Graduate Management Admission Council®

Designing Optimal Item Pools for Computerized Adaptive Tests with Sympson-Hetter Exposure Control

Lixiong Gu

Educational Testing Service

Mark D. Reckase

Michigan State University

Presented at the Item Exposure Paper Session, June 7, 2007

Abstract

Computerized adaptive testing requires a well-designed item pool containing an appropriate number of items to build individualized tests that match the examinees' ability levels. An optimal item pool should also contain well-balanced items that will achieve optimal item usage and lower the cost of item creation. One of the methods for designing the blueprint for an item pool is Reckase's method (2003), which is a Monte Carlo method to determine the properties of an optimal item pool. This study extended the method for designing item pools calibrated with the three-parameter logistic model and applied it to situations where the Sympson-Hetter procedure is used to control the item exposure rate. The procedures for designing the item pool and two approaches for simulating test items are presented. The performance of simulated item pools are evaluated along with an operational item pool.

Acknowledgments

This research was based on the first author's dissertation work completed in Michigan State University. The statements made and opinions expressed are solely the responsibility of the authors and not necessarily those of Educational Testing Service. Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Gu, L. & Reckase, M.D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Mark D. Reckase, 461 Erickson Hall, Michigan State University,
East Lansing MI 48824-1034, U.S.A. Email: reckase@msu.edu

Designing Optimal Item Pools for Computerized Adaptive Tests With Sympson-Hetter Exposure Control

Item pools play an important role in computerized adaptive testing (CAT). Items in the pool are indexed, structured, or otherwise assigned information that can be used to facilitate their selection for a test. Item pools have been called "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). For conventional paper-and-pencil tests, a well-designed item pool provides test developers a convenient yet powerful tool to produce high quality tests. The concept of an item pool is expanded in CAT. 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 for the testing application. Another kind is the operational item pool, which is a smaller subset of the master pool, and by design it has to be small enough so that the computer can easily retrieve the items and, when necessary, 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 can be limited. After a certain number of uses, items might need to be retired and put back into the master pool. Some items might be able to be reused 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 can be a problem in situations where learners have access to the item pool. Larger pools also mean more that items that match content, item format, and statistical requirements are available (Millman & Arter, 1984). The caveats, however, are: (1) the items added to the pool should be well written, content valid, and statistically fit; and (2) the total number of items should be manageable and easily retrievable.

An often-overlooked issue in item pool design is how to design and develop item pools in a more systematic and empirical manner, constructing 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.

The item writing process is usually guided by appropriately designed test specifications that outline the content attributes and their distributions. Requirements for statistical attributes, such as the range of difficulty, might 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 specifications 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 CAT. 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 the composition of an existing item bank. Subsequently, several methods for the construction of rotating item pools have been demonstrated in empirical studies, some of which achieved the design goal with integer programming methods (for a review of these methods, see Ariel, Veldkamp, & van der Linden, 2004).

Veldkamp and van der Linden (1999) described five steps to design an optimal blueprint for a CAT item pool with a mathematical programming method:

1. A set of specifications for the CAT is analyzed and all item attributes figuring in the specifications are identified.
2. Using the specifications, an integer programming model for the assembly of the shadow tests in the CAT simulation is formulated.
3. The population of examinees is identified and an estimate of its ability distribution is obtained; for example, from historical data.
4. A CAT simulation is carried out using the integer programming model for the shadow tests and sampling simulees from the ability distribution. Counts are collected of the number of times items from the cells in the classification table are used.
5. The blueprint is calculated from these counts, adjusting them to obtain optimal projections of the item exposure rates.

The advantage of this 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 this mathematical programming method is closely tied with the shadow test procedure in item selection 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 mathematical programming. This approach does not assume pre-existing items. Instead, items are simulated (in terms of IRT parameters) to match the current ability estimates to provide sufficiently optimal information. Reckase's method first partitions the target item pool into smaller pools based on different non-statistical attributes, such as content. Then the CAT process is simulated to construct the small item pools simultaneously. The simulation starts with an examinee randomly drawn from the expected examinee distribution to receive the CAT. 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.

Research Questions

The present study reports the development of optimal item pools for CATs using two different strategies. A modified version of Reckase's method is applied to designing optimal item pools calibrated with the three-parameter logistic model. Sympson-Hetter exposure control methods are investigated (Sympson & Hetter, 1985).

For the purpose of this research, it was desired to have an operational pool of items measuring an empirically significant dimension of ability. An operational item pool for a 15-item section of a large-scale aptitude test was 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 distributions. The simulation study took into consideration the distribution of the examinee population and the expected precision of ability estimates.

The following research questions were investigated in this study:

1. What does the optimal item pool designed for a CAT look like when the item selection procedure imposes no exposure control or when it incorporates the Sympson-Hetter method?
2. Do optimal item pools designed by a Monte Carlo simulation perform better than the real operational item pool in terms of empirical criteria?

Conceptual Framework

Reckase (1989) listed four major components of a CAT: 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 constrain the item selection so that items are selected not only by their statistical characteristics but also by 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 procedures, and target item exposure and overlap rates (Bergstrom & Lunz, 1999).

Item Pool

The adaptive feature of CAT makes it unnecessary to use pre-designed test forms as are used in paper-and-pencil tests. Rather, it requires an item pool from which all tests will be drawn. An item pool is not only a reservoir of items, but also an organized collection 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.

Owen's Bayesian Ability Estimation Procedure

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 CAT 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 item, the mean and variance are updated using a statistical procedure that combines the information in the prior distribution with the observed score (correct or incorrect) on the most recent test item, and the parameters of that item'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 item. 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. Owen's equation for updating the prior mean is as follows:

$$\mu(\theta_i \mid u_i) = \frac{\int \theta P(u_i \mid \theta) h(\theta) d\theta}{\int P(u_i \mid \theta) h(\theta) d\theta} \tag{1}$$

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

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

$$\hat{\sigma}_i^2 = \hat{\sigma}_{i-1}^2 - \left(\hat{\sigma}_{i-1}^2\right)^2 \left(a_i^{-2} + \hat{\sigma}_{i-1}^2\right)^{-1} \lambda_n, \tag{3}$$

where

u_i is the item response (answer is correct when $u_i = 1$ and incorrect 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,

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

$B_i = \left(b_i - \hat{\theta}_i\right) \Big/ \left(a_i^{-2} + \hat{\sigma}_{i-1}^2\right)^{-1/2}$, 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 item. 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 an incorrect prior can result in the need to administer more items to recover, and a regression toward the mean in ability estimation tends to occur.

Maximum Information 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 maximum information (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):

$$I_j(\theta) = \left[\frac{\partial P_j(\theta)}{\partial \theta}\right]^2 \Big/ P_j(\theta)Q_j(\theta) = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)Q_j(\theta)} \tag{4}$$

where $P_j(\theta)$ is the probability of a correct response, given θ, and $Q_j(\theta)$ is the probability of an incorrect response. Substituting the model specification in Equation 4, it can be simplified for the dichotomous three-parameter logistic item response model (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980) as:

$$I_j(\theta) = \frac{D^2 a_j^2 (1-c_j)}{(c_j + e^{DL_j})(1+e^{-DL_j})^2} \tag{5}$$

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 θ examinee correctly answering the item). Equation 5 indicates that the item information increases as b_j approaches θ, as a_j increases, and as c_j approaches 0 (Hambleton et al., 1991).

Unconstrained MI selection chooses an item i that maximizes the Fisher information evaluated at $\hat{\theta}_i$, the provisional θ estimate for the examinee after n 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, MI item selection is often based on a previously computed table in which items are sorted by the information they provide at each number of θ values (an "information table"). Item selection is equivalent for all θs 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 is needed to select the next item, it need only be evaluated once for each item at each tabulated

point. Item selection based on an information table is slightly less efficient but less computationally burdensome than MI item selection. These statistically motivated item selection procedures can be tempered by practical considerations such as item exposure rates.

