and 2.5. The MTI pool performs better for ability levels below -2.0 and the PM pool performs better for ability levels over 2.5.

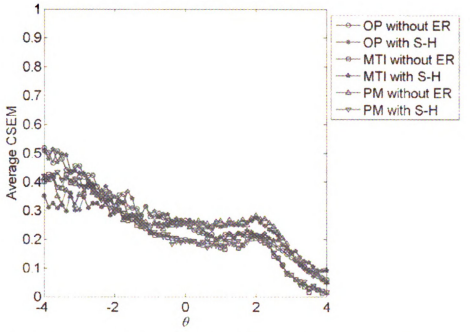


Figure 6.5 Conditional standard error of measurement (CSEM)

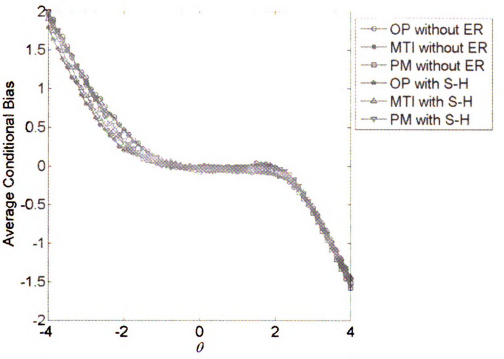


Figure 6.6 Conditional bias

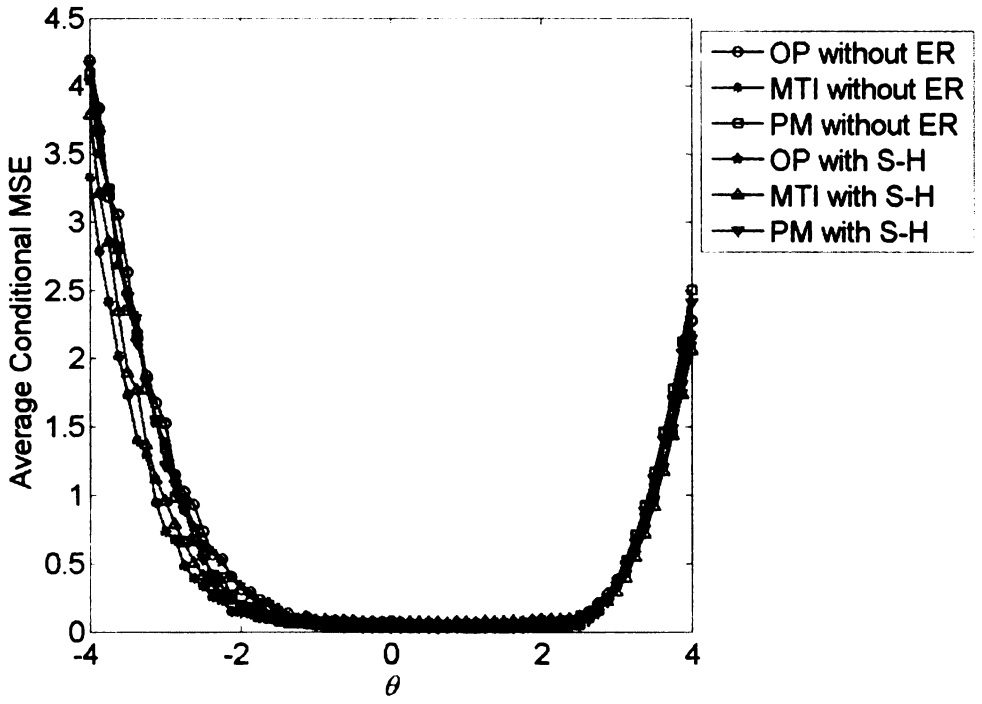


Figure 6.7 Conditional mean square error (CMSE)

### *6.2 Item Pools for Tests with Content Balance*

As shown in Figure 6.8, the optimal item pools for tests with content balancing display the same pattern as those without content balancing. Items are added to the bins where items tend to have more than the desired exposure rates. Because content 3 appears only once in a test, no additional items are added in comparison to the pools without exposure control. An interesting fact is that in optimal item pools, content 2 has slightly more items than content 1 does, although both contents appear in a test with an equal number of items.

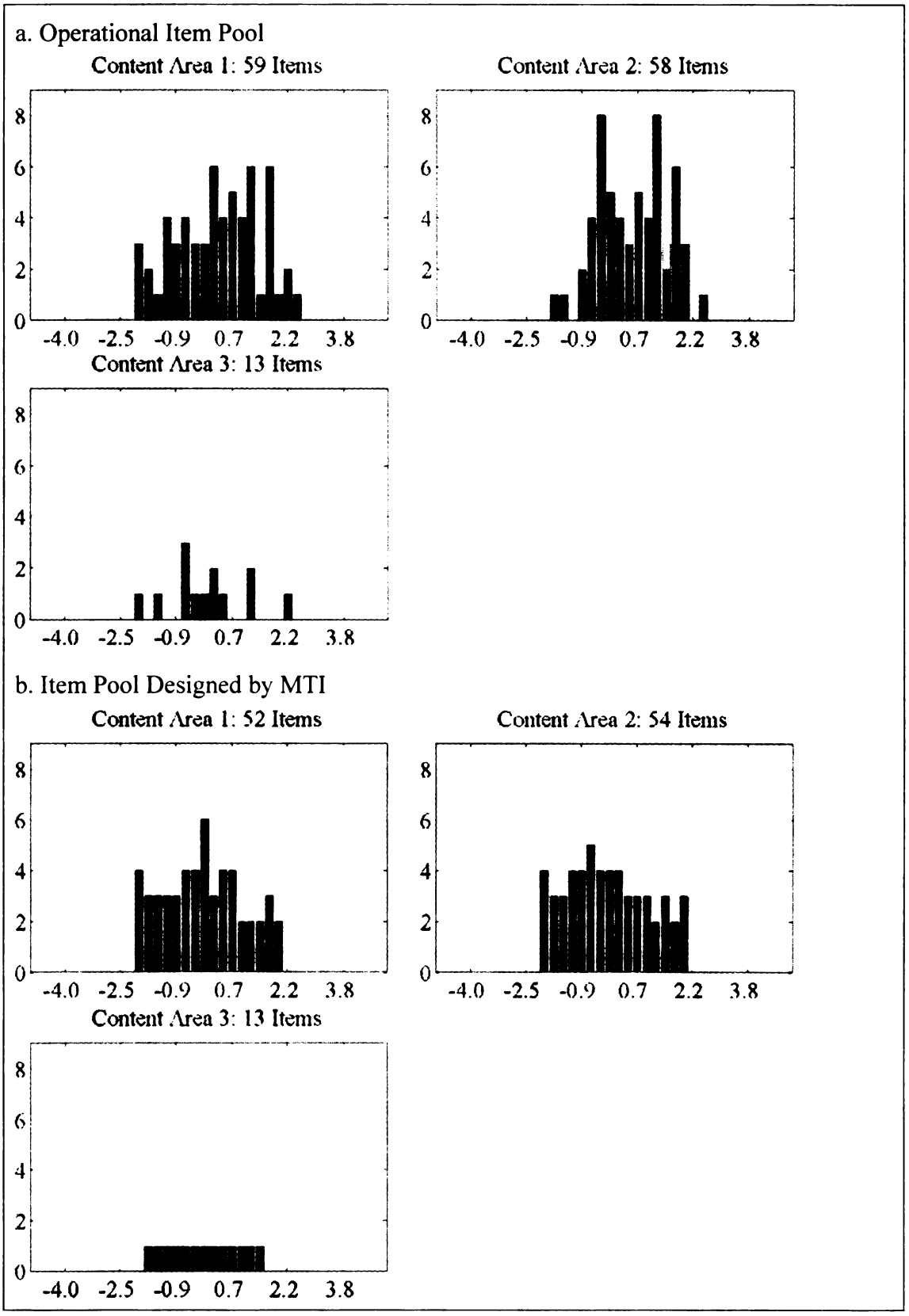


Figure 6.8 Item distribution for item pools with Sympson-Hetter exposure control

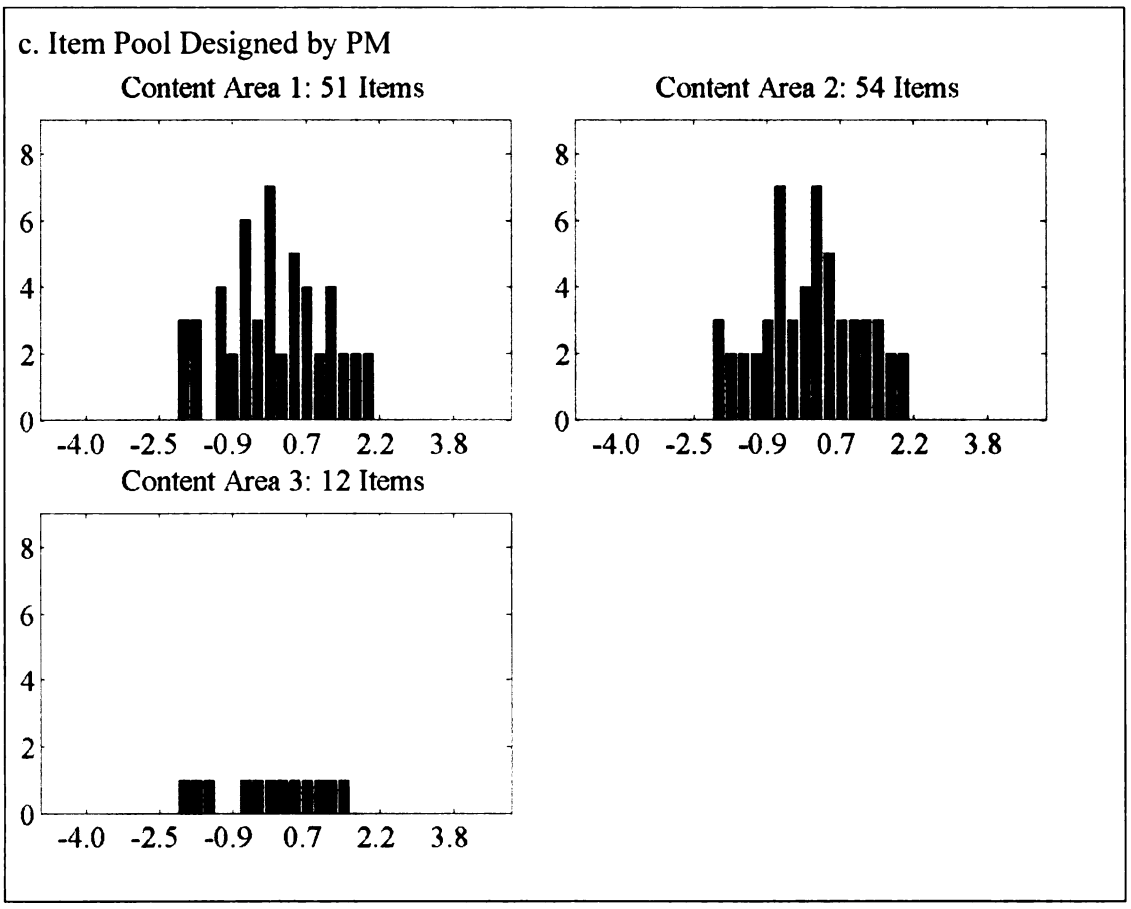


Figure 6.8 (Conti'd) Item distribution for item pools with Sympson-Hetter exposure control

Table 6.3 shows the item pool size and the summary statistics of the item parameters within each pool. The total numbers of items in the two optimal item pools are similar while both sizes are smaller than the operational pool. Compared to the item pools without exposure control, the MTI pool has 12 more items and the PM pool has 18 more items. On average, items in the optimal pools have higher $a$-parameters and relatively lower $b$-parameters than the operational pool.

Table 6.3 Item Pool Size and Item Parameter Statistics for General Science with Sympson-Hetter Exposure Control

| Pool | Pool Size | *a* Mean | *a* SD | *a* Max. | *a* Min. | *b* Mean | *b* SD | *b* Max. | *b* Min. | *c* Mean | *c* SD | *c* Max. | *c* Min. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Content 1 | | | | | | | | | | | | | |
| OP | 59 | 1.167 | 0.354 | 2.314 | 0.628 | 0.099 | 1.189 | 2.211 | -2.211 | 0.224 | 0.067 | 0.396 | 0.091 |
| MTI | 52 | 1.602 | 0.105 | 1.784 | 1.271 | -0.152 | 1.122 | 1.960 | -2.012 | 0.253 | 0.074 | 0.452 | 0.092 |
| PM | 51 | 1.589 | 0.286 | 1.994 | 0.899 | -0.033 | 1.067 | 2.048 | -1.992 | 0.221 | 0.057 | 0.343 | 0.087 |
| Content 2 | | | | | | | | | | | | | |
| OP | 58 | 1.215 | 0.377 | 2.732 | 0.564 | 0.350 | 0.997 | 2.215 | -1.863 | 0.218 | 0.069 | 0.459 | 0.078 |
| MTI | 54 | 1.586 | 0.098 | 1.750 | 1.278 | -0.161 | 1.142 | 1.981 | -2.059 | 0.271 | 0.077 | 0.426 | 0.116 |
| PM | 54 | 1.623 | 0.260 | 1.994 | 0.916 | -0.029 | 1.035 | 1.989 | -1.938 | 0.215 | 0.061 | 0.452 | 0.107 |
| Content 3 | | | | | | | | | | | | | |
| OP | 13 | 1.087 | 0.282 | 1.776 | 0.726 | -0.195 | 1.128 | 1.917 | -2.137 | 0.220 | 0.058 | 0.305 | 0.115 |
| MTI | 13 | 1.620 | 0.114 | 1.786 | 1.371 | -0.105 | 1.000 | 1.524 | -1.620 | 0.262 | 0.050 | 0.371 | 0.203 |
| PM | 12 | 1.317 | 0.204 | 1.521 | 0.788 | -0.145 | 1.095 | 1.305 | -2.005 | 0.223 | 0.080 | 0.370 | 0.134 |

Table 6.4 Summary Statistics of the Performance of the Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0125 | 0.0113 | 0.0107 |
| MSE | 0.1323 | 0.0881 | 0.0891 |
| Correlation | 0.9301 | 0.9553 | 0.9557 |
| Skewness of item exposure rate | 8.5601 | 6.9623 | 6.6554 |
| Item overlap rate | 0.2725 | 0.2345 | 0.2343 |
| Pct of items with item exposure Rate> 1/3 | 13.08% | 4.20% | 4.27% |
| Pct of items with item exposure Rate<.02 | 43.85% | 20.17% | 25.64% |
| Pool Size | 130 | 119 | 117 |



Table 6.5 Number of Over- and Under-Exposed Items by Content

| | Content 1 | | | Content 2 | | | Content 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Statistic | OP | MTI | PM | OP | MTI | PM | OP | MTI | PM |
| Pct of items with item exposure Rate> 1/3 | 15.25% | 5.77% | 1.96% | 13.79% | 3.70% | 7.41% | 0.00% | 0.00% | 0.00% |
| Pct of items with item exposure Rate<.02 | 44.07% | 23.08% | 19.61% | 41.38% | 16.67% | 25.93% | 53.85% | 23.08% | 50.00% |
| Pool Size | 59 | 52 | 51 | 58 | 54 | 54 | 13 | 13 | 12 |

Table 6.4 and 6.5 display the overview of the evaluation results for these item pools. The performance shows the same pattern as the item pools without content balancing. The MSE from the optimal item pools is smaller than that from the operational pool. The PM item pool results in the highest correlation coefficient.

