Technical Report

A Review of Exposure Control Strategies for CAT and Potential Applications in MST

Prepared For:

The American Institute of Certified Public Accountants (AICPA)

Ting Xu

University of Pittsburgh

August 2010

INTRODUCTION

Computerized adaptive testing offers the advantages of more precise and efficient ability estimation and more flexible scheduling for testing when compared with conventional linear tests. Both item-level and testlet-level computerized adaptive tests have proved to be useful alternatives to the conventional paper-and-pencil tests. However, a potential problem is test security. In computerized adaptive testing, different examinees are presented with different sets of items and they are allowed to take the exam at different times. Consequently, some examinees may obtain knowledge of some test items prior to testing from those who have already taken the exam and seen the items. Thus, these examinees would be advantaged on the test and their test scores are no longer valid measures of their true trait levels. To maintain test security, attempts need to be made to avoid repeatedly administering the same items to examinees.

Various exposure control strategies have been proposed to control overexposure of test items for item-level adaptive tests. Moreover, some of these strategies also aim at improving the usage of items with a tendency to serious underexposure. The purpose of this study is to provide a brief review of item exposure control strategies for item-level computerized adaptive tests (CATs). Furthermore, based on existing exposure control strategies, this study attempts to propose two designs that extend exiting exposure control strategies to multistage adaptive testing (MST). The MST model considered in this study is a three-stage computerized sequential adaptive testing (CAST) design, which is the administration model of the CPA exams.

The rest of the paper is organized as follows: Section 1 provides a literature review of existing item exposure control strategies for CAT. Based on these strategies, section 2 proposes two designs that incorporate exposure control procedures for CAT into MST for the three-stage

CAST model of the CPA exams. A pilot study using a master pool of the CPA exams is presented. Finally, the concluding remarks are made in Section 3.

Section 1: Item Exposure Control Strategies for Computerized Adaptive Testing

Many high-stakes examinations are now being routinely administered in a computerized adaptive testing (CAT) (e.g., the Graduate Record Examination GRE, Armed Services Vocational Aptitude Battery) or multistage adaptive testing (MST) format (e.g., the Uniform Certified Public Accountants Examination, the Law School Admission Test). The main advantage of the CAT format is that it provides more efficient and precise estimation of the examinee's ability than does the conventional paper-and-pencil exam. In CAT, each successive item is selected so that the difficulty level of the item is tailored to the currently estimated examinee ability, thereby providing improved ability estimation. A practical advantage of CAT is that it allows examinees to take the exam based on a flexible schedule. However, the convenience for examinees and the item selection procedure adopted by CAT can constitute threats to test security if item exposure is not well controlled.

Reckase (2010) addressed five major components of a CAT: 1) an item response theory (IRT) model; 2) a calibrated item pool; 3) an item selection algorithm; 4) a trait estimation method; and 5) a stopping rule. In addition, for high-stakes testing, issues such as content balancing and item exposure also need to be considered (Reckase, 2010). Most of the CAT designs are based on IRT (Folk & Smith, 2002). The focus of this review is on IRT-based adaptive tests. A calibrated item pool contains all items calibrated through an IRT model and they are ready to be administered to the examinees. An ideal item pool should contain sufficient number of items so that multiple test forms can be assembled for a range of examinee abilities (Davey & Nering, 2002). To date, most adaptive testing programs have implemented fixed-length stopping rule.

In the current CAT applications, the maximum information (MI) method is a commonly used strategy to select items (Wainer, 2000). This method selects the unused item of the pool that provides the most information at the last estimated ability. Under the three-parameter IRT model, item information increases as the item difficulty approaches the ability level of the examinee, as the discrimination parameter increases, and as the chance of pseudo-guessing approaches zero (Hambleton & Swaminathan, 1985). Accordingly, given a provisional estimate of the examinee's trait level, the MI algorithm tends to select the next item with a high discrimination value if its difficulty level is close to the provisional estimated value. Consequently, items with high discrimination values are frequently administered (overexposure), while some others may never or rarely be selected (underexposure).