Sympson-Hetter Exposure Control Method

The Sympson-Hetter (S-H) 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 selection determined by an iterative CAT simulation. Items with high administration frequencies are assigned smaller exposure control parameters, which range from 0 to 1. During test administration, the exposure control parameter of the selected item is compared to a uniform random number, which also ranges from 0 to 1. 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 simulations 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. The procedure is as follows:

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 items administered from items 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.

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 θ levels can be controlled, the same control might not hold for examinees at a particular level of ability. For instance, even though the exposure of an item might be controlled such that it is administered to no more than 30% of the examinees overall, it might 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 θ 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 have been proposed. Parshall, Davey, and Nering (1998) developed a conditional S-H 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 θ level. This addition to the S-H technique (often referred to as the

conditional S-H technique when a multinomial model is not used) is desirable because it overcomes the major disadvantages of the SH method by establishing an exposure control parameter for each item at a number of different θ levels.

Item Pool Design and Components of CAT

Parshall, Davey, and Nering (1998) discussed 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 might 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. When item security is a consideration, the more examinees taking the test, the more items that should be in the item pool. The CAT item selection procedure selects 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 examinees. 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 examinees, i.e., the number of examinees at each ability level.

With respect to the same item response patterns, different estimation methods might lead to slightly different ability estimates and, in turn, influence the choice of the best suitable item. Different item selection rules, such as selecting the item that maximizes the information or minimizes posterior variance at the current ability estimate, might choose different items to be the most appropriate item 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 content areas. 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 that take the same item, the more likely that item would be disclosed to the public. Item under-use potentially increases item development 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 exposure control and cause very little decrease in the efficiency of ability estimation.

Reckase's Simulation Method and Extensions to the Three-Parameter Logistic Model (3PLM)

Basic Concepts

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 CATs are administered to the examinees, assuming that each item administered to the examinee has the item parameters best suitable for the provisional ability estimate. After a certain number of examinees have taken the test, the union of the "virtual" items is the optimal item pool for the CAT program.

Theoretically, every θ estimate is unique and the items optimally suitable for the estimate have unique item parameters. The simulation process described above would lead to as many items in the item pool as the total number of items administered to examinees, i.e., the test length multiplied by the number of examinees. In practice, however, items function very similarly to those items with parameters that 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 the one-parameter logistic model (1PLM), the item difficulty parameter (b parameter) controls the selection of test items. The bins are defined as ranges on the IRT θ scale. For example, two consecutive bins with width of 0.2 on the θ scale are denoted as (0.0:0.2) and (0.2:0.4). Items with b parameters 0.11 and 0.13 are exchangeable in CAT item selection because they all belong to the bin (0.0:0.2). The item pool blueprint is simplified 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 that 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 might vary in their usefulness for estimating the ability level. The approach taken to determine bin width used here is to identify the range of the θ 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.

The end product of the optimal item pool design is an array of integers ($x_1, x_2, \ldots, x_B$), 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 might contain more items so that the shared exposure rates for items from the highly exposed bins are below the target exposure rate.

Reckase's Method for an Optimal Item Pool Calibrated With the 1PLM

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

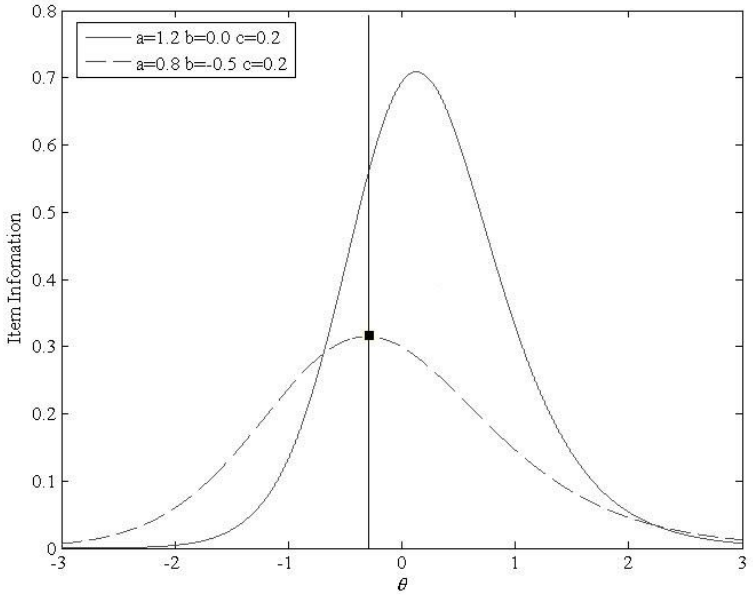
1. Understand clearly the characteristics of the CAT program, because item pool design must model the test procedure as closely as possible.
2. Identify the categorical attributes required for the items, such as content area, and divide the item pool into smaller pools according to these attributes. 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.
3. 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 θ metric. The first item is the same for all examinees. It is an item with maximum information at a θ of zero. The next optimal item will be based on the examinee's response that item and the θ estimate. Subsequent items are selected to have maximum information at the most recent θ estimate. If items are calibrated with the 1PLM, the optimal item is the one with a b value the same as the current θ estimate. As the test items are selected and administered, they are tallied in bins based on their b values.
4. The tallied items into bins form the distribution of items for one examinee. To form the item pool for the full test, the union of the distributions for the sample of examinees is produced. The results of the union operation is the item pool for this CAT design and the specified sample of examinees.

This strategy works well for item pools calibrated with the 1PLM, 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 a θ estimate will always provide maximum information at the θ estimate. Therefore, they are always the optimal items at the θ estimate compared to items with b parameters different from the θ estimate. When items are calibrated with the 2PLM or 3PLM, they might differ in the amount of information they provide even with the same b parameters, simply because they have different a or c parameters.

Reckase's Method Applied to the 3PLM

In the 3PLM, the information an item could provide at a θ 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 θ level, given that the *b* parameter is close to the θ 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 θ level, an item reaching the maximum information it could provide is not necessarily the item providing maximum information at the θ level. On the other hand, an item providing its highest information at one θ level might provide more information than any other items in the item pool over a range of θ levels. As demonstrated in Figure 1**,** an item with parameters $a = 1.2$, $b = 0.0$, and $c = 0.2$ provides more information at θ 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 θ level.

Figure 1. Item Information Provided By Two Different Items



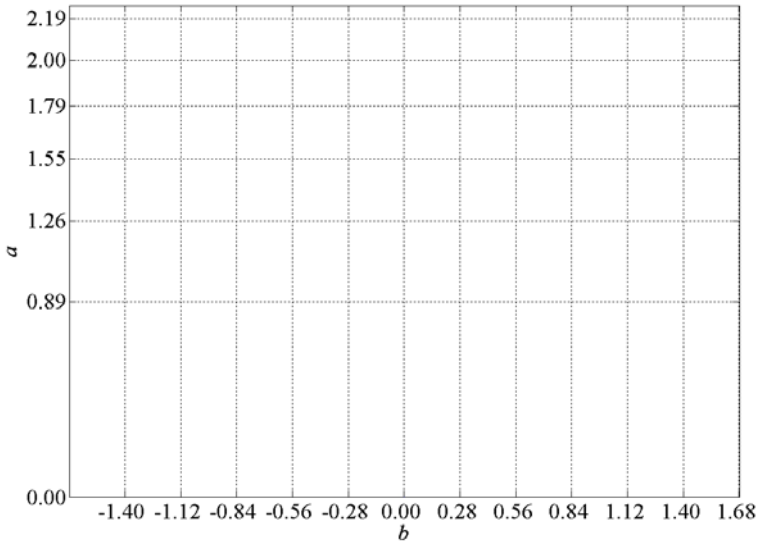
Therefore, the optimal item for a θ level should not be defined as the item providing its most information at the θ level. In addition, it is unrealistic to define the optimal item pool as the one that contains items with the highest possible *a* parameters. Instead, the optimal item pool should contain items with a range of discrimination 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 3PLM.