Table 6.4 also shows that optimal item pools have smaller test-retest overlap rate despite having fewer items. The plots for conditional test-retest overlap rates in Figure 5.2 reveal that the operational pool has the lowest overlap rates at ability levels below -1.5 and the MTI pool has the lowest overlap rates at levels beyond approximately 1.00. This is consistent with the findings from item pools without exposure control, but different from the ones for item pools without content balancing, where the operational pool has the lowest overlap rates at the two extremes of the ability levels.

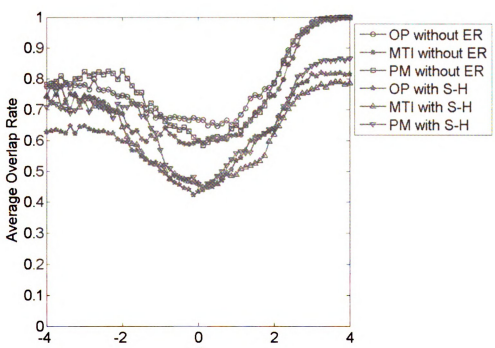


Figure 6.9 Test-retest overlap rate conditional on $\theta$

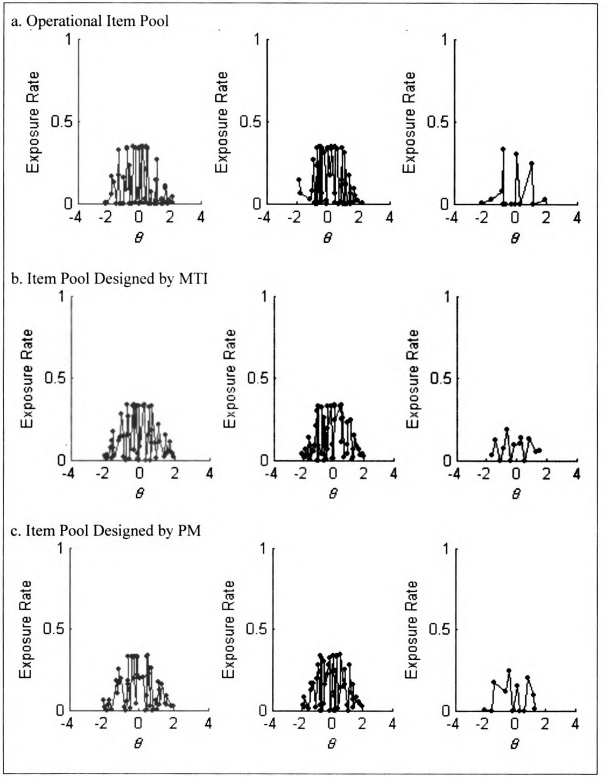


Figure 6.10 Item exposure rate by difficulty level

Table 6.4 and 6.5 also shows that, compared to the operational pool, both optimal item pools have significantly smaller numbers of under-exposed, as well as over-exposed, items. The number of under-exposed items is similar to the pools without exposure control, but the number of over exposed items decreases substantially.

Figure 6.10 shows the item exposure rate for individual items ordered by their difficulty levels in each item pool. Extremely easy and extremely difficult items tend to have less exposure rates, but under-exposed items are across all difficulty levels, especially for the operational item pool. As indicated in Table 6.4, the MTI item pool has the fewest under-exposed items, and utilizes the extreme difficulty level items more often.

Figure 6.11 presents the conditional test information plots. The plots for item pools with exposure control look very similar to those without exposure control. The optimal pools with exposure control yield similar or more information than the pool without exposure control, except that the MTI pool with exposure control yields a smaller amount of information at the ability levels between 0 and 1.5. The operational pool produces less information at the ability levels between -1.0 to 1.0 when compared to the same pool without exposure control.

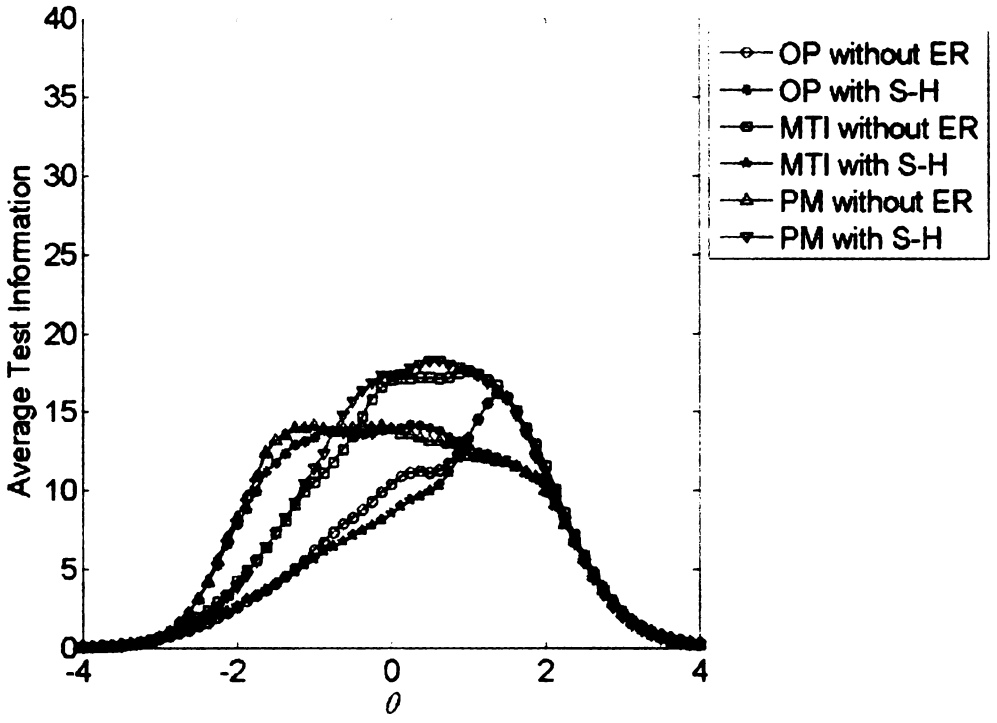


Figure 6.11 Average test information conditional on true $\theta$

Figure 6.12 presents the CSEM for three item pools. It shows that the MTI pool performs better at ability levels below -2 and the PM pool performs better at ability levels over 0.0. The operational pool has the largest SEM at ability levels between -2.0 and 1.5.

Figure 6.13 and 6.14 show that the MTI pool yields the smallest bias and the smallest MSE at most ability levels. The operational pool and PM pool produce similar MSE.

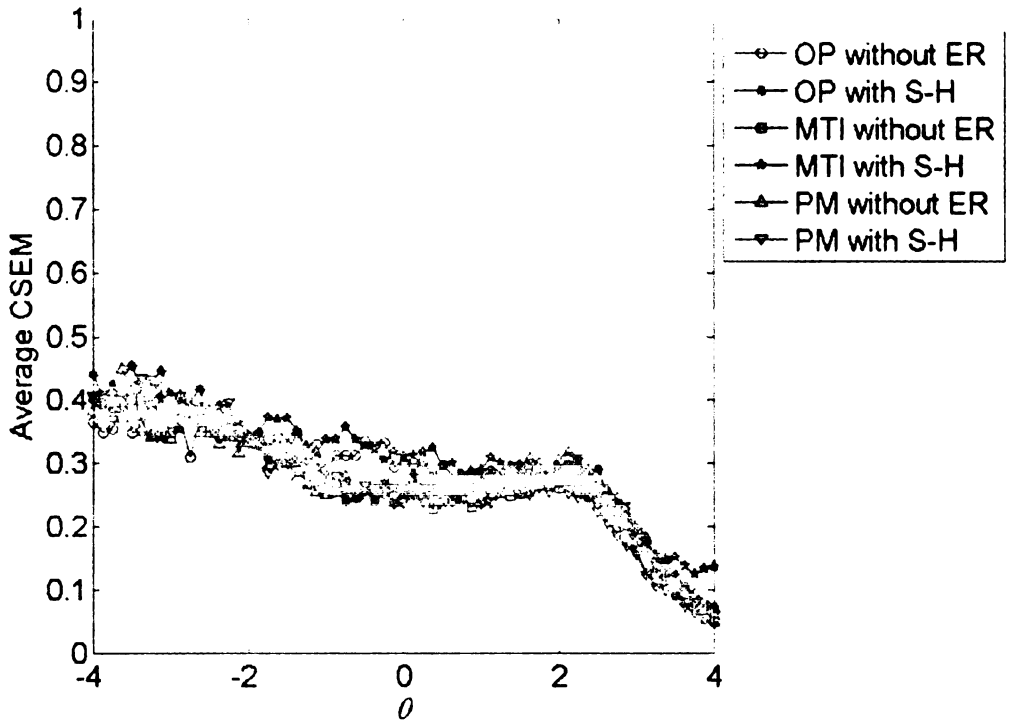


Figure 6.12 Conditional standard error of measurement (CSEM)

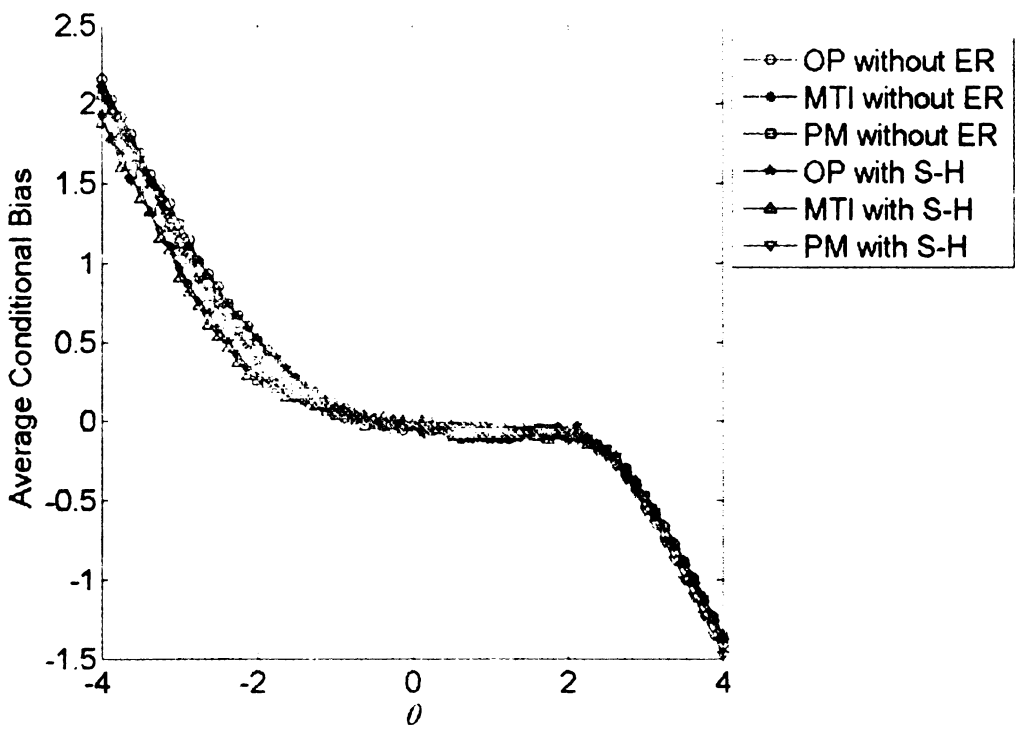


Figure 6.13 Conditional bias

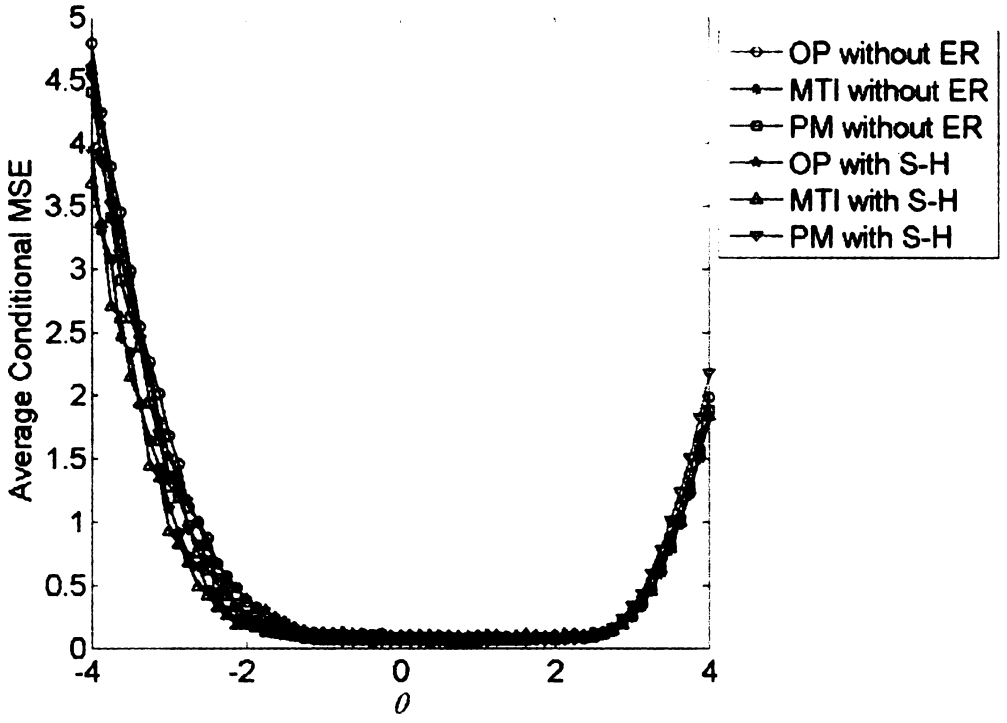


Figure 6.14 Conditional mean square error (CMSE)

### *6.3 Summary*

The results suggest all optimal pools, compared to operational pools, save about 10 or more items while performing better based on pool size, test security and measurement accuracy. Tests assembled from optimal pools have smaller test-retest overlap rates. In addition, optimal pools have significantly lower percentages of items with exposure rate below 0.02.