Item over- and under-exposure may produce some undesirable results. If an item is frequently administered, a large number of examinees will see the item, which increases the risk that some examinees will memorize and share it with those who have not yet taken the test. Once an examinee has prior knowledge about the item, it will no longer be a valid measure of the latent trait. In the case of underexposed items, if a large proportion of items are rarely administered, the item pool is not well utilized. Georgiadou, Triantafillou, and Economides (2007) concluded that "In short, all the items from the item pool should be used for economic reason while no item should be overused for security reason" (p.6).

As such, researchers have developed various exposure control mechanisms, including item-exposure rate control, item pool design, and item pool rotation (Ariel, et al., 2004). According to Revelta and Ponsoda (1998), item exposure control strategies aim at achieving two major goals: (a) preventing overexposure of some items, and (b) improving the usage of items that are never or rarely selected. Item exposure rate is the main index that is used to track item exposure in

CATs. It refers to the proportion of times an item is administered to the total number of CATs that have been administered.

Item Exposure Control Methods

Georgiadou, Triantafillou, and Economides (2007) classified exposure control procedures in CAT into five categories: 1) randomized item-selection strategies; 2) conditional selection strategies; 3) stratified strategies; 4) combined strategies; and 5) multiple stage adaptive test (MST) designs. The randomized item-selection procedures add a random component to the item selection process. They identify multiple items near the optimal level of maximum information, rather than the most informative item, and then randomly select one for administration. While simple and easy to implement, randomized strategies do not directly control the maximum item exposure rate. Alternatively, the conditional selection strategies incorporate an exposure control parameter for each item so that the maximum exposure rate is controlled under a prespecified target value. A major disadvantage of the conditional selection strategies is that the process of finding the exposure control parameter value for each item is time-consuming. The stratified strategies administer items with low discrimination at the early stage of CAT and then items with higher discrimination as the test progresses. The purpose is to produce more balanced item pool usage without losing measurement precision. Combining different exposure control methods to form a new method in certain conditions appear to perform better than using a single strategy. MST designs are partially adaptive where adaptation occurs at the level of sets of items. In MST designs, tests are pre-constructed and exposure controls may be achieved during the test assembly process rather than during the adaptive testing. As far as we know, none of the item exposure strategies currently used in formal CAT has been applied in MST.

In addition to the above methods, strategies for managing item pool have also been developed to address item exposure in CAT, such as methods of item pool design (e.g., van der Linden et al., 2006), item pool rotations (e.g., Ariel et al., 2004; Way et al., 1998; Stocking & Swanson, 1998), etc. These methods, though not as widely used as the conditional or the randomization strategies in CAT, have shown to perform as well as some conditional selection strategies in certain conditions (Barrada et al., 2008).

The Randomized Item-Selection Strategies

The randomized item-selection strategies are the early attempts to control item exposure rates. Instead of always drawing the optimal items, this group of strategies incorporates a random component in the item selection process. For example, the 5-4-3-2-1 procedure of McBride and Martin (1983) (MM) selects the first item randomly from five most informative items identified by the selection algorithm. The second item is randomly selected from four most informative items, and so on. For the fifth and the rest of the items, selection is entirely based on the information criteria. Kingsbury and Zara (1989) developed the randomesque procedure (also named the Kingsbury-Zara procedure) which has the selection algorithm to identify a prespecified number of optimal items and then randomly selects one for administration. Instead of selecting among an arbitrary number of items, the .10 logits procedure of Lunz and Stahl (1998) identifies all items within .10 logits of the target difficulty level and then randomly selects one among them for administration. The Progressive method (Revuelta & Ponsoda, 1998) adds a random component to the maximum information method. The influence of the random component on item selection is gradually diminished while the importance of test information becomes more prominent as the test progresses.