Extending the "Bin" Concept

Under the framework of the 3PLM, the maximum amount of information an item could provide is determined 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 might provide less information at a θ level where $\hat{\theta}$ is far from the examinee's true θ. An item with smaller *c* parameters provides more information at its maximum level, but for a well developed item pool, the *c* parameters usually vary slightly across items so for the case considered here they have little influence on the amount of information items provide. Therefore *a* and *b* parameters are the two primary factors that determine how much information an item is capable of providing at a θ level. Items that have similar information functions have similar *a* and *b* parameters. This leads to the extension of the "bin" concept introduced in item pool simulation with the 1PLM, where it is defined to be the interval of *b* parameter values within which items provide similar amounts of information over a range of θ levels.

With the 3PLM, the boundary of a "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 2, 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 an *a*-bin and the marginal total across each column is denoted as a *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 θ, and provide maximum information at the θ level around the boundary of the bin in which they are located.

Figure 2. Bins Defined by Both *a* and *b* Parameters



While the boundaries of the *b*-bins are determined by dividing the θ metric (or equivalently, the 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 6 shows the relationship between the *a* parameter and the maximum information, M_i, an item provides.

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}[1-20c_i-8c_i^2+(1+8c_i)^{3/2}] \quad (6)$$

It can be further shown that the differences between the maximum information function (ΔM) for items with different *a* parameters is

$$\Delta M = \frac{D^2[1-20c-8c^2+(1+8c)^{3/2}]}{8(1-c)^2}\Delta a^2 \quad (7)$$

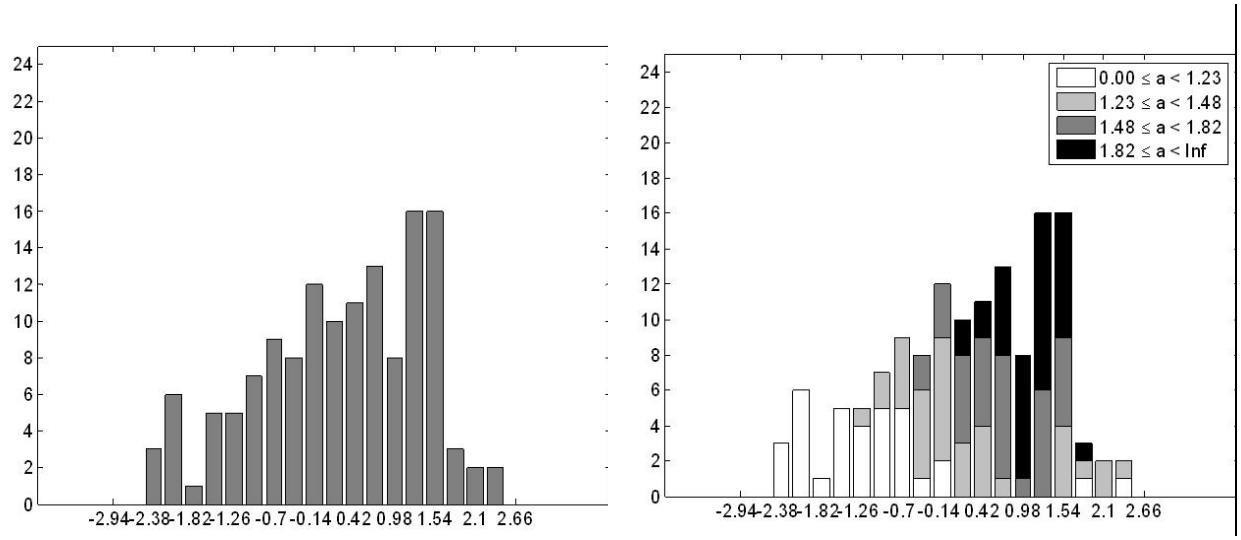
Using the average *c* parameter of the existing items, which is around 0.2, the resulted constant is 0.5, therefore

$$\Delta M = 0.5\ \Delta a^2 \quad (8)$$

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. 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.

Figure 3. Item Distribution By *b*-Bins and *ab*-Bins

***a.* *b*-Bins** **b. *ab*-Bins**



Distinctions are made, however, with respect to the functions of *b*-bins and *a*-bins. As mentioned above, the closeness of the *b* parameters to θ 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 θ level where it functions the best. With the MI item selection approach, if an item with high information at a θ

level is available, it will be selected 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 θ 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 $\vec{x} = (x_1, x_2, \ldots, x_B)$, which tells how many items are needed in each *b*-bin, and a matrix $X = (X_1, X_2, \ldots, X_B)$, where each element X_B is a integer vector (y_{B1}, $y_{B2}, \ldots, y_{BA}$) 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_Bs from *ab*-bins with the lowest item discrimination are set to zero so that $\Sigma\, y_{B\cdot} = X_B$ 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, where Figure 3a shows how many items in each *b*-bin are needed for the optimal pool and Figure 3b distinguishes different *ab*-bins with gray-scales and shows the number of items needed for each *ab*-bin within a *b*-bin.

Strategies to Generate Items for Item Pool Simulation with the 3PLM

During the item pool simulation, each item generated for the current θ estimate is assumed to provide its most information at the θ 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 1PLM situation where it is assumed that the optimal item is the one with *b* parameters close to the current θ estimate and which provides the most item information. Note that with the 3PLM this item is not necessarily the item that provides more information at the current θ 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 θ level close to the other bin. However, by assuming that optimal items are those that provide their most information at the θ estimates, recording the *ab*-bin items belong to is equivalent to recording approximately how much information is needed at the θ 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 the next section.

Generating Item Parameters During the Item Pool Simulation Process

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). 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).

$$a' = \log(a_i) = \beta_0 + \beta_1 b_i + \varepsilon_i \quad (9)$$

$$a = exp(a') \quad (10)$$

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 with 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:

1. After each response, obtain the estimate of θ ($\hat{\theta}$) and use it as the approximation of the *b* parameter for the next optimal item.
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.
3. Generate the *c* parameter from the beta distribution. Re-compute the *b* parameter so that the item provides maximum information at $\hat{\theta}$:

$$b_i = \hat{\theta}_{ij} - \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \tag{11}$$

With the PM strategy, the -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 θ 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.

Minimum test information (MTI) strategy. The MTI strategy posits that the item pool is optimal when CATs assembled from it can provide just sufficient information to measure 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 discriminating items.

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

The MTI strategy generates items in three steps:

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

$$I_i = \frac{I_{target} - I_{adm}}{l_{test} - l_{adm}}. \quad (12)$$

To mimic the way CAT selects items, i.e., selecting the items providing the largest information at the current θ estimate, the target information could be manipulated so that target information starts with a reasonably large number, 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 (Chang & Ying, 1999), 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.

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 $\hat{\theta}$ is

$$\hat{\theta}_i = b_i + \frac{1}{Da_i} \ln\left(\frac{1+\sqrt{1-8c_i}}{2}\right), , \quad (13)$$

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

$$M_i = \frac{D^2 a_i^2}{8(1-c_i)^2}\left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{3/2}\right] \quad (14)$$

By rearranging the equation and replacing M_i with I_i, it can be shown that

$$a_i = \sqrt{\frac{8(1-c_i)^2 I_i}{D^2\left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{3/2}\right]}} \quad (15)$$

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 θ estimate.