Chapter VII The Performance of the Item Pools with

*a*-Stratified Exposure Control

### *7.1 Item Pools for Tests without Content Balance*

The *a*-stratified exposure control selects item by the closeness of *b*-parameter to the provisional ability estimate instead of maximum information. Therefore, it is not necessary to adjust the item pool after CAT simulations. The item distribution for the OP, MTI, and PM pools are displayed in Figure 7.1.

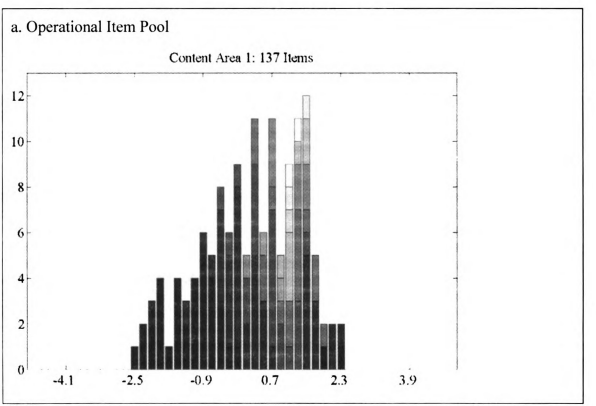


Figure 7.1 Item distribution for item pools without content balancing and with *a*-stratified exposure control

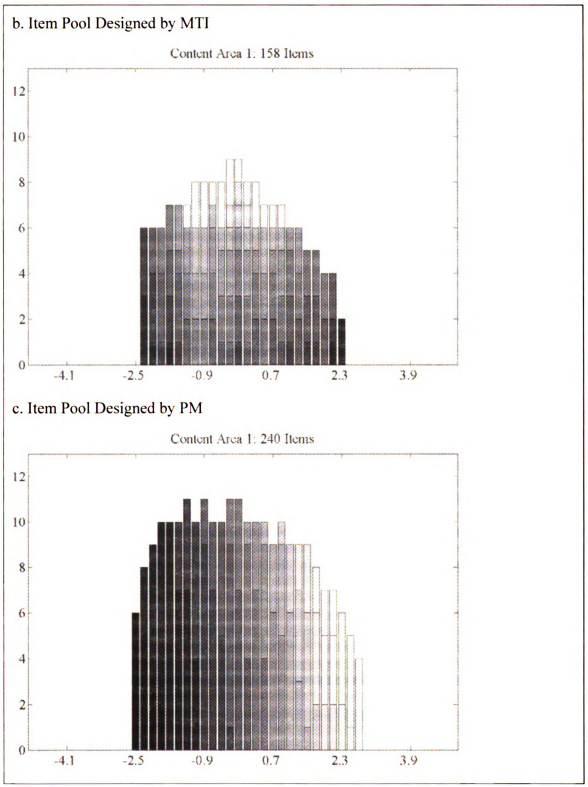


Figure 7.1 Item distribution for item pools without content balancing and with *a*-stratified exposure control

As can be seen, the shapes of the distribution look similar for the MTI pool and the PM pool, although the MTI pool is much smaller. The PM pool has more difficult items with high *a*-parameters, while the MTI pool has more moderately difficult items with high *a*-parameters. Both optimal pools seem to require more items with moderate to low *b*-parameters and fewer items with high *b*-parameters.

Table 7.1 presents the sizes of three pools and the summary statistics for the item parameters within each pool. The size of either optimal item pools is larger than the size of the operational pool by 40 or more. The PM pool has the smallest *b*-parameter range and the largest average *a*-value with a maximum of 3.520 and minimum of 1.146. The range of *a*-parameter values for items in the MTI pool is from 1.275 to 3.394 and from 0.746 to 3.141 for the operational pool.

The overview of the evaluation results for these item pools are presented in Table 7.2. The average bias of the ability estimates is positive from the operational pool and the MTI pool and negative from the PM pool, but the magnitudes of the bias are negligible. Optimal item pools yield smaller MSE and higher correlation coefficient than the operational pool does. Of the item pools, the MTI pool results in the highest correlation coefficient and the smallest MSE.

Table 7.2 also shows that the MTI pool has the lowest test-retest overlap rate. The PM pool, however, has the highest overlap rate despite having the largest pool size. The plots of conditional test-retest overlap rate for item pools shown in Figure 7.2 reveal that the MTI pool has the lowest overlap rate at most ability levels.

Table 7.1 Item Pool Size and Item Parameter Statistics for Arithmetic Reasoning with *a*-Stratified Exposure Control

| Pool | Pool Size | *a* | | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. |
| OP | 137 | 1.556 | 0.487 | 3.141 | 0.746 | 0.115 | 1.170 | 2.343 | -2.625 | 0.186 | 0.063 | 0.328 | 0.038 |
| MTI | 183 | 1.814 | 0.310 | 3.394 | 1.275 | -0.214 | 1.335 | 2.302 | -2.563 | 0.207 | 0.069 | 0.424 | 0.049 |
| PM | 237 | 2.027 | 0.572 | 3.520 | 1.146 | -0.022 | 1.416 | 2.725 | -2.545 | 0.186 | 0.064 | 0.393 | 0.040 |

Table 7.2 Summary Statistics of the Performance of the Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0048 | 0.0123 | -0.0064 |
| MSE | 0.1100 | 0.0757 | 0.0936 |
| Correlation | 0.9420 | 0.9622 | 0.9523 |
| Skewness of item exposure rate | 34.9510 | 25.8286 | 102.8691 |
| Item overlap rate | 0.3650 | 0.2230 | 0.4973 |
| Pct of items with item exposure Rate> 1/3 | 10.22% | 3.28% | 6.33% |
| Pct of items with item exposure Rate<.02 | 27.74% | 26.23% | 56.12% |
| Pool Size | 137 | 183 | 237 |

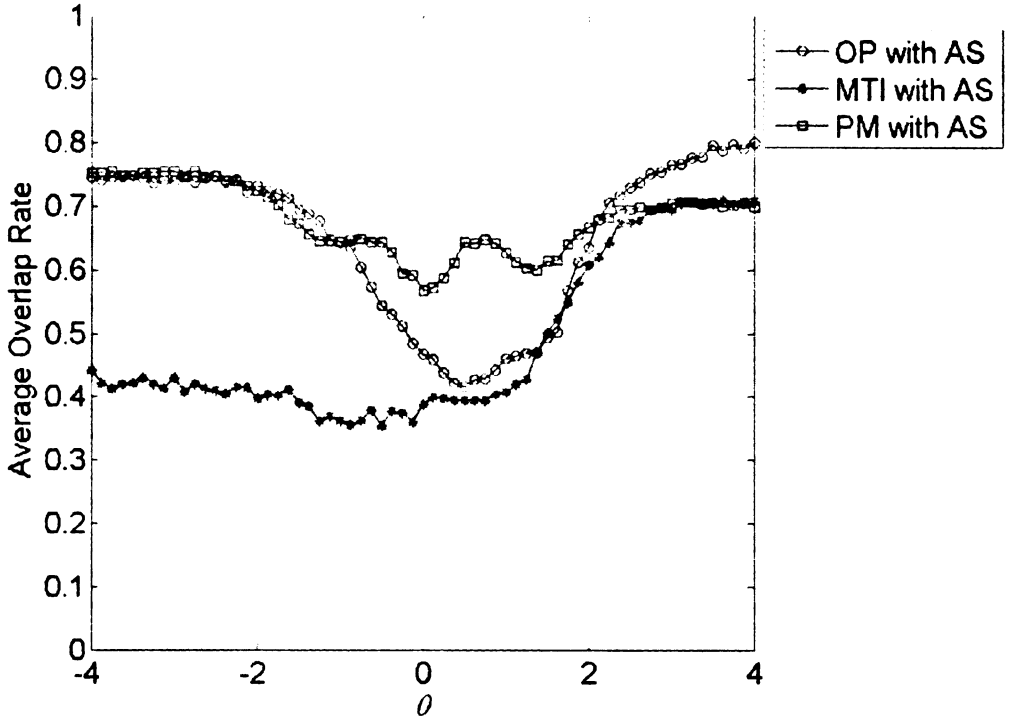


Figure 7.2 Test-retest overlap rate conditional on $\theta$

As can be seen in Table 7.2, the MTI pool has the lowest percentage of items with an exposure rate below .02 and the lowest percentage of items with an exposure rate higher than the target rate 1/3. The PM pool has a lower percentage of over-exposed items, but has over half of the items under-exposed. Individual item exposure rates are shown in Figure 7.3. It is clear that items in the MTI pool are used more evenly. By contrast, a few items in the PM item pool are used more often than most items, leading to a highly skewed item exposure rates for PM pool.

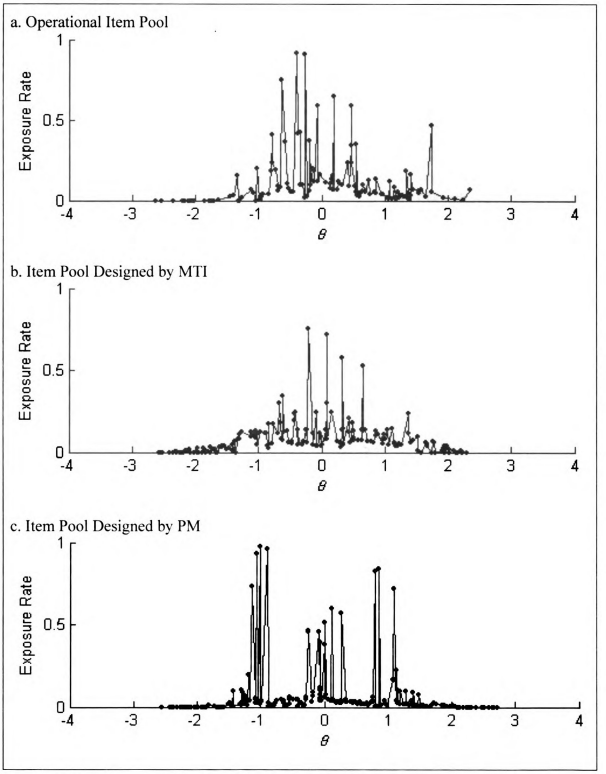


Figure 7.3 Item exposure rate by difficulty level

Figure 7.4 shows the average test information at the fixed ability levels. The plots for the PM pool and the operational pool look similar in shape, but the PM pool provides

more information at most ability levels. The MTI pool provides similar amount of information between ability levels of -1.5 to 1.5, and provides information exceeding 10 over the range between -2.0 to 2.0.

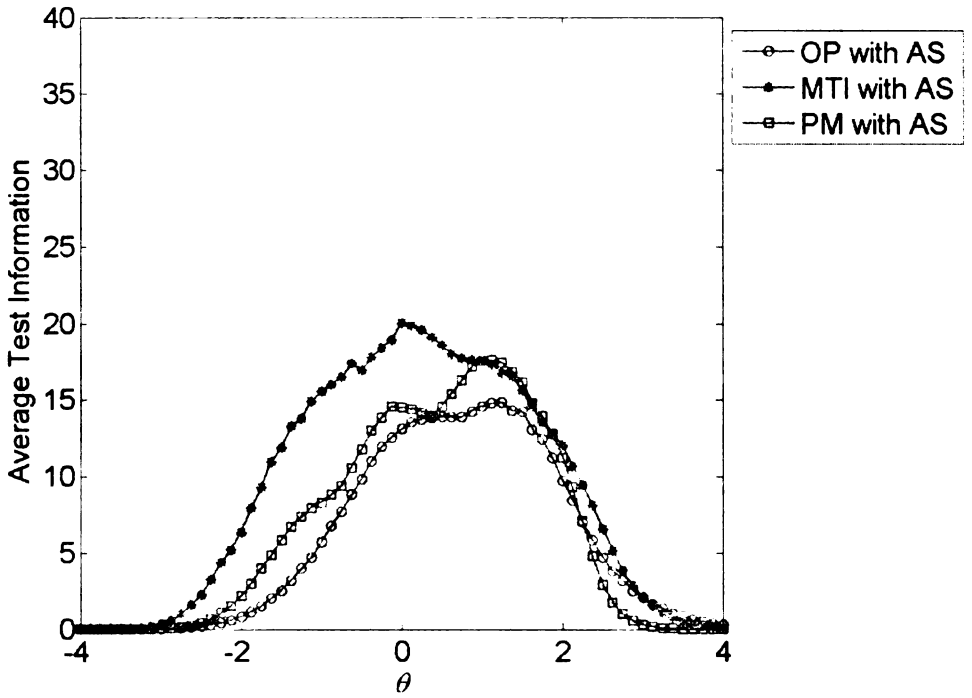


Figure 7.4 Average test information conditional on true $\theta$

Figure 7.5 to 7.7 present the CSEM, conditional bias, and CMSE for three item pools. The charts show that all three item pools yield similar SEM, bias, and MSE over the ability level between -2.0 to 2.0 and the results are mixed for ability levels beyond ±2.0.