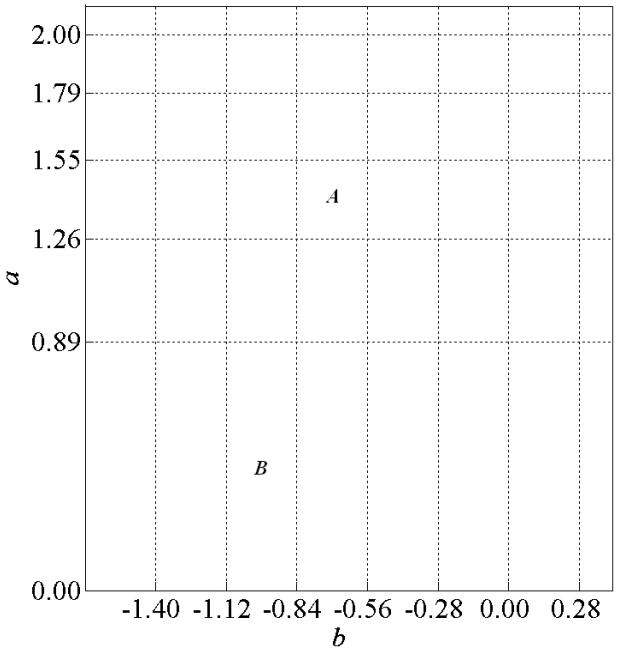
3. Calculate the *b* parameter with *a* and *c* parameters from Equation 12 so that the generated item provides its most information at $\hat{\theta}$.

Post-Simulation Adjustment

The results from the item pool simulation, the vector $\vec{x}$ 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* parameters 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 θ level within the other bin. For example, Figure 4 shows that items from *ab*-bin A (–0.84:–0.56, 1.26:1.55) might provide more information at θ'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 might provide their most information over the same θ range. In other words, the item selection

procedure would choose the item in bin A over the item in bin B for θ estimates around −1.12 to −0.84.

Figure 4. 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 that resulted from item pool simulation will have redundant items.

These items are trimmed by using an information 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 midpoints of the *b*-bins are treated as the anchor θ level, and the midpoints of both the *a*-bin and *b*-bin as *a* parameters and *b* parameters, respectively, representing the bin that the item comes from. For example, the θ 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 information table. Sufficient items are drawn in this way to represent the number of items needed in each bin.

As shown in Figure 5, 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 they provide 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.

Figure 5. Items in the Order of Highest Information in Each *b*-Bin

-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

One caveat of this procedure is that because items provide more information over a range of θ levels, they might be administered more times than the test developers desire. 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 an item 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 6 shows the expected item usage within each *b*-bin for 8,000 examinees ordered by the information each item provides. Item 109 in Figure 5 is expected to be selected 11,800 times (8,000 + 2,471 + 1,329), 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 rate. Four more items from the same *ab*-bin, therefore, need to be added.

Figures 7 and 8 demonstrate an item pool blueprint before and after post-simulation adjustment.

Figure 6. Item Usage in the Order of Highest Information in Each *b*-*B*in

-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 7. Item Distribution for Optimal Item Pool Before Adjustment

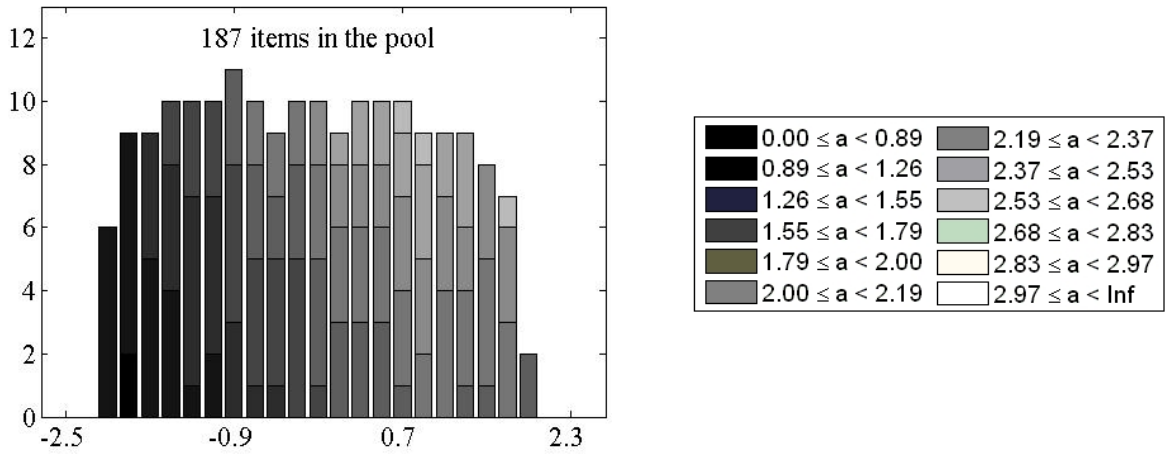
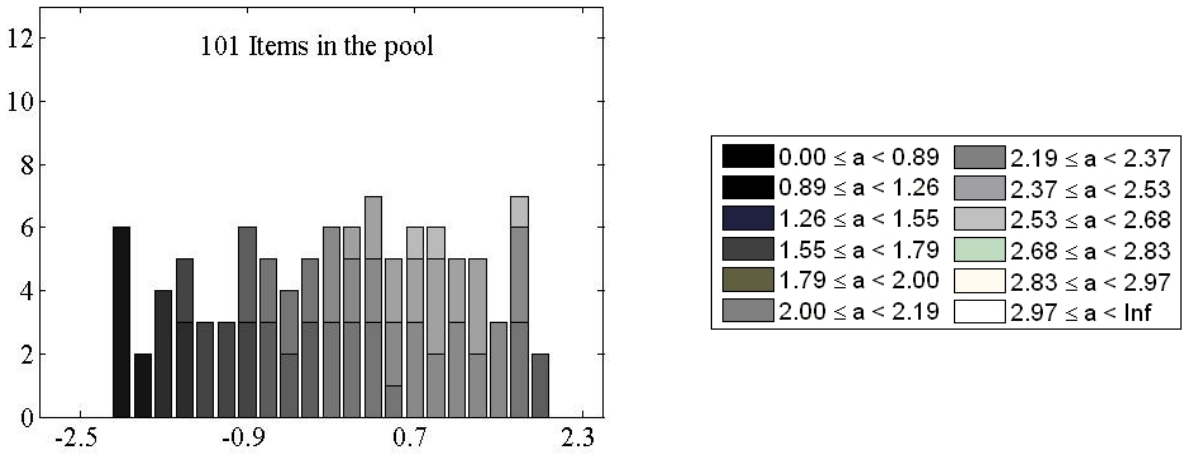


Figure 8. Item Distribution for the Optimal Item Pool After Post-Simulation Adjustment



Design Adjustments to Exposure Control Method

The 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 a bin, which is shared by the number of items needed in the bin. If the CAT has no item-exposure control, the design blueprint for the item pool after post-simulation adjustment follows directly from these exposure rates.

The Sympson-Hetter (1985) 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 that 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.

Method

This study was composed of two closely related parts. In the first part, Monte Carlo simulations were used to design optimal item pools for a large-scale aptitude test. The second part evaluated the optimal item pools with empirical criteria and compared them to the operational item pool on performances.

Operational Item Pool

The operational item pool for a section of a large aptitude test was used as the target and benchmark of this study. The section was a 15-item test measuring the ability to solve basic arithmetic word problems. Items were selected using MI. To save computation time, an information "look-up" table was 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 θ estimate during test administration. For each test, the prior θ distribution had a mean of 0.0 and a standard deviation of 1.0. The θ estimate after the last item was used as the score for the test. Each test was terminated after a fixed number of items.

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 was to investigate the optimal item pool design for a specific testing program, the simulation procedure followed closely the psychometric procedure used in the operations.

Test length was the same as the operational test, which was a 15-item fixed length section. Content balancing was not considered. MI and the information table were used to select items. Owen's approximation to the posterior mean (Owen, 1975) was used to update the θ estimate during test administration. For each test, the prior distribution of θ had a mean of 0.0 and a standard deviation of 1.0.

Step 2: Generating the examinee population. The operational item pool was designed to serve examinees whose θ distribution was assumed to be 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 a θ level of 0. After each response, optimal items were generated for the current θ estimate. It was assumed that items were calibrated by the 3PLM. Therefore, *a*, *b,* and *c* parameters were generated by one of the two methods (PM and MTI) described above. 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 θ estimate and method used (PM or MTI). The *b* parameters were generated so that the item provided its most information at $\hat{\theta}$.

Step 4: Generating response data. Examinee responses were generated following each item generation according to the 3PLM, where the probability of examinee *i* correctly answering item *j* is expressed as:

$$P_i(\theta_j) \equiv c_i + (1 - c_i)\frac{1}{1 + \exp\left[-1.7a_i\left(\theta_j - b_i\right)\right]} \tag{16}$$

$P_i(\theta_j)$ is the probability that a person $j = 1, \ldots, J$ with an ability parameter θ_j gives a correct response to item $i = 1, \ldots, I$; 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 θ 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 5: Post-simulation adjustment. Five replications were conducted for each combination of methods and control variables so that a relatively stable approximation 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.

Control Variables

Two independent variables were controlled for in all item pool designs: design method and exposure control method. The target exposure control rate was 1/3 for the Sympson-Hetter method, which was the the same as the operational procedure. The simulation design is illustrated in Table 1.

Table 1. Simulation Design

Test length	15
Examinee distribution	N(0,1)
Exposure control	No exposure control
	Sympson-Hetter (target exposure rate was 1/3)
Design Method	Prediction Model
	Minimum Test Information
Bin width	*b*-bin: 0.20 *a*-bin: $\Delta a^2 = 2\Delta I_{Maximum} = 0.8$
Content Balancing	Single content

Evaluating Simulated and Operational Item Pools

Two types of distributions were considered in the item pool evaluation: (1) 6,000 θ's were simulated from N(0,1), and these values were treated as the true abilities for the examinees, and (2) 65 fixed values of θ ranging from –4 to 4 with an interval of 0.125 were selected (i.e., θ = –4.0, –3.875, . . . , 3.875, 4.0). Five hundred examinees were set to have an identical latent ability at each θ level. The former was to evaluate general performance, and the latter was to compute statistics conditional on θ.

The item pool evaluation criteria used by Chang and Ying (1999) and Reckase (2005) were adopted for this study. Precision of θ estimation included average test information at each θ level, bias, mean square error (MSE), and correlation coefficients between estimated and true θs. Test security indicators included 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, it can be taken as an 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 θ point, the standard error of measurement (SEM) was calculated by

$$SEM(\theta_i) = \sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}\left(\hat{\theta}_{ij} - \bar{\theta}_i\right)} \tag{17}$$

where $N_i = 500$ was the number of replications (i.e., the number of CATs administered) at each fixed θ point, and $\bar{\theta}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\hat{\theta}_i$ is the mean of the θ estimates over the N_i replications at θ_i.

Bias and mean square error (MSE). These quantities were defined as follows:

$$Bias = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right) \tag{18}$$

and

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - \theta_j\right)^2 , \tag{19}$$

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

Conditional bias and conditional mean square error (CMSE). These quantities were defined as

$$Conditional\ Bias = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\hat{\theta}_{j_i} - \theta_i\right) , \tag{20}$$

and

$$CMSE = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\hat{\theta}_{j_i} - \theta_i\right)^2 , \tag{21}$$

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

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

$$\chi^2 = \sum_{i=1}^{n} \frac{(r_i - L/n)^2}{L/n}, \tag{22}$$

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 22 captures the discrepancy between the observed and the ideal item exposure rates, and it quantifies the efficiency of item pool usage. A low χ^2 value implies that most of the items were fully used.

The ratio of χ^2 measures follows an F distribution and can be used to compare the exposure rates of two methods:

$$F_{\text{method1, method2}} = \chi^2_{\text{method 1}} / \chi^2_{\text{method 2}} \tag{23}$$

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 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 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 some CAT programs.

Percentage of underexposed items. A low item exposure rate means that an 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 was considered as underexposed.

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):

$$\overline{T} = \frac{\sum_{i=1}^{n} \binom{m_i}{2}}{L\binom{N}{2}} = \frac{\sum_{i=1}^{n} m_i(m_i - 1)}{LN(N-1)} \tag{24}$$

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.

Results

Performance of the Item Pools without Exposure Control

Figure 9 compares the distributions of the operational item pool and two optimal item pools designed by MTI and PM, assuming no exposure control. Table 2 presents the pool sizes and the summary statistics of the item parameters for the three pools. The optimal item pools consisted of the fewest items. This is not surprising, partly because both assumed no exposure control, while the operational pool was designed for tests with Sympson-Hetter exposure control.

Table 2 indicates that all item pools had items that spanned a wide range of difficulty levels, roughly from –2.5 to 2.5. However, the items in the optimal item pools had slightly smaller ranges. The operational pool had a large number of items with b parameters between 0.0 and 1.5, while the optimal pools displayed a more even distribution across b-bins. The MTI pool consisted of the fewest items and their a parameters were more concentrated, ranging from 1.275 to 1.781. The PM pool had item parameters similar to the operational pool, in which difficult items tended to have high a parameters and easy items tended to have moderate to low a parameters.

The overview of the evaluation results for these item pools are presented in Table 3. The θ estimates from all pools exhibited a certain level of positive bias; however, the magnitudes of the bias were negligible. MSEs from optimal item pools were smaller than that from the operational pool. The MTI pool and PM pool resulted in a higher correlation coefficient than the operational pool. Table 3 also shows that optimal item pools had a smaller test-retest overlap rate despite having fewer items. It also indicates that the magnitude of the item overlap rate might not be related to the pool size with the optimal combinations of the items in the pool.

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

Figure 10 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 3 indicates that the MTI pool has the fewest under-exposed items and Figure 10a shows that items with extreme difficulty levels are utilized more often in MTI pools.

As shown in Figure 11, the three item pools resulted in quite different average test information plots at various fixed θ levels. The plot for the PM item pool looks similar to the one for the operational pool, but provides more information over most θ levels. The MTI item pool provides significantly less information over the θ range between approximately –1.5 and 2.0, but the amount of information it provides over a wide range of θ levels exceeds the target information, which is 10.0 between θ levels ±2.0, and 8.0 beyond θ levels ±2.0.

Figure 9. Item Distribution for Item Pools Without Exposure Control

a. Operational Item Pool **b. Item Pool Designed by MTI** **c. Item Pool Designed by PM**