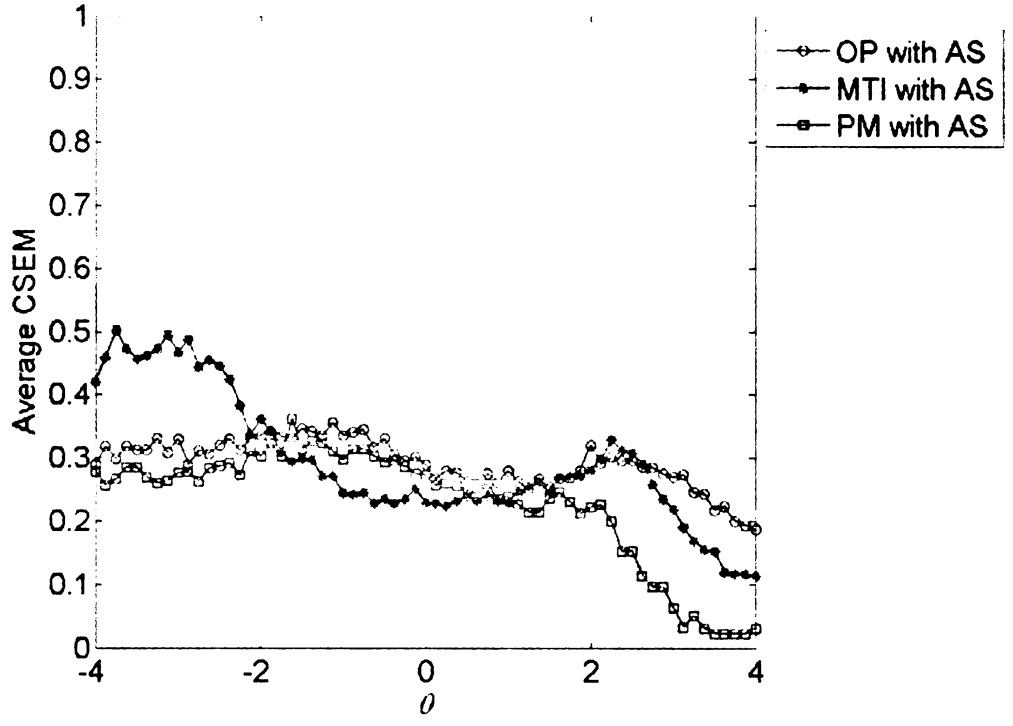


Figure 7.5 Conditional standard error of measurement (CSEM)

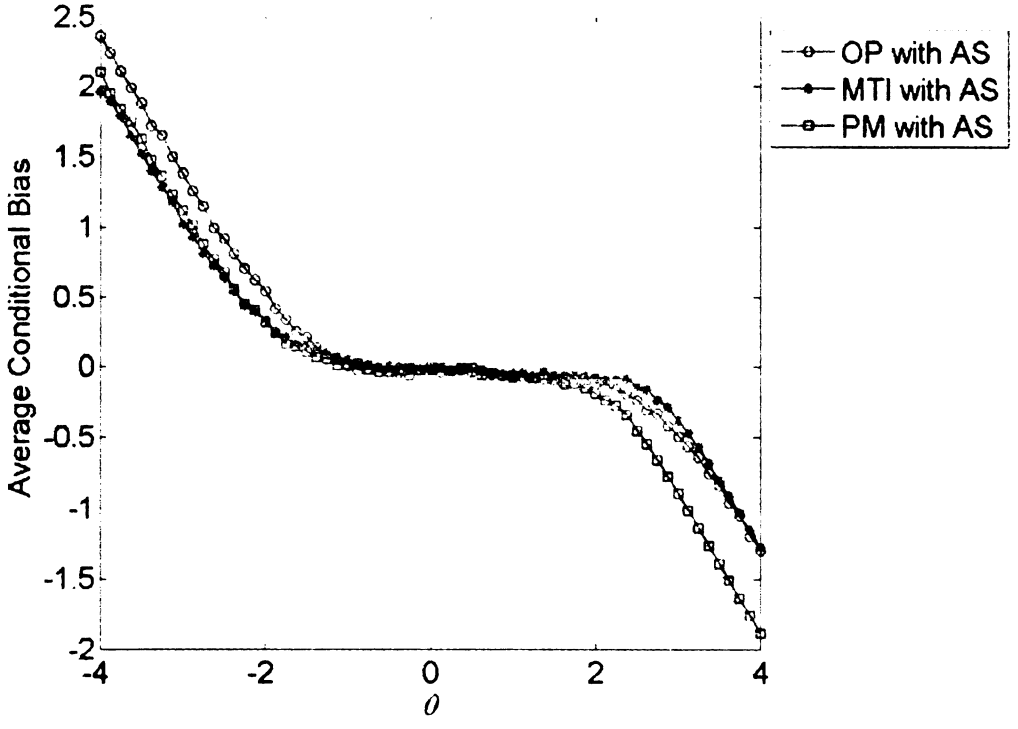


Figure 7.6 Conditional bias

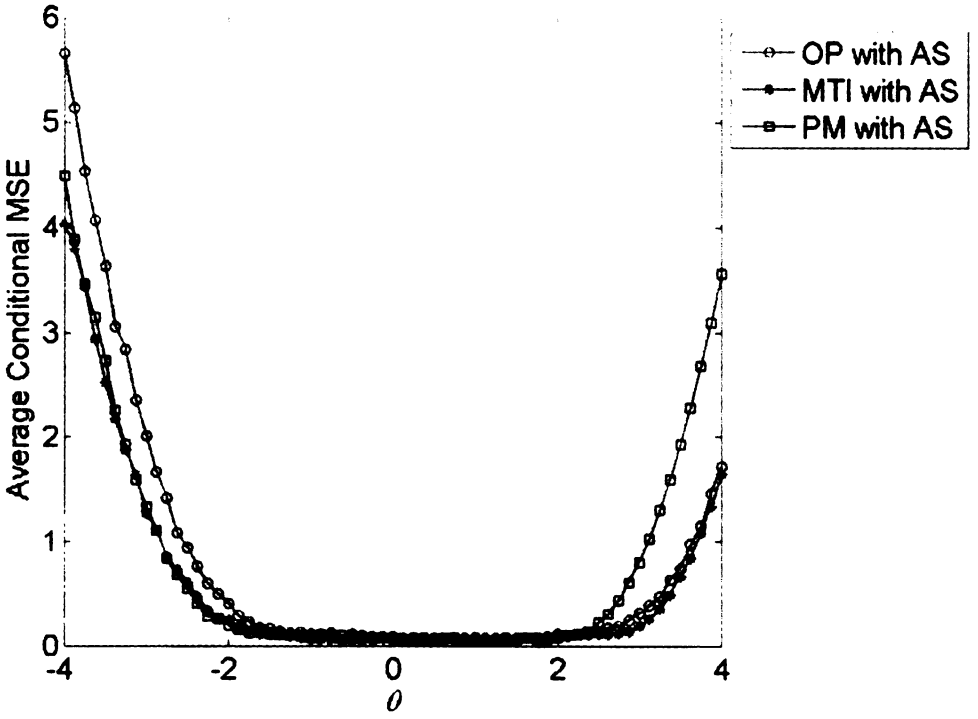


Figure 7.7 Conditional mean square error (CMSE)

### *7.2 Item Pools for Tests with Content Balance*

The item distributions for the OP, MTI, and PM pools are displayed in Figure 7.8. The results are similar to those without content balancing. The shape of the item distributions looks similar for the MTI pool and the PM pool. The PM pool has more highly discriminating and difficult items in both Content 1 and Content 2, while the MTI pool has more moderately difficult items with high discriminating parameters in all three contents. Unlike the operational pool, which has more items with both high *a*-parameters and high *b*-parameters, both optimal pools seem to require more items with moderate to low *b*-parameters and fewer items with high *b*-parameters.

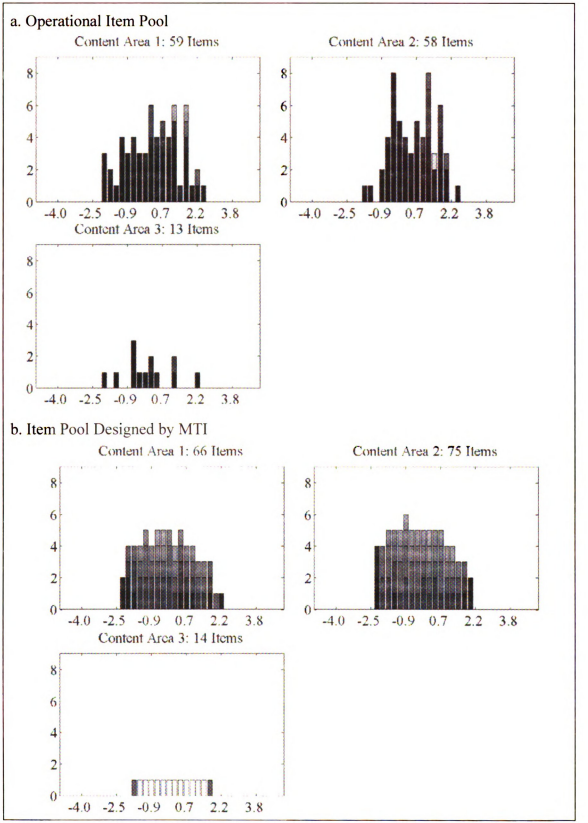


Figure 7.8 Item distribution for item pools with content balancing and without exposure control

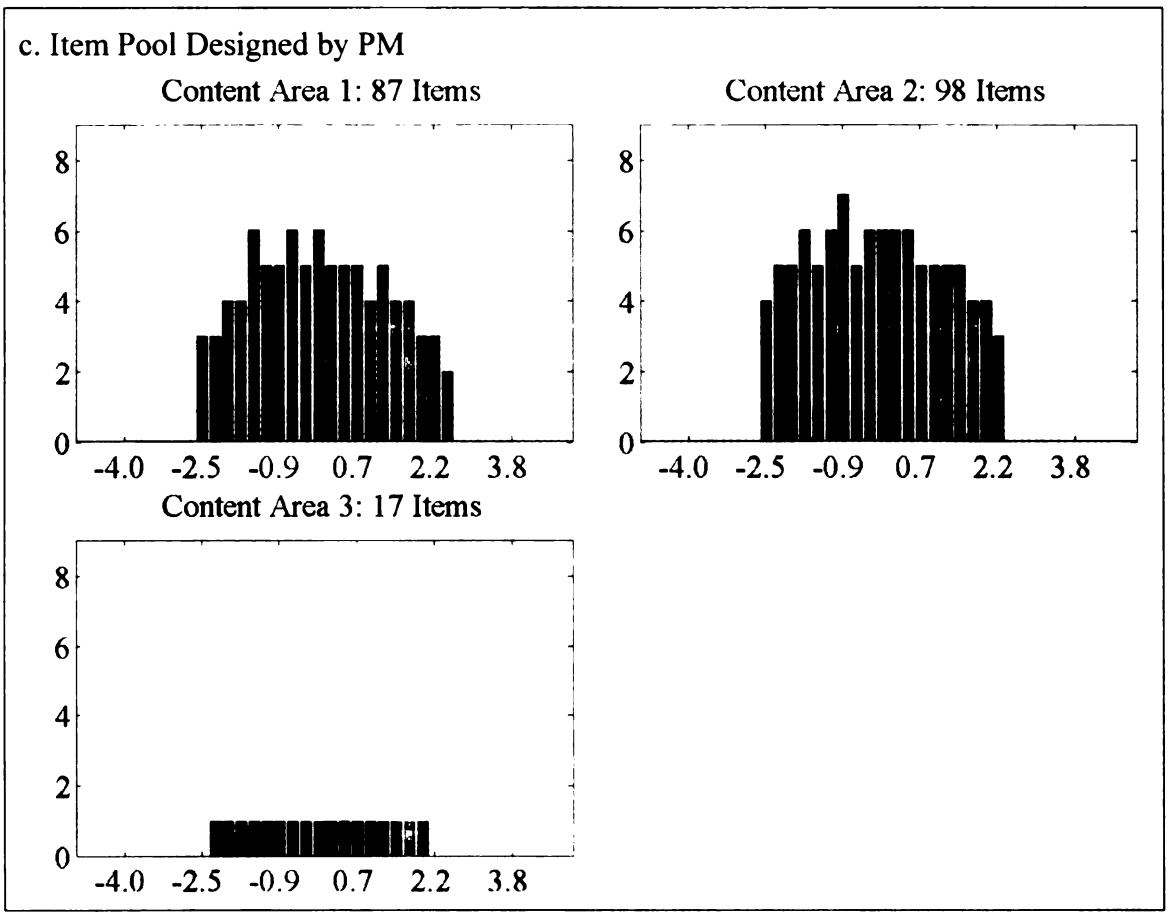


Figure 7.8 (Conti'd) Item distribution for item pools with content balancing and without exposure control

Table 7.3 presents the item pool sizes for the three pools and the summary statistics for item parameters within each pool. Results are similar to ones where no content balancing was present. The sizes for both optimal item pools are larger than the size of the operational pool by 40 or more. The MTI pool has the largest average *a*-parameter values with a maximum of 4.046 and minimum of 1.244. The operational pool has the smallest *b*-parameter range. All three item pools have similar numbers of items in Content 3. Interestingly, although Content 1 and 2 appear the same number of times in a test, the optimal pools require more items in Content 2 than Content 1.

Table 7.3. Item Pool Size and Item Parameter Statistics for General Science with *a*-Stratified Exposure Control

| Pool | Pool Size | *a* | | | | *b* | | | | *c* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. | Mean | SD | Max. | Min. |
| Content 1 | | | | | | | | | | | | | |
| OP | 59 | 1.167 | 0.354 | 2.314 | 0.628 | 0.099 | 1.189 | 2.211 | -2.211 | 0.224 | 0.067 | 0.396 | 0.091 |
| MTI | 78 | 1.645 | 0.200 | 2.154 | 1.283 | -0.158 | 1.220 | 2.238 | -2.302 | 0.250 | 0.070 | 0.432 | 0.110 |
| PM | 86 | 1.525 | 0.333 | 2.186 | 1.003 | -0.053 | 1.301 | 2.406 | -2.266 | 0.222 | 0.068 | 0.386 | 0.082 |
| Content 2 | | | | | | | | | | | | | |
| OP | 58 | 1.215 | 0.377 | 2.732 | 0.564 | 0.350 | 0.997 | 2.215 | -1.863 | 0.218 | 0.069 | 0.459 | 0.078 |
| MTI | 85 | 1.715 | 0.233 | 2.176 | 1.244 | -0.188 | 1.242 | 2.197 | -2.252 | 0.227 | 0.067 | 0.392 | 0.095 |
| PM | 100 | 1.516 | 0.330 | 2.129 | 0.981 | -0.117 | 1.350 | 2.511 | -2.534 | 0.233 | 0.066 | 0.405 | 0.101 |
| Content 3 | | | | | | | | | | | | | |
| OP | 13 | 1.087 | 0.282 | 1.776 | 0.726 | -0.195 | 1.128 | 1.917 | -2.137 | 0.220 | 0.058 | 0.305 | 0.115 |
| MTI | 14 | 2.906 | 0.804 | 4.046 | 1.759 | 0.037 | 1.062 | 1.683 | -1.585 | 0.233 | 0.076 | 0.402 | 0.134 |
| PM | 17 | 1.496 | 0.316 | 1.897 | 0.947 | -0.128 | 1.303 | 1.876 | -2.303 | 0.225 | 0.055 | 0.349 | 0.143 |

Table 7.4 Summary Statistics of the Performance of the Ideal Item Pools

| Statistic | OP | MTI | PM |
|---|---|---|---|
| Bias | 0.0062 | 0.0085 | -0.0092 |
| MSE | 0.1631 | 0.0898 | 0.1359 |
| Correlation | 0.9149 | 0.9544 | 0.9280 |
| Skewness of item exposure rate | 9.4318 | 7.7723 | 16.8795 |
| Item overlap rate | 0.3042 | 0.1865 | 0.3934 |
| Pct of items with item exposure Rate> 1/3 | 7.69% | 3.39% | 9.36% |
| Pct of items with item exposure Rate<.02 | 27.69% | 24.29% | 53.69% |
| Pool Size | 130 | 177 | 203 |



Table 7.5 Percentage of Over- and Under-Exposed Items by Content

| | Content 1 | | | Content 2 | | | Content 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Statistic | OP | MTI | PM | OP | MTI | PM | OP | MTI | PM |
| Pct of items with item exposure Rate> 1/3 | 5.08% | 2.56% | 9.30% | 8.62% | 2.35% | 9.00% | 15.38% | 14.29% | 11.76% |
| Pct of items with item exposure Rate<.02 | 20.34% | 19.23% | 46.51% | 24.14% | 20.00% | 56.00% | 76.92% | 78.57% | 76.47% |
| Pool Size | 59 | 78 | 86 | 58 | 85 | 100 | 13 | 14 | 17 |

The overview of the evaluation results for these item pools are presented in Table 7.4. The results are similar to those without content balancing. The PM pool shows slightly negative bias for the ability estimates. The average bias from the operational pool and the MTI pool are positive but the magnitudes of the bias are negligible. Optimal item pools yield smaller MSE and higher correlation coefficient than the operational pool does. Of all the pools, the MTI pool resulted in the highest correlation coefficient and the smallest MSE.

The plots of conditional test-retest overlap rate shown in Figure 7.2 draw a consistent picture as the plots for item pools without content balancing. The MTI pool has the lowest overlap rate at all ability levels. Summary statistics in Table 7.2 also shows that the MTI pool has the lowest average test-retest overlap rate. The PM pool, however, has the highest overlap rate despite having the largest pool size.

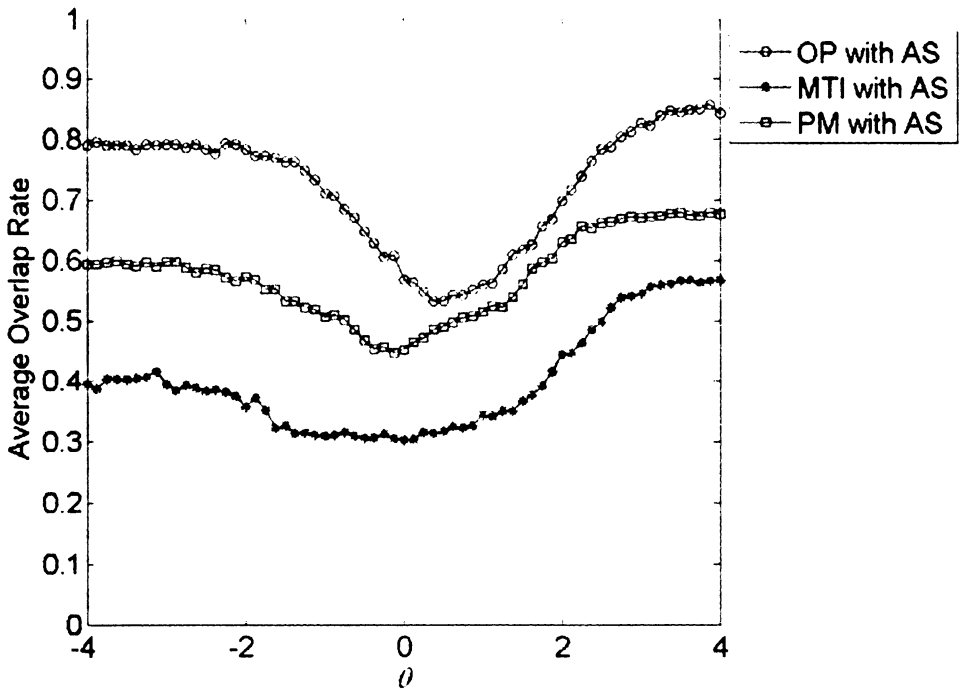


Figure 7.9 Test-retest overlap rate conditional on $\theta$

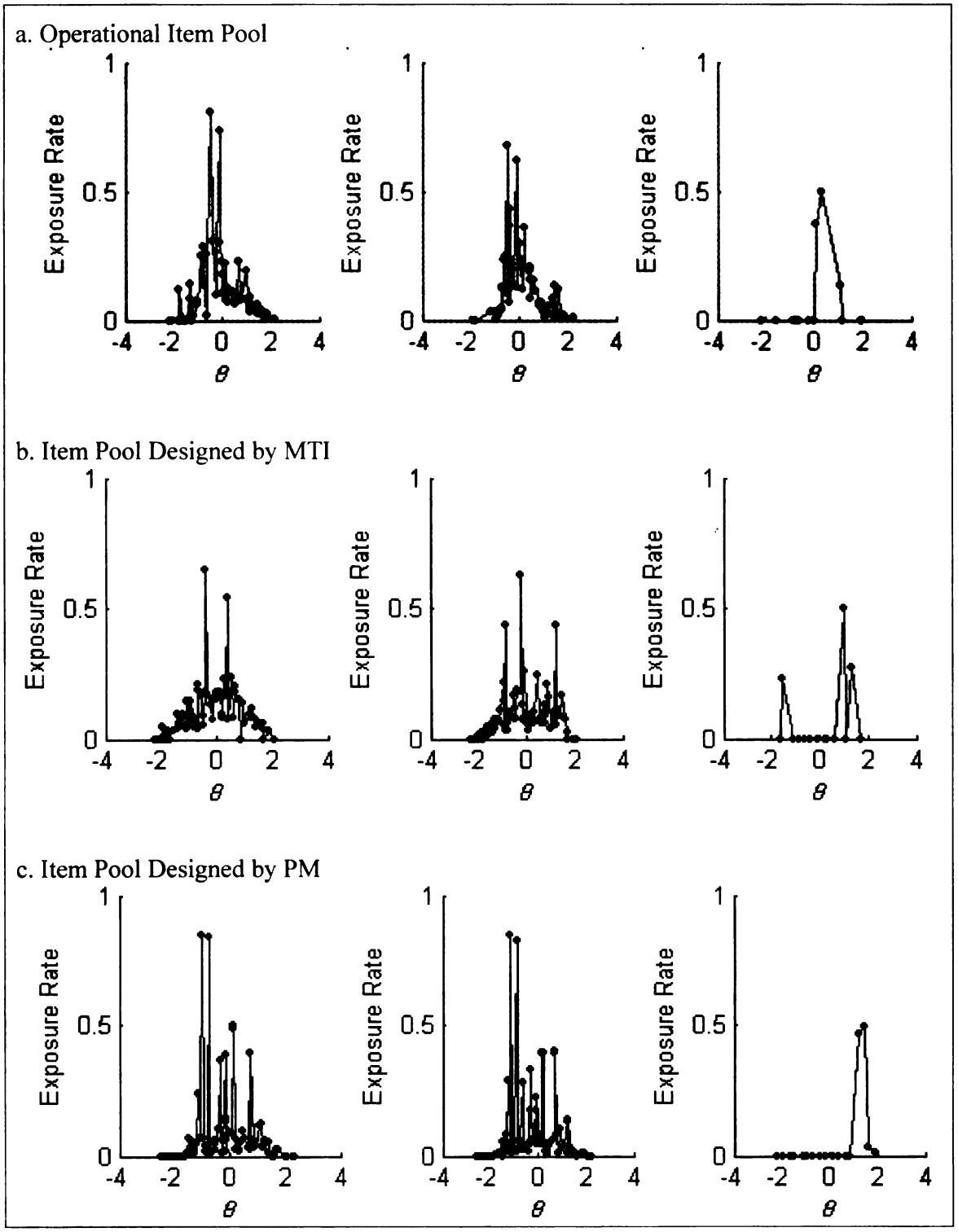


Figure 7.10 Item exposure rate by difficulty level

Table 7.2 also shows that the MTI pool has the lowest percentage of items with exposure rate below 0.02, the lowest percentage of items with exposure rate higher than

the target rate 1/3, and the lowest skewness of the item exposure rate. The PM pool has over half of the items under-exposed and is the most skewed pool in terms of item usage. Individual item exposure rates are shown in Figure 7.3. It is clear that items in the MTI pool are used more evenly and only a few items in the PM item pool are used more often while others are less used.

Figure 7.11 shows the average test information over the ability levels ranging from -4.0 to 4.0. The amount of information the operational pool provides seems quite low. Even at its peak levels, it is below 10. The plot for the PM item pool looks similar to that for the operational pool but the PM pool provides more information over a range of ability levels. The MTI pool provides a similar amount of information between ability levels -2.0 to 1.5, and provides information exceeding 10 at the range between -2.0 to 2.0.

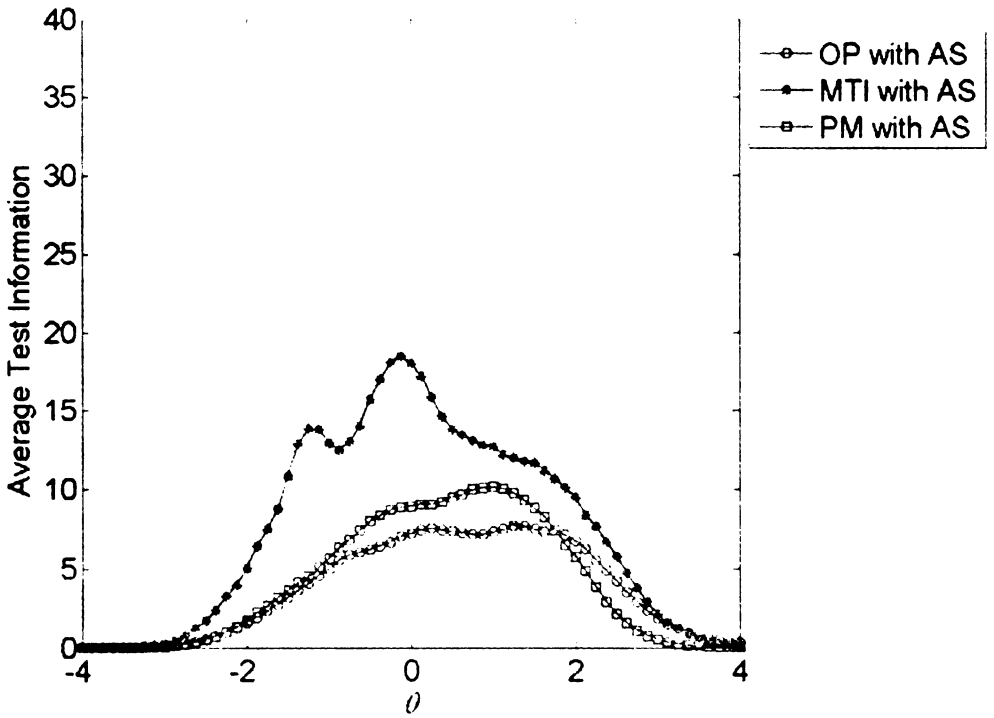


Figure 7.11 Average test information conditional on true $\theta$

Figure 7.12 to 7.14 present the CSEM, conditional bias, and CMSE for three item pools. The charts show that all three item pools yield similar SEM, bias, and MSE over the ability level between -2.0 to 2.0. The MTI pool performs better at ability level below -2.0 and over 2.0.