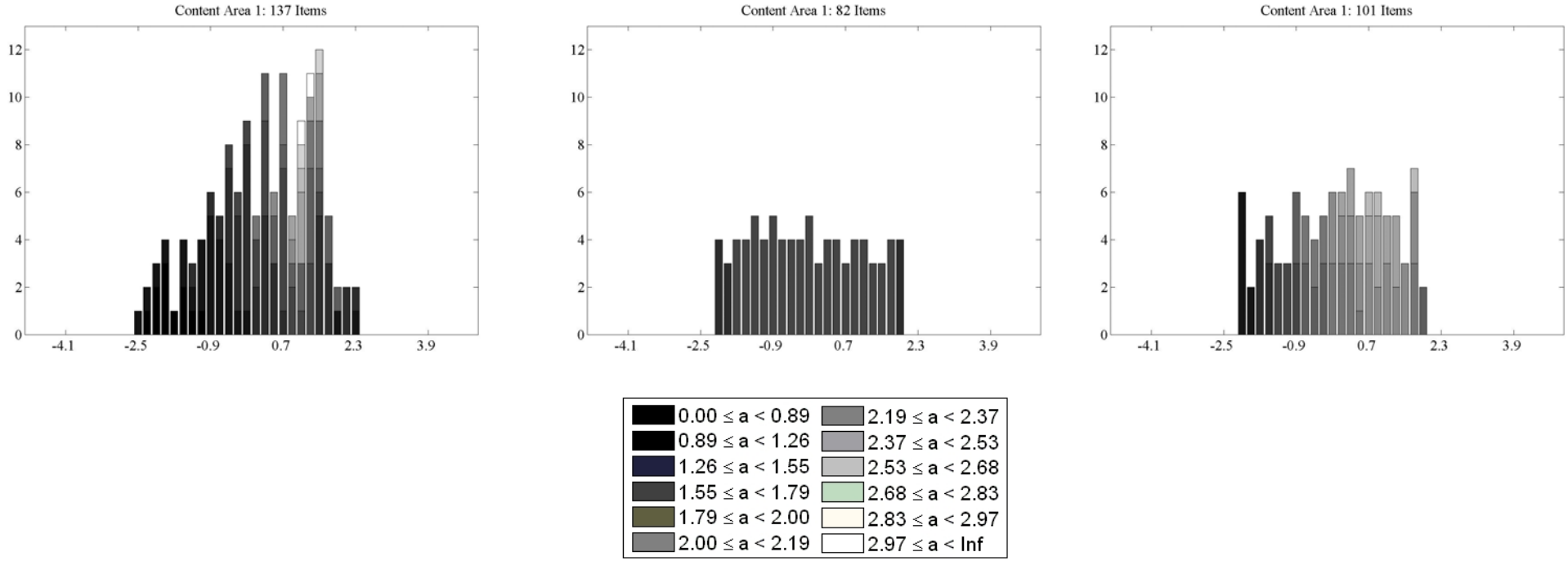


Table 2. Item Pool Size and Item Parameter Statistics for Arithmetic Reasoning Without Exposure Control

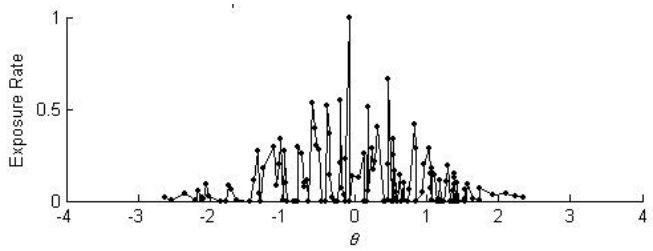
	Pool	*a*				*b*				*c*			
Pool	Size	Mean	SD	Max.	Min.	Mean	SD	Max.	Min.	Mean	SD	Max.	Min.
OP	137	1.556	0.487	3.141	0.746	0.115	1.170	2.343	-2.625	0.186	0.063	0.328	0.038
MTI	82	1.601	0.105	1.781	1.275	-0.159	1.194	1.942	-2.186	0.218	0.075	0.423	0.069
PM	101	2.000	0.427	2.638	0.932	-0.031	1.143	1.943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

Table 3. Summary Statistics of the Performance of the Item Pools

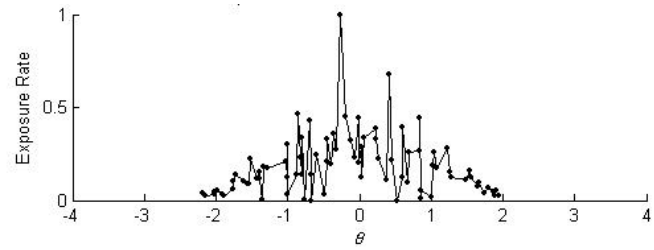
Statistic	OP	MTI	PM
Bias	0.0025	0.0022	0.0114
MSE	0.0857	0.0739	0.0576
Correlation	0.9563	0.9636	0.9703
Skewness of item exposure rate	31.3822	12.0199	15.0003
Item overlap rate	0.3385	0.3294	0.2969
Pct of items with item exposure Rate > 1/3	8.76%	14.63%	8.91%
Pct of items with item exposure rate < .02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

Figure 10. Item Exposure Rate by Difficulty Level

a. Operational Item Pool



b. Item Pool Designed by MTI



c. Item Pool Designed by PM

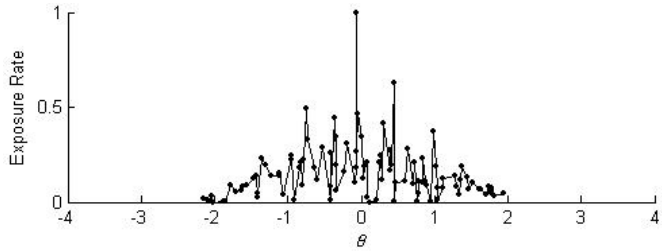


Figure 11. Average Test Information Conditional on True θ

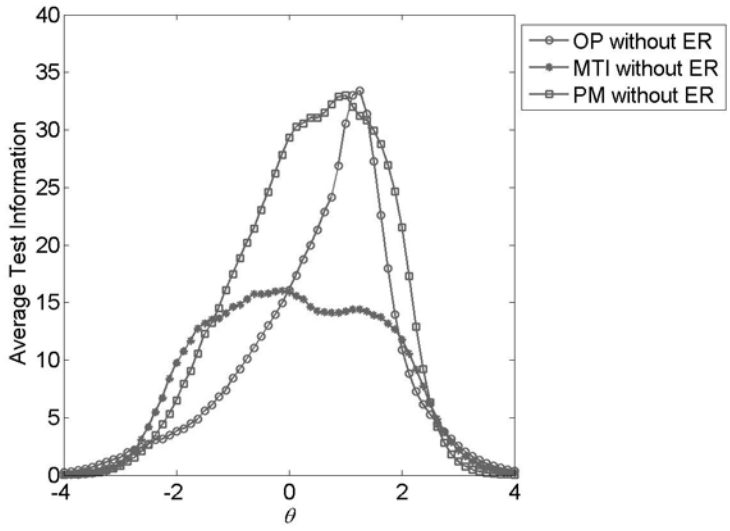


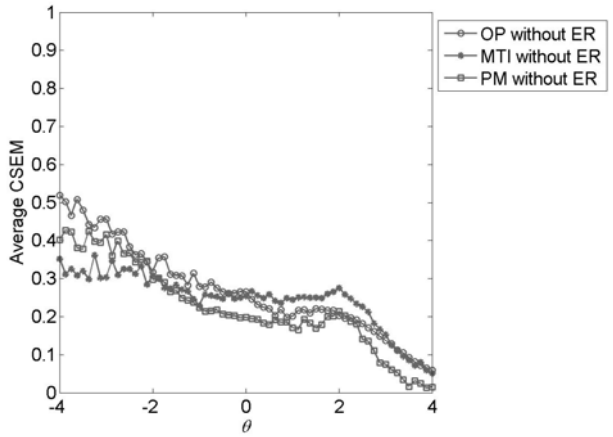
Figure 12 present the CSEM, conditional bias, and CMSE for the three item pools. Figure 5b shows a significant increase in the bias of θ estimation, which is positive for θ levels below around –2.0 and negative for θ levels above around 2.0. It is not surprising because of the short test length and the Bayesian estimation method. MTI performed better for θ levels below –2.0 and PM performed better for θ levels over 2.5.

Summary. The results suggest that regardless of the constraints of content balancing, the optimal item pools performed better than the operational item pool based on pool size, test security, and measurement accuracy, although each design method had its preferable features. The operational item pool performed better over a given range of θ levels because a large number of items, including very discriminating items, were clustered around these levels. Optimal item pools, especially those designed with MTI, provided information more evenly over most θ levels and provided sufficient measurement precision with a minimum number of items. All optimal pools, compared to operational pools, saved about 20 or more items and yielded better correlations. In addition, optimal pools had a significantly lower percentage of items with exposure rates below 0.02. With or without content balancing, PM item pools resulted in the highest correlation and the lowest item overlap rate.

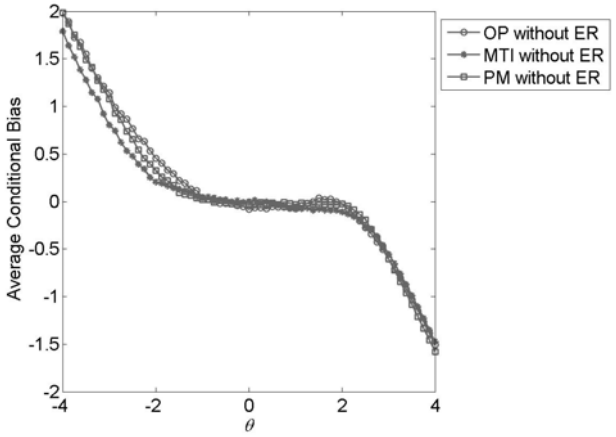
Overall, an item pool designed with the MTI method performed the best, which indicates that the optimal item pool needs the fewest items to achieve desirable precision if all the items have moderate item discrimination and distribute roughly uniformly over a wide range of difficulty levels.