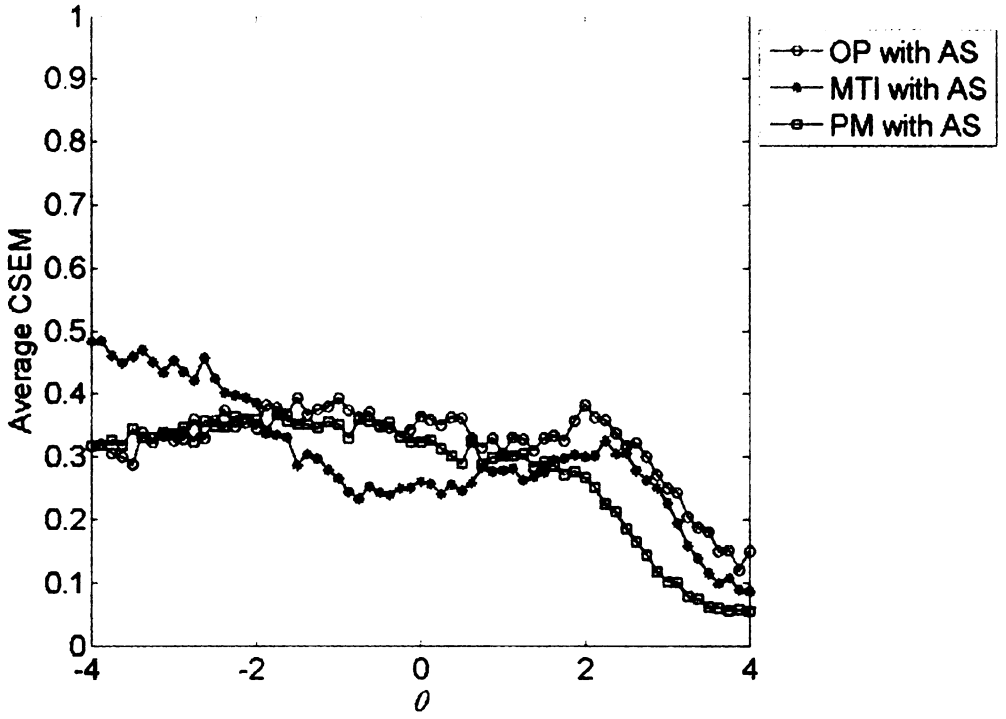


Figure 7.12 Conditional standard error of measurement (CSEM)

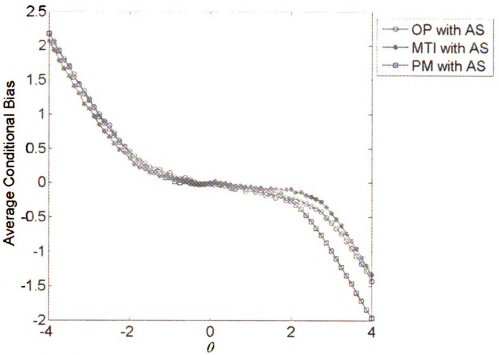


Figure 7.13 Conditional bias

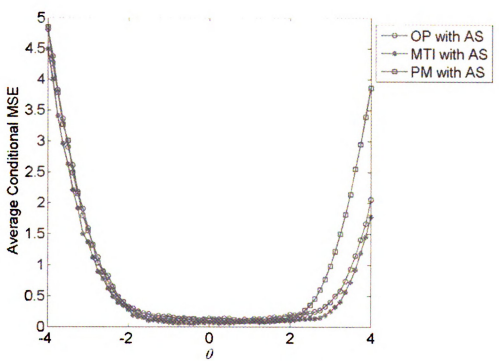


Figure 7.14 Conditional mean square error (CMSE)

### *7.3 Summary*

The results suggest the optimal pool designed with MTI performs the best, based on test security and measurement accuracy. It does require more items and some highly discriminating items. Tests assembled from MTI pools have smaller test-retest overlap rates and significantly lower percentages of over- and under-exposed items.

The item pool designed by the PM did not perform as well as the operational pool and the pool designed by MTI, despite having more items. One possible reason is that the way items are stratified is slightly different from the way items are simulated. It might yield better results if items are stratified with Chang and van der Linden's (2003) 0-1 linear programming method during the evaluation stage.

# Chapter VIII Discussion

This chapter first revisits the definition of "optimal" in item pool design and discusses how this study successfully addresses the criteria of optimality the definition implies. It then presents the implication of Reckase's method on the practice of item pool development and maintenance. Finally, the limitations of this study and the expected future research are discussed.

## *8.1 A Revisit to the Definition of "Optimal"*

This study investigated two approaches to design optimal blueprints for CAT item pools. Except for the item pool designed for *a*-stratified exposure control by the PM method, optimal pools designed by either method perform better than the operational pools no matter whether exposure control and content balancing are considered or not.

Optimal item pool design looks for the most desirable or favorable combination of items to form an item pool that would support the assembly of a large number of individualized computerized adaptive tests. There is, however, no single pool that is absolutely optimal, as it is constrained by a number of factors and different compositions of the items that may yield similar measurement precisions. That is why the two "optimal" pools look quite different and each is still optimal in some sense.

A general objective for an optimal item pool design is to meet the three criteria described by van der Linden (1999): 1) sufficiently large to allow several thousand overlapping subtests to be drawn from its items; 2) consisting of items spanning the entire range of item difficulty relative to the population of interest; and 3) consisting of an appropriate mix of high and low discriminating items to lower the item creation cost

while meeting the needs of test precision. It is not hard to meet the first criteria, where minimum size can be translated to the test length divided by the target exposure rate (1.0 if no exposure rate is considered). Simulation studies can easily realize the first two requirements by randomly sampling a large number of examinees from the expected examinee population, simulating a CAT administration to them, and tallying the number of items needed in different difficulty levels. All optimal item pools designed in this study are at least five times the test length and span a wide range of item difficulties. Acceptable precision can be achieved with difficulty ranges slightly smaller than the items in the operational pools in that the ranges of the item difficulty level for optimal pools are all smaller.

The third criterion, item creation cost, was not estimated directly in this study. However, mechanisms to indirectly approximate the item writing cost and the effort to minimize it were utilized in Reckase's item pool design method. Item generations with both MTI and PM methods imply the relationship between item parameters and the cost of item creations. The difference between PM and MTI can be primarily the difference between the assumptions. The MTI method assumes that highly discriminating items (i.e., items with high *a*-parameters) are more difficult to write and cost more to create, therefore, the MTI design tries to limit the number of high discrimination items by simulating items to meet the minimum test information requirement. The PM method assumes that the more the items with certain characteristics among the existing items, the less expensive to create items with the same characteristics. Therefore, it minimizes the item creation cost by modeling the item characteristics (i.e., the relationship between IRT parameters) and simulating items that are more like existing items.

Overall, the optimally designed item pools perform better than the operational pools obtained from CAT-ASVAB. The results show that the MTI design generally leads to smaller pools and contain items with lower *a*-parameters. The PM pools, on the other hand, maintain the correlation between item parameters but do not perform as well as the MTI pool.

The operational pools, on the other hand, provide more measurement precision over some range of latent ability levels. A closer look at the Arithmetic Reasoning pool finds more highly discriminating items at the range of *b*-parameters between 0 and 1.5. In practice, when operational item pools retire frequently, such high discriminating items may be difficult to replace. It introduces doubts on whether or not the same performance over similar ability levels can be easily duplicated. Item pools designed with Reckase's method have more items evenly distributed over a wider range of ability levels. As a result, optimal pools perform better than the operational ASVAB item pools at most latent ability levels.

The results imply that improvement may be made to the operational item pools by adjusting the item distributions to make them closer to what optimal item pool blueprints demand. For example, the arithmetic reasoning pool may perform better by adding more moderately discriminating items in the lower end of the latent ability levels and taking out some highly discriminating items in the higher end of the latent ability levels for future use.

### *8.2 Implications on the Practice of Item Pool Development*

This study is based on the assumption that examinees are normally distributed with a population mean ability of 0 and variance of 1. However, in reality, examinee

distributions are not always normal, and the expected distribution may not match the exact examinee distribution, which can only be decided when the tests are administered. The question raised is how robust the design is to the violation of the distributions. There are two situations, and, thus, two treatments are required. In the case where the expected distribution is not normal, it is possible to sample the examinees from a predefined examinee distribution, which can be constructed from previous test administrations. On the other hand, since it is a simulation study, violation to the assumptions may threaten the validity of the study and impact the results. The extent of the potential impacts could be a study of interest for future research.

The end product of the item pool design is a blueprint listing the number of needed items in each bin, that is, items with the *a*- and *b*- parameters in a certain range. Similar to the function of a test blueprint for the paper-and-pencil test, the item pool blueprint serves as a guide for item selection or item creation for the item pool. It portrays the optimal item composition an item pool should have and, therefore, is a target item developers should try to match. Items with desired content coverage and statistics can be either selected from previously written ones or created by item writers.

This method has not been tested in the practice. In this study, all items required by the design method are assumed available when comparing the optimally designed item pools to the operational pools. It seems hard to produce items with exactly the same item parameters required by the item pool design blueprint. However, with the advance in item modeling research, it will be more and more feasible to create large numbers of similar items with the desired psychometric properties. When the PM pool takes into account the correlations between *a*- and *b*-parameters, the blueprint designed by the PM

method may be easier to fulfill. The MTI pools achieve acceptable measurement precision with a minimum number of items, but it is uncertain how hard it would be to find or to create the proper items. On the other hand, improvement on the design method, such as combining the two methods to take advantages of the good features of each method, will make the design more practical. In addition, it should be pointed out that by defining the width of "bins", the blueprint requires similar items within a certain range instead of with exact item parameters. Future studies are needed to investigate how hard it is to fulfill the required items of the blueprint.

### *8.2 Implications on Item Pool Management*

In practice, operational item pools are not static. In most testing programs, tests are administered from the bank and new items are pretested on a continuous basis. Obsolete items are removed from time to time. Thus, monitoring item usage and replenishing new items are two important tasks of item pool management (van der Linden & Veldkamp, 2000). The item pool design methods presented here can easily be adapted for use in item pool management, both at the master pool level and at the operational pool level.

The master item pool is a union of operational item pools. The distribution of the optimal master pool could be simply a number of replications of the operational pool distribution. In other words, if the master pool will support ten smaller operational pools, the optimal item distribution of the master pool in each bin is simply ten times the item distribution in the optimal pool designed by the simulation method. Alternatively, the union method can take into consideration the expected exposure rates for the items in each bin, where the number of items needed in each bin for the master pool can be expressed as

$$X'_{AB}=Max(RX_{AB}r_{AB}, X_{AB})$$

where $R$ is the number of operational item pools a master item pool can support, and $r_{AB}$ is the expected exposure rate for the numbers in each bin. In this way, the master item pool has more items in the most exposed bins and fewer items in the least exposed bins.

### *8.3 Reckase's Method vesus the Mathematical Programming Method*

The results show that the extensions to Reckase's method work well in designing optimal item pools in situations where items are calibrated with 3PL. Compared to the mathematical programming method, Reckase's method simulates the CAT procedure straightforwardly and, therefore, is more flexible in adapting different item selection and ability estimate process and is easier to implement. Constraints on non-statistical attributes (e.g., content balancing) are absorbed into the first stage of the design by partitioning the target pool into smaller ones. There is no special software needed. The mathematical programming method is more mathematically structured by quantifying all the constraints and searching for the optimal solutions with linear programming, but it also requires the use of a "shadow test" item selection approach in CAT simulation. Reckase's method emphasizes the randomness of the item parameters in simulation, while the mathematical programming method focusing on optimizing predefined "pseudo" items. In the end, when they are all modeling the same CAT process, the simulation results should be similar.

While taking different approaches, Reckase's method and the mathematical programming method are similar in many ways. One of the important similarities is between the PM item simulation approach and the mathematical programming method in the way item costs are minimized in item pool design process. The mathematical

programming method defined a cost function, which is an inverse of the number of real items with certain combination of the attributes, including IRT parameters. It assumes that the more real items with the combination of item parameters, the less cost it is to create items with this item parameter combination. The idea is essentially the same as the PM method, in which the simulation would more likely generate items along the regression line of *b*-parameters on *a*-parameters where more real items are clustered.

Either method may be able to borrow some ideas from the other to improve the item pool design. No literature has described the design of item pools with *a*-stratified exposure control by the mathematical programming method. Chang and van der Linden described the 0-1 linear programming method to optimize the stratification of *a*-parameters for an existing item pool but not for the "pseudo items". As explored in this study, it may be possible to simulate the item pool design by varying the target information at different stages of the test.

### *8.4 Limitations and future studies*

Due to the limited resources, the prediction models in this study were based on one operational item pool. In practice, it is possible to use multiple recent item pools to get a more accurate estimation of the attributes of the items written for the testing programs.

Previous research showed that the bin width might influence the number of items required in the optimal pool. With post-simulation adjustment utilized in this study, item pools would trim unnecessary items in the bins, so the bin width might not influence the size of the final pool. However, future studies are needed to investigate the impact.

The optimal item pool blueprints designed for computerized adaptive testing with *a*-stratified exposure control appear to require more items than those designed for CAT

with Sympson-Hetter exposure control. One of the reasons is post-simulation adjustment was not applied due to different item selection procedure *a*-stratified exposure control uses. Future research is expected to explore the appropriate post-simulation adjustment for item pools with item selection procedure different from maximum information based ones.

While this study investigated the optimal item pool design with the PM and MTI methods separately, both methods have their shortcomings. MTI method tends to result in items with low correlations between their *a*- and *b*-parameters. The PM method, while maintaining similar correlation as the items in operational pools do, tends to perform better over some ability levels than others. It is important for future research to explore ways to combine the two design methods so the item generation would take into account the item parameter correlations while meeting the minimum information requirement.

# APPENDIX

# APPENDIX

Table A.1 Item Distribution for the Operational Item Pool – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -2.40 | -2.20 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -2.20 | -2.00 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -2.00 | -1.80 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.80 | -1.60 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.60 | -1.40 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.40 | -1.20 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.20 | -1.00 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.00 | -0.80 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.80 | -0.60 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.60 | -0.40 | 0 | 3 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| -0.40 | -0.20 | 0 | 1 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.20 | 0.00 | 0 | 1 | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 0.00 | 0.20 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.20 | 0.40 | 0 | 0 | 5 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 0.40 | 0.60 | 0 | 0 | 3 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.60 | 0.80 | 0 | 0 | 1 | 6 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 0.80 | 1.00 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 5 |
| 1.00 | 1.20 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 3 | 1 | 1 | 0 | 1 | 9 |
| 1.20 | 1.40 | 0 | 0 | 0 | 3 | 4 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 11 |
| 1.40 | 1.60 | 0 | 0 | 5 | 1 | 1 | 2 | 0 | 2 | 0 | 1 | 0 | 0 | 12 |
| 1.60 | 1.80 | 0 | 0 | 3 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.80 | 2.00 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.00 | 2.20 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.20 | 2.40 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 11 | 26 | 40 | 22 | 16 | 8 | 2 | 7 | 1 | 2 | 0 | 2 | 137 |

Table A.2 Item Distribution for Item Pool Designed by MTI Method and without Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.40 | -2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.20 | -2.00 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -2.00 | -1.80 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.80 | -1.60 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.60 | -1.40 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.40 | -1.20 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.20 | -1.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.00 | -0.80 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.80 | -0.60 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.60 | -0.40 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.40 | -0.20 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.20 | 0.00 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.00 | 0.20 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.20 | 0.40 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.40 | 0.60 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.60 | 0.80 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.80 | 1.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.00 | 1.20 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.20 | 1.40 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.40 | 1.60 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.60 | 1.80 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.80 | 2.00 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2.00 | 2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.20 | 2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 11 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 82 |

Table A.3 Item Distribution for Item Pool Designed by PM Method and without Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.40 | -2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.20 | -2.00 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -2.00 | -1.80 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.80 | -1.60 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.60 | -1.40 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.40 | -1.20 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.20 | -1.00 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.00 | -0.80 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.80 | -0.60 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.60 | -0.40 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.40 | -0.20 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.20 | 0.00 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.00 | 0.20 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 6 |
| 0.20 | 0.40 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 2 | 0 | 0 | 0 | 0 | 7 |
| 0.40 | 0.60 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 5 |
| 0.60 | 0.80 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 1 | 0 | 0 | 0 | 6 |
| 0.80 | 1.00 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 1 | 0 | 0 | 0 | 6 |
| 1.00 | 1.20 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 5 |
| 1.20 | 1.40 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 5 |
| 1.40 | 1.60 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.60 | 1.80 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 0 | 0 | 0 | 7 |
| 1.80 | 2.00 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.00 | 2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.20 | 2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 8 | 7 | 11 | 13 | 19 | 25 | 15 | 3 | 0 | 0 | 0 | 101 |

Table A.4 Item Distribution for Item Pool Simulated with MTI Method and with Sympson-Hetter Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.40 | -2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.20 | -2.00 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -2.00 | -1.80 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.80 | -1.60 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.60 | -1.40 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.40 | -1.20 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.20 | -1.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.00 | -0.80 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.80 | -0.60 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.60 | -0.40 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.40 | -0.20 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.20 | 0.00 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 0.00 | 0.20 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.20 | 0.40 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.40 | 0.60 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.60 | 0.80 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.80 | 1.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.00 | 1.20 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.20 | 1.40 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.40 | 1.60 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.60 | 1.80 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.80 | 2.00 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2.00 | 2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.20 | 2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 11 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 95 |

Table A.5 Item Distribution for Item Pool Simulated with PM Method and with Sympson-Hetter Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.40 | -2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.20 | -2.00 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -2.00 | -1.80 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.80 | -1.60 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.60 | -1.40 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.40 | -1.20 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.20 | -1.00 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.00 | -0.80 | 0 | 0 | 0 | 3 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| -0.80 | -0.60 | 0 | 0 | 0 | 0 | 3 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| -0.60 | -0.40 | 0 | 0 | 0 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.40 | -0.20 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.20 | 0.00 | 0 | 0 | 0 | 0 | 0 | 3 | 6 | 0 | 0 | 0 | 0 | 0 | 9 |
| 0.00 | 0.20 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 2 | 0 | 0 | 0 | 0 | 7 |
| 0.20 | 0.40 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 5 | 0 | 0 | 0 | 0 | 10 |
| 0.40 | 0.60 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 0 | 0 | 0 | 0 | 6 |
| 0.60 | 0.80 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 2 | 0 | 0 | 0 | 7 |
| 0.80 | 1.00 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 2 | 0 | 0 | 0 | 7 |
| 1.00 | 1.20 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 5 |
| 1.20 | 1.40 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 5 |
| 1.40 | 1.60 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.60 | 1.80 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 0 | 0 | 0 | 7 |
| 1.80 | 2.00 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.00 | 2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.20 | 2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 8 | 7 | 11 | 16 | 25 | 28 | 20 | 5 | 0 | 0 | 0 | 120 |

Table A.6 Item Distribution for Item Pool Simulated with MTI Method and with *a*-Stratified Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.40 | -2.20 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -2.20 | -2.00 | 0 | 0 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -2.00 | -1.80 | 0 | 0 | 0 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -1.80 | -1.60 | 0 | 0 | 1 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| -1.60 | -1.40 | 0 | 0 | 0 | 1 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 7 |
| -1.40 | -1.20 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 7 |
| -1.20 | -1.00 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 2 | 8 |
| -1.00 | -0.80 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 2 | 8 |
| -0.80 | -0.60 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 8 |
| -0.60 | -0.40 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 2 | 8 |
| -0.40 | -0.20 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 2 | 9 |
| -0.20 | 0.00 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 9 |
| 0.00 | 0.20 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 1 | 8 |
| 0.20 | 0.40 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 2 | 8 |
| 0.40 | 0.60 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 1 | 7 |
| 0.60 | 0.80 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 7 |
| 0.80 | 1.00 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 1 | 7 |
| 1.00 | 1.20 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.20 | 1.40 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.40 | 1.60 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.60 | 1.80 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.80 | 2.00 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2.00 | 2.20 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2.20 | 2.40 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 8 | 15 | 36 | 39 | 26 | 14 | 3 | 0 | 0 | 17 | 158 |

Table A.7 Item Distribution for Item Pool Simulated with PM method and with *a*-Stratified Exposure Control – Arithmetic Reasoning

| *b* \ *a* | | 0.00 | 0.89 | 1.26 | 1.55 | 1.79 | 2.00 | 2.19 | 2.37 | 2.53 | 2.68 | 2.83 | 2.97 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.89 | 1.26 | 1.55 | 1.79 | 2.00 | 2.19 | 2.37 | 2.53 | 2.68 | 2.83 | 2.97 | ∞ | Total |
| -∞ | -3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -3.00 | -2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.80 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.40 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -2.40 | -2.20 | 0 | 2 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| -2.20 | -2.00 | 0 | 2 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| -2.00 | -1.80 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| -1.80 | -1.60 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -1.60 | -1.40 | 0 | 0 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -1.40 | -1.20 | 0 | 0 | 6 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -1.20 | -1.00 | 0 | 0 | 3 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -1.00 | -0.80 | 0 | 0 | 1 | 8 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| -0.80 | -0.60 | 0 | 0 | 0 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -0.60 | -0.40 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -0.40 | -0.20 | 0 | 0 | 0 | 0 | 8 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| -0.20 | 0.00 | 0 | 0 | 0 | 0 | 8 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 0.00 | 0.20 | 0 | 0 | 0 | 0 | 3 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 10 |
| 0.20 | 0.40 | 0 | 0 | 0 | 0 | 1 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 10 |
| 0.40 | 0.60 | 0 | 0 | 0 | 0 | 0 | 3 | 5 | 2 | 0 | 0 | 0 | 0 | 10 |
| 0.60 | 0.80 | 0 | 0 | 0 | 0 | 0 | 2 | 6 | 2 | 0 | 0 | 0 | 0 | 10 |
| 0.80 | 1.00 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 4 | 1 | 0 | 0 | 0 | 10 |
| 1.00 | 1.20 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 2 | 0 | 0 | 0 | 9 |
| 1.20 | 1.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 2 | 0 | 0 | 9 |
| 1.40 | 1.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 9 |
| 1.60 | 1.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 2 | 0 | 8 |
| 1.80 | 2.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 2 | 8 |
| 2.00 | 2.20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 3 | 7 |
| 2.20 | 2.40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 6 |
| 2.40 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 5 |
| 2.60 | 2.80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 |
| 2.80 | 3.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.00 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 10 | 47 | 34 | 30 | 22 | 21 | 18 | 14 | 11 | 13 | 17 | 237 |

Table A.8 Item Distribution for the Operational Item Pool – General Science Content 1

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.82 | -1.56 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.56 | -1.30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.04 | -0.78 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.52 | -0.26 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.26 | 0.00 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.00 | 0.26 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.26 | 0.52 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.52 | 0.78 | 0 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.04 | 1.30 | 0 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.30 | 1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 | 1.82 | 0 | 3 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.82 | 2.08 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2.08 | 2.34 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.34 | 2.60 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 12 | 30 | 10 | 3 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 59 |

Ta

Table A.9 Item Distribution for the Operational Item Pool – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.82 | -1.56 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 | -1.30 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.04 | -0.78 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -0.78 | -0.52 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.52 | -0.26 | 3 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| -0.26 | 0.00 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.00 | 0.26 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.26 | 0.52 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.52 | 0.78 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.04 | 1.30 | 0 | 1 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 1.30 | 1.56 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 1.56 | 1.82 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.82 | 2.08 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 9 | 26 | 14 | 6 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 58 |

Table A.10 Item Distribution for the Operational Item Pool – General Science Content 3

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.82 | -1.56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.56 | -1.30 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.04 | -0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -0.78 | -0.52 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.52 | -0.26 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 | 0.00 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 | 0.26 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 0.26 | 0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 | 0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.78 | 1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.04 | 1.30 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.30 | 1.56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 | 2.34 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 3 | 7 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |

Table A.11 Item Distribution for the Optimal Item Pool Designed by MTI and without Exposure Control – General Science Content 1

| *b* \ *a* | | 0.00 | 0.89 | 1.26 | 1.55 | 1.79 | 2.00 | 2.19 | 2.37 | 2.53 | 2.68 | 2.83 | 2.97 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.89 | 1.26 | 1.55 | 1.79 | 2.00 | 2.19 | 2.37 | 2.53 | 2.68 | 2.83 | 2.97 | ∞ | |
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 | -1.30 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.30 | -1.04 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.04 | -0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.52 | -0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.26 | 0.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.00 | 0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.26 | 0.52 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.52 | 0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.30 | 1.56 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.56 | 1.82 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.82 | 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 6 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 46 |

Table A.12 Item Distribution for the Optimal Item Pool Designed by MTI and without Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 | -1.30 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.30 | -1.04 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.04 | -0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.52 | -0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.26 | 0.00 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.00 | 0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.26 | 0.52 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 0.52 | 0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.30 | 1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.56 | 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 | 2.08 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 7 | 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 |

Table A.13 Item Distribution for the Optimal Item Pool Designed by MTI and without Exposure Control – General Science Content 3

| *b* \ *a* | 0.00 – 0.89 | 0.89 – 1.26 | 1.26 – 1.55 | 1.55 – 1.79 | 1.79 – 2.00 | 2.00 – 2.19 | 2.19 – 2.37 | 2.37 – 2.53 | 2.53 – 2.68 | 2.68 – 2.83 | 2.83 – 2.97 | 2.97 – ∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ – -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 – -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 – -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 – -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 – -1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.82 – -1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 – -1.30 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 – -1.04 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.04 – -0.78 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.78 – -0.52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.52 – -0.26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 – 0.00 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 – 0.26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.26 – 0.52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 – 0.78 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.78 – 1.04 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 – 1.30 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 – 1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 – 1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.82 – 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 – 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 – 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 – 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 – ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 0 | 0 | 2 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |

Table A.14 Item Distribution for the Optimal Item Pool Designed by PM and without Exposure Control – General Science Content 1

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.82 | -1.56 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 | -1.30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.30 | -1.04 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.04 | -0.78 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -0.78 | -0.52 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.52 | -0.26 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -0.26 | 0.00 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.00 | 0.26 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 0.26 | 0.52 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.52 | 0.78 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.30 | 1.56 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.56 | 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 | 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 6 | 8 | 18 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 41 |