Figure 12. Comparison of Item Pool Performance Conditional on θ Estimates

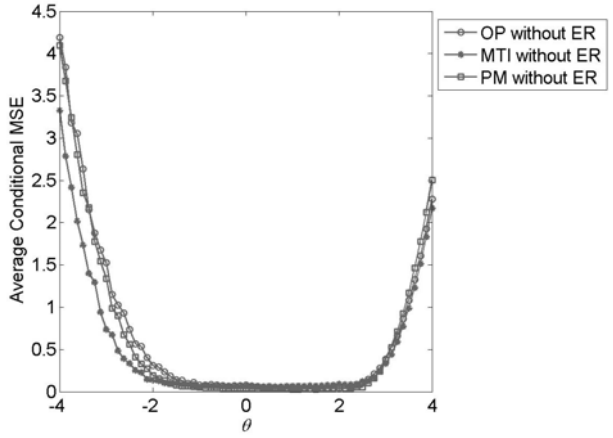
a. CSEM



b. Conditional Bias



c. CMSE



Performance of the Item Pool With Sympson-Hetter Exposure Control

The blueprints of the optimal item pools with Sympson-Hetter exposure control were built upon the blueprint of those without exposure control. Specifically, more items were added in the bins where items tended to be selected more often than the desired exposure rates. This relationship is reflected in the item distribution illustrated in Figure 13, where compared to the optimal pools without exposure control, there were noticeably more items with *b* parameters in the range −1.0 to 1.0 in the optimal pools designed by both MTI and PM.

Table 4 shows the item pool size and the summary statistics of the item parameters within each pool. The MTI pool consisted of the fewest items and their *a* parameters varied within 1.307 and 1.777, a smaller range than the other two pools. The optimal item pool designed by PM had more high *a*-parameter items. However, items in the operational pool had the maximum *a* parameter value of 3.141, compared to 1.777 for items in the MTI pool and 2.633 for items in the PM pool.

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

Table 5 lists the performance overview of the item pools. On average, all three pools yielded slightly positive bias for θ estimates. The operational pool displayed the smallest bias, but the difference from the optimal pools was 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, consisted of the fewest items, which was 42 items less than the operational pool and 25 less than the PM pool. In addition, the optimal item pools had slightly smaller test-retest overlap rate despite having fewer items.

Figure 14 shows the item exposure rate for individual items in each pool in the order of item difficulty. It can be seen that the exposure control method worked very well, with the exposure rates for all individual items around or below the target exposure rate. The MTI pool seemed to utilize items more evenly and have the fewest under-exposed items. The operational item pool seemed to have large numbers of difficult items underexposed.

A closer look at the measurement precision at the individual θ levels is displayed with the conditional test information plots in Figure 15. The plots for item pools with exposure control look very similar to those without exposure control. Because of the added items, optimal pools with SH exposure control yielded more information at some θ levels and closely matched the information provided at other levels with the optimal pool without exposure control. The operational pool, on the other hand, produced less information at θ levels between −0.5 to 0.75 when SH exposure control was used.

Figure 13. Item Distributions for Item Pools With Sympson-Hetter Exposure Control

a. Operational Item Pool **b. Item Pool Designed by MTI** **c. Item Pool Designed by PM**

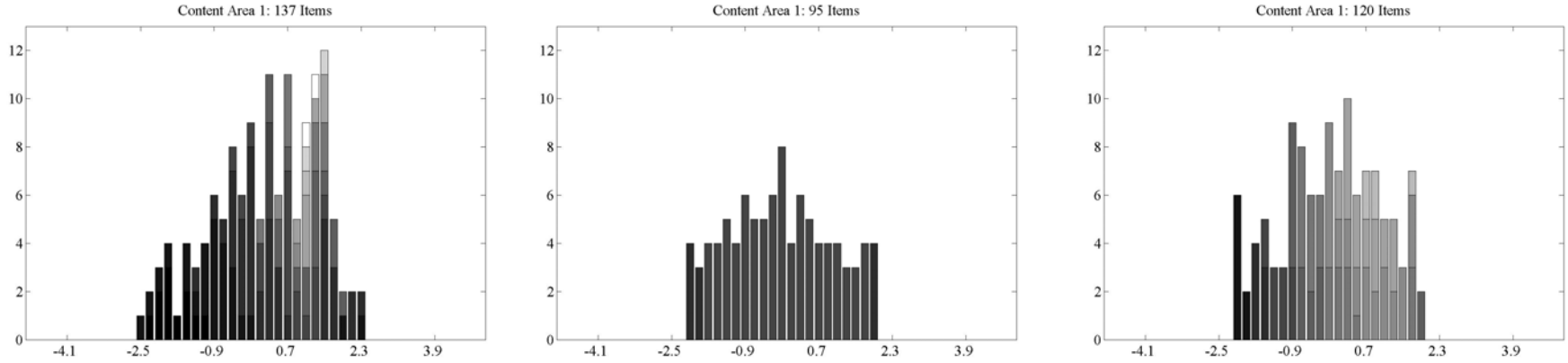


Table 4. Item Pool Size and Item Parameter Statistics for Arithmetic Reasoning with Sympson-Hetter Exposure Control

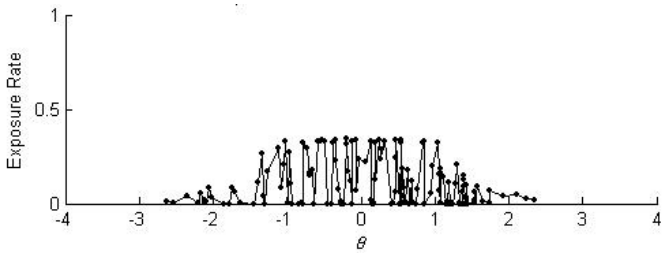
	Pool	*a*				*b*				*c*			
Pool	**Size**	**Mean**	**SD**	**Max.**	**Min.**	**Mean**	**SD**	**Max.**	**Min.**	**Mean**	**SD**	**Max.**	**Min.**
OP	137	1.556	0.487	3.141	0.746	0.115	1.170	2.343	-2.625	0.186	0.063	0.328	0.038
MTI	95	1.616	0.092	1.777	1.307	-0.141	1.130	1.935	-2.172	0.228	0.076	0.498	0.082
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 5. Summary Statistics of the Performance of the Item Pools

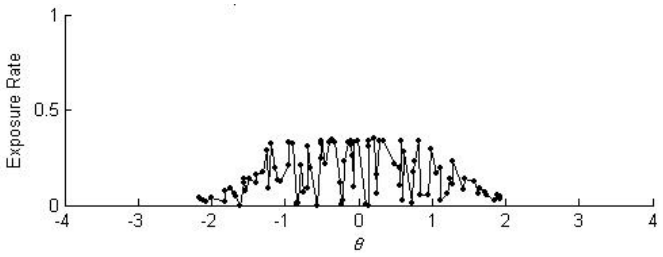
Statistic	OP	MTI	PM
Bias	0.0073	0.0104	0.0105
MSE	0.0929	0.0823	0.0564
Correlation	0.9525	0.9593	0.9728
Skewness of item exposure rate	18.9078	8.5813	10.7972
Item overlap rate	0.2474	0.2481	0.2149
Pct of items with item exposure rate > 1/3	5.11%	11.58%	4.17%
Pct of items with item exposure rate < .02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