Table A.15 Item Distribution for the Optimal Item Pool Designed by PM and without Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.82 | -1.56 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.56 | -1.30 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.30 | -1.04 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.04 | -0.78 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.52 | -0.26 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -0.26 | 0.00 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.00 | 0.26 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.26 | 0.52 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.52 | 0.78 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.04 | 1.30 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.30 | 1.56 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.56 | 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 | 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 5 | 11 | 17 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 46 |

Table A.16 Item Distribution for the Optimal Item Pool Designed by PM and without Exposure Control – General Science Content 3

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.82 | -1.56 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 | -1.30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.04 | -0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -0.78 | -0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.52 | -0.26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 | 0.00 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 | 0.26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.26 | 0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 | 0.78 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.78 | 1.04 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 | 1.30 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 | 1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 1 | 2 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |

Table A.17 Item Distribution for the Optimal Item Pool Designed by MTI and with Sympson-Hetter Exposure Control – General Science Content 1

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 | -1.30 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.30 | -1.04 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.04 | -0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.52 | -0.26 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.26 | 0.00 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.00 | 0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.26 | 0.52 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.52 | 0.78 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.78 | 1.04 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.30 | 1.56 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.56 | 1.82 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.82 | 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 6 | 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 52 |

Table A.18 Item Distribution for the Optimal Item Pool Designed by MTI and with Sympson-Hetter Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 | -1.30 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.30 | -1.04 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.04 | -0.78 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.78 | -0.52 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.52 | -0.26 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -0.26 | 0.00 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.00 | 0.26 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.26 | 0.52 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.52 | 0.78 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.30 | 1.56 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.56 | 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 | 2.08 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 7 | 47 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 54 |

Table A.19 Item Distribution for the Optimal Item Pool Designed by MTI and with Sympson-Hetter Exposure Control – General Science Content 3

| *b* \ *a* | 0.00 – 0.89 | 0.89 – 1.26 | 1.26 – 1.55 | 1.55 – 1.79 | 1.79 – 2.00 | 2.00 – 2.19 | 2.19 – 2.37 | 2.37 – 2.53 | 2.53 – 2.68 | 2.68 – 2.83 | 2.83 – 2.97 | 2.97 – ∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ – -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 – -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 – -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 – -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 – -1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.82 – -1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 – -1.30 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 – -1.04 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.04 – -0.78 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.78 – -0.52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.52 – -0.26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 – 0.00 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 – 0.26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.26 – 0.52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 – 0.78 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.78 – 1.04 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 – 1.30 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 – 1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 – 1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.82 – 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 – 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 – 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 – 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 – ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 0 | 0 | 2 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |

Table A.20 Item Distribution for the Optimal Item Pool Designed by PM and with Sympson-Hetter Exposure Control – General Science Content 1

| *b* \ *a* | 0.00 – 0.89 | 0.89 – 1.26 | 1.26 – 1.55 | 1.55 – 1.79 | 1.79 – 2.00 | 2.00 – 2.19 | 2.19 – 2.37 | 2.37 – 2.53 | 2.53 – 2.68 | 2.68 – 2.83 | 2.83 – 2.97 | 2.97 – ∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ – -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 – -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 – -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 – -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 – -1.82 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.82 – -1.56 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.56 – -1.30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.30 – -1.04 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.04 – -0.78 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -0.78 – -0.52 | 0 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.52 – -0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.26 – 0.00 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| 0.00 – 0.26 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 0.26 – 0.52 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.52 – 0.78 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.78 – 1.04 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.04 – 1.30 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.30 – 1.56 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.56 – 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 – 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 – 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 – 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 – 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 – ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 0 | 6 | 9 | 23 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 51 |

Table A.21 Item Distribution for the Optimal Item Pool Designed by PM and with Sympson-Hetter Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -1.82 | -1.56 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.56 | -1.30 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.30 | -1.04 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| -1.04 | -0.78 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.78 | -0.52 | 0 | 0 | 2 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| -0.52 | -0.26 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -0.26 | 0.00 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 0.00 | 0.26 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| 0.26 | 0.52 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.52 | 0.78 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0.78 | 1.04 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.04 | 1.30 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.30 | 1.56 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.56 | 1.82 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1.82 | 2.08 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 5 | 11 | 21 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 54 |

Table A.22 Item Distribution for the Optimal Item Pool Designed by PM and with Sympson-Hetter Exposure Control – General Science Content 3

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.82 | -1.56 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 | -1.30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.04 | -0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -0.78 | -0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.52 | -0.26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 | 0.00 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 | 0.26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.26 | 0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 | 0.78 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.78 | 1.04 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 | 1.30 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 | 1.56 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 1 | 2 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |

Table A.23 Item Distribution for the Optimal Item Pool Designed by MTI and with *a*-Stratified Exposure Control – General Science Content 1

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| -2.08 | -1.82 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.82 | -1.56 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.56 | -1.30 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.30 | -1.04 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.04 | -0.78 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.78 | -0.52 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.52 | -0.26 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.26 | 0.00 | 0 | 0 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.00 | 0.26 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.26 | 0.52 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.52 | 0.78 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.30 | 1.56 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.56 | 1.82 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 1.82 | 2.08 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 21 | 35 | 21 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 78 |

Table A.24 Item Distribution for the Optimal Item Pool Designed by MTI and with *a*-Stratified Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -2.08 | -1.82 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -1.56 | -1.30 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.30 | -1.04 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.04 | -0.78 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.78 | -0.52 | 0 | 0 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.52 | -0.26 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.26 | 0.00 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.00 | 0.26 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.26 | 0.52 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.52 | 0.78 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.04 | 1.30 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.30 | 1.56 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.56 | 1.82 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.82 | 2.08 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 1 | 16 | 35 | 21 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 85 |

Table A.25 Item Distribution for the Optimal Item Pool Designed by MTI and with *a*-Stratified Exposure Control – General Science Content 3

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.08 | -1.82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1.82 | -1.56 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 | -1.30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| -1.30 | -1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| -1.04 | -0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| -0.78 | -0.52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| -0.52 | -0.26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| -0.26 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0.00 | 0.26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0.26 | 0.52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0.52 | 0.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0.78 | 1.04 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 | 1.30 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 | 1.56 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 | 1.82 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | 14 |

Table A.26 Item Distribution for the Optimal Item Pool Designed by PM and with *a*-Stratified Exposure Control – General Science Content 1

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -2.08 | -1.82 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.82 | -1.56 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.56 | -1.30 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -1.30 | -1.04 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.04 | -0.78 | 0 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.78 | -0.52 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.52 | -0.26 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -0.26 | 0.00 | 0 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.00 | 0.26 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.26 | 0.52 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.52 | 0.78 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.04 | 1.30 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.30 | 1.56 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 21 | 23 | 18 | 14 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 86 |

Table A.27 Item Distribution for the Optimal Item Pool Designed by PM and with *a*-Stratified Exposure Control – General Science Content 2

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -2.34 | -2.08 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -2.08 | -1.82 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| -1.82 | -1.56 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -1.56 | -1.30 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| -1.30 | -1.04 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -1.04 | -0.78 | 0 | 1 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| -0.78 | -0.52 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.52 | -0.26 | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| -0.26 | 0.00 | 0 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.00 | 0.26 | 0 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.26 | 0.52 | 0 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 0.52 | 0.78 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 0.78 | 1.04 | 0 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 1.04 | 1.30 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.30 | 1.56 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 25 | 29 | 20 | 13 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

Table A.28 Item Distribution for the Optimal Item Pool Designed by PM and with *a*-Stratified Exposure Control – General Science Content 3

| *b* \ *a* | | 0.00<br>0.89 | 0.89<br>1.26 | 1.26<br>1.55 | 1.55<br>1.79 | 1.79<br>2.00 | 2.00<br>2.19 | 2.19<br>2.37 | 2.37<br>2.53 | 2.53<br>2.68 | 2.68<br>2.83 | 2.83<br>2.97 | 2.97<br>∞ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -∞ | -2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.86 | -2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.60 | -2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -2.34 | -2.08 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -2.08 | -1.82 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.82 | -1.56 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.56 | -1.30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.30 | -1.04 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -1.04 | -0.78 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.78 | -0.52 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.52 | -0.26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| -0.26 | 0.00 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.00 | 0.26 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.26 | 0.52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.52 | 0.78 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0.78 | 1.04 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.04 | 1.30 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.30 | 1.56 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.56 | 1.82 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1.82 | 2.08 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2.08 | 2.34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.34 | 2.60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.60 | 2.86 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.86 | ∞ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | 0 | 4 | 4 | 4 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 |

# REFERENCES

# REFERENCES


Ariel, A., Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41*(4), 345-360.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-359). Hillsdale, NJ: Erlbaum.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS RR-96 13). Princeton, NJ: ETS.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3). http://www.jtla.org.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*(2), 129 137.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191-244.

Boekkooi-Timminga, E. (1991). *A method for designing Rasch Model-based item banks.* Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (No. 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Chang, H. H. (2004). Understanding computerized adaptive testing: From Robins-Monro to Lord and beyond. In David Kaplan (Ed.) *The Sage handbook of quantitative methodology for the social sciences.* (pp. 117-136). Thousand Oaks, CA: Sage Publications, Inc.

Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). *Performance of item exposure control methods in computerized adaptive testing: Further explorations.* Paper presented at the Annual Meeting of the American Educational Research Association.

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

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37-52.

Chang, H. H., & van der Linden, W. J. (2003). Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*(4), 262-274.

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

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

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing* (No. ACT-RR-99-5): American College Testing Program, Iowa City, IA.

Davey, T., & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Davey, T., & Thomas, L. (1996). *Constructing adaptive tests to parallel conventional programs*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Davis, L. L. (2002). Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items. Unpublished doctoral dissertation, University of Texas, Austin.

Davis, L. L., & Dodd, B. G. (2001). *An examination of testlet scoring and item exposure constraints in the verbal reasoning section of the MCAT.* MCAT Monograph Series: Association of American Medical Colleges.

Deane, P., Graf, E. A., Higgins, D., Futagi, Y., Lawless, R. (2006). *Model analysis and model creation: Capturing the task-model structure of quantitative item domains* (No. ETS-RR-06-11). Educational Testing Service, Princeton, NJ.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement. 23,*355-368.

Embretson, S. E. (2001). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum Publishers.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1976). *Computer methods for mathematical computations*, Prentice-Hall, 1976.

Graf, E. A., Peterson, S., Steffen, M., Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models* (No. ETS-RR-05-25). Educational Testing Service, Princeton, NJ.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

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

Hau, K. T., Wen, J. B., & Chang, H. H. (2002). *Optimum number of strata in the a-stratified computerized adaptive testing design.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

Hetter, R. D., & Sympson J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Grist, S., Rudner, L., & Wise, L. (1989). *Computer adaptive tests* (ERIC Digest 107). Washington, DC: American Institute for Research and ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC No. ED 315 425)

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement items. *Journal of Educational Measurement, 5*, 275-290.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 192-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.

Luecht, R. M. (1998). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the the annual meeting of the National Council on Measurement in Education.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

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

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223-226). New York: Academic Press.

MathWorks. (2005). MATLAB: The language of technical computing, Version 7, Release 14, Student Version [Computer software]. Natick, MA: Author

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*(4), 315-330.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287-304.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurements, 28,* 95-104.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (RB-69-92). Princeton, NJ: Educational Testing Service

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*(350), 351-356.

Parshall, C., Davey, T., & Nering, M. (1998). *Test development exposure control for adaptive tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego CA.

Patsula, L. N., & Steffan, M. (1997). *Maintaining item and test security in a CAT environment: A simulation study.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Prosser, F. (1974). Item banking. In G. Lippey (Ed.), *Computer-assisted test construction* (pp. 29-66). Englewood Cliffs, NJ: Educational Technology Publications.

Reckase, M. D. (1974). *An application of the Rasch simple logistic model to tailored testing.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

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

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests.* Paper presented at the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D. & He, W. (2004). *The ideal item pool for the NCLEX-RN examination--Report to NCSBN*: Michigan State University, East Lansing, MI.

Reckase, M. D. & He, W (2005). *Ideal item pool design for the NCLEX-RN exam.* Michigan State University, East Lansing, MI.

Rudner, L. (1998). Item banking. *Practical Assessment, Research & Evaluation, 6*(4).

Segall, D. O., Moreno, K. E., & Hetter, R. D. (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.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine (Ed.), *Item generation for test development* (pp. 361 384). Mahwah, NJ: Lawrence Erlbaum Associates.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (No. ETS-RR-94-5): Educational Testing Service, Princeton, NJ.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. v. d. Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Netherlands: Kluwer Academic Publishers.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., & Lewis, C. (1995). *A new method for controlling item exposure in computer adaptive testing (Research Report 95-25)*. Princeton, NJ: Educational Testing Service.

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 atapitve testing.* Paper presented at the 27th Annual Meeting of the Military Testing Association, San Diego, CA.

Thomasson, G. L. (1995). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the Annual Meeting of the Psychometric Soceity, Minneapolis, MN.

Thissen, D., & Mislevy, R. J. (2000). Testing Algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

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.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*, 195-210.

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

Veldkamp, B. P., & van der Linden, W. J. (1999). *Designing item pools for computerized adaptive testing. Research report 99-03*: Twente Univ , Enschede (Netherlands) Faculty of Educational Science and Technology.

Wainer, H. (2000). Rescuing Computerized Testing by Breaking Zipf's Law. *Journal of Educational and Behavioral Statistics, 25*(2), 203-224.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 271-299). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1990). Computerized adaptive testing: A primer. In. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24-35). Washington, DC: United States Civil Service Commission.

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

Wetzel, C.D., & McBride, J. R. (1986). *Reducing the predictability of adaptive item sequences*. Paper presented at the annual conference of the Military Testing Association, San Diego, CA.

Yao, T. (1991). CAT with a poorly calibrated item bank. *Rasch Measurement Transactions 5:2*, 141.