Figure 14. Item Exposure Rate by Difficulty Level

a. Operational Item Pool



b. Item Pool Designed by MTI



c. Item Pool Designed by PM

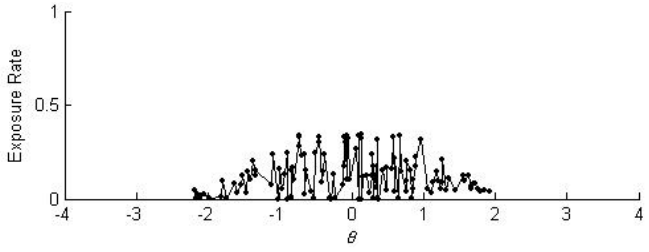
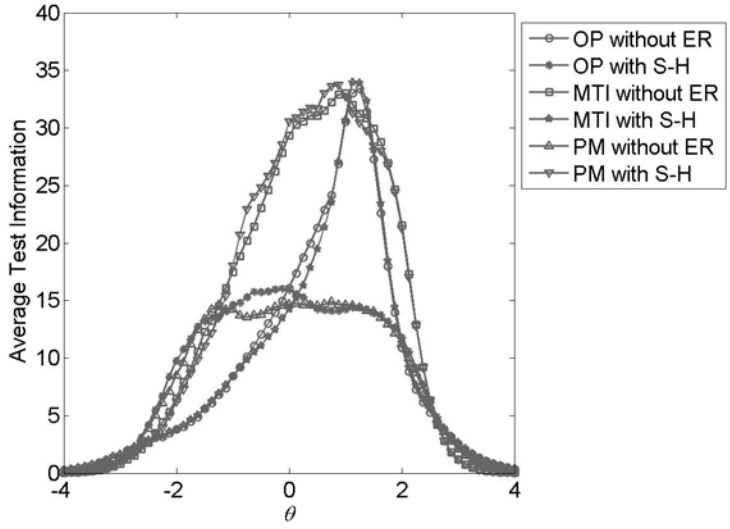


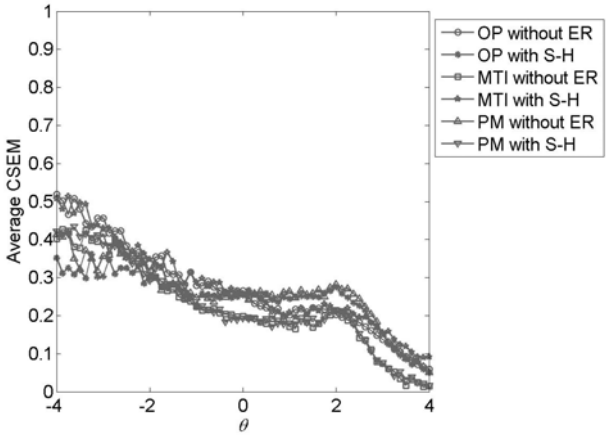
Figure 15. Average Test Information Conditional on True θ



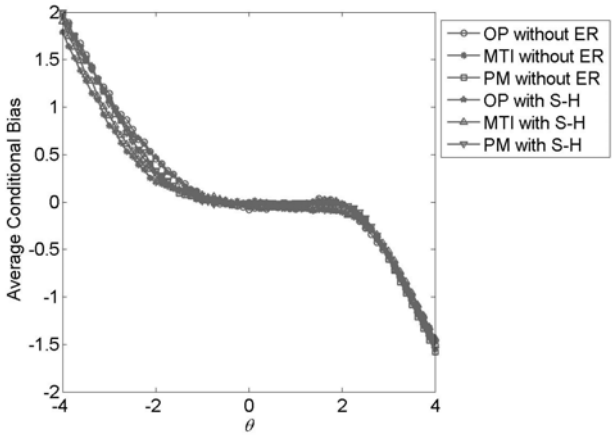
The plots for conditional SEM, conditional bias, and conditional MSE are presented in Figure 16. Smaller values indicate better accuracy in θ estimates. The plots indicate that all item pools yielded similar performance at θ levels between –2.0 and 2.5. The MTI pool performed better for θ levels below –2.0 and the PM pool performed better for θ levels over 2.5.

Figure 16. Comparison of Item Pool Performance Conditional on θ Estimates

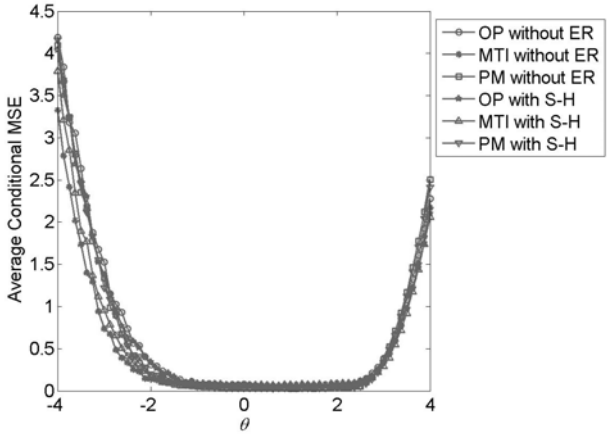
a. CSEM



b. Conditional Bias



c. CMSE



Summary

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

Discussion and Conclusions

Implications for the Practice of Item Pool Development

This study investigated two approaches to designing optimal blueprints for CAT item pools. Both approaches designed item pools performing better than the operational pools whether exposure control was considered or not. The results showed that the MTI design generally led to smaller pools that contained items with lower *a* parameters. PM pools maintained the correlation between item parameters but did not perform as well as the MTI pool. The operational pool, on

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

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 CATs. 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 might yield similar measurement precision. That is why the two "optimal" pools looked quite different and each was still optimal in some sense.

This study was 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 might 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 of the assumptions might 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 a 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 items or created by item writers.

This method has not been tested in practice. In this study, all items required by the design method were assumed available when comparing the optimally designed item pools to the operational pools. It seems difficult to produce items with exactly the same item parameters required by the item pool design blueprint. However, with advances in item modeling research, it will be more and more feasible to create large numbers of similar items with the desired psychometric properties. Because the PM pool takes into account the correlations between *a* and *b* parameters, the blueprint designed by the PM method might be easier to fulfill. The MTI pools achieved acceptable measurement precision with a minimum number of items, but it is uncertain how difficult 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 difficult it is to fulfill the required items of the blueprint.

Implications for 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

$$X'_{AB}=Max(RX_{AB}r_{AB}, X_{AB}) \quad (25)$$

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.

Reckase's Method Versus the Mathematical Programming Method

The results showed that the extensions to Reckase's method worked well in designing optimal item pools in situations where items are calibrated with the 3PLM. Compared to the mathematical programming method, Reckase's method simulated the CAT procedure straightforwardly and, therefore, is more flexible in adapting different item selection and ability estimation processes 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 pools. 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 focuses 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 between the PM item simulation approach and the mathematical programming method is in the way item costs are minimized in the item pool design process. The mathematical programming method defines a cost function, which is an inverse of the number of real items with certain combinations 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 might be able to borrow some ideas from the other to improve 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 might be possible to simulate the item pool design by varying the target information at different stages of the test.

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 obtain a more accurate estimation of the attributes of the items written for a testing program.

Previous research showed that the bin width might influence the number of items required in the optimal pool. With the 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 its impact.

While this study investigated optimal item pool design with the PM and MTI methods separately, both methods have their shortcomings. The 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.

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*, 345-360.

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.

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

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

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.

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

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.

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

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*, 315-330.

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*, 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.

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.

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.

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

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 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., 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 99-0). Enschede, The Netherlands: Twente University, Faculty of Educational Science and Technology.

Appendix

Table A.1 Item Distribution for the Operational Item Pool

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	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 the MTI Method Without Exposure Control

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	∞	
-∞	-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 the PM Method Without Exposure Control

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	∞	
-∞	-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

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	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 by the PM Method and With Sympson-Hetter Exposure Control

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	∞	
-∞	-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 the Item Pool Simulated by the MTI Method and With *a*-Stratified Exposure Control

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
-∞	-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