The randomized procedures are simple and easy to implement. Given a sufficient number of items at each difficulty level, the procedures can perform well if the test length is not limited (Fork & Smith, 2002). Sympson and Hetter (1985) pointed out that the problem with randomized strategies is that the probability of administering an item given selection is the same for all items. Consequently, a fixed administration rate is applied for all items, which can be too high for very popular items but too low for items that are rarely selected (Davey & Nering, 2002).

The Conditional Selection Strategies

The Sympson-Hetter (1985) Procedure (SH)

Sympson and Hetter (SH) (1985) developed a procedure allowing exposure rate to vary across items. In the SH procedure, an exposure parameter, a value between zero and one, is assigned to each item in the pool. When an item is selected by the CAT selection algorithm criteria, a random number randomly drawn from a uniform [0, 1] distribution is compared with the item's exposure parameter. If the random number is less than the exposure parameter, the item will be administered. Otherwise, the next best item selected by the CAT selection algorithm criteria will be considered. Exposure parameters are obtained through a series of iterative simulations so that the maximum exposure rate is below a specified target rate. Items that are frequently administered will have a low exposure control parameter, whereas items which are infrequently administered will have a high exposure control parameter.

The van der Linden (2003) Alternatives

van der Linden (2003) addressed several disadvantages of the SH procedure. Firstly, this procedure requires iterative simulations to set values for item exposure parameters, which is very time-consuming. Moreover, if some items are added or dropped from the pool, the simulation process has to start all over again. Secondly, the iterative simulation process may not converge to a stable stage. Occasionally, it is impossible to keep all exposure rates below a reasonable target value. The researcher therefore developed several alternative formulas for the SH procedure in the step of updating item exposure parameters. Some of the alternatives appeared to overcome some of the problems.

The Stocking and Lewis (1998) Conditional Procedure (SLC)

Although the SH (1985) procedure works well for controlling the overall exposure rate of an item, it does not guarantee that the exposure rate for examinees at a given trait level is well controlled below a certain rate. Because same items are likely to be selected and administered to examinees with similar trait estimates, an item may have a high exposure rate for examinees at a given trait level even though it overall exposure rate is quite low across trait levels. This is particularly true for very easy and very difficult items, which tend to be administered whenever selected, resulting substantial test overlap among extreme examinees (Davey & Nering, 2002).

Stocking and Lewis (1998) proposed the conditional multinomial method (SLC) to control item exposure rate at individual trait levels. Instead of employing a single exposure control parameter for each item, this procedure derives a series of exposure control parameters for an item at each specific trait level. Accordingly, item exposure rates conditional on trait levels can be well controlled. The exposure control parameters are obtained through iterative simulations, which are very time-consuming.

The Davey and Parshall (1995) Procedure (DP)

The Davey and Parshall (1995) (DP) procedure calculates an exposure parameter for each item conditioned on all items previously administered to the examinee. The purpose is to reduce the chance that clusters of items appear together. As a result, the extent of test overlap can be reduced and test security may be ensured. This procedure derives a table of exposure control parameters through series of simulations, which is time-consuming, especially when the item pool is large.

The SH Procedure with Test Overlap Control (SHT)

Chen and Lei (2005) have shown that the average between-test overlap rate ($\overline{T}$) can be expressed as a linear function of the mean and variance of the item exposure rates ($\overline{r}$ and S_r^2). For a given pool-size to fixed-test-length ratio, $\hat{\overline{T}}$ can be expressed as:

$$\hat{\overline{T}} = \frac{1}{\overline{r}} \times S_r^2 + \overline{r} .$$

Accordingly, the average between-test overlap can be reduced via controlling the mean and variance of the item exposure rates. Based on this functional relationship, they developed a modified SH procedure to simultaneously control item exposure and test overlap rates. Some of the alternatives appear to be more effective than the original SH procedure.

The Shadow-Test Approach

In the shadow-test approach (van der Linden, 2000), before selecting an item for administration, a shadow test is assembled for the examinee based on the current ability estimate. Shadow tests are full-length tests that meet all the content specifications, contain all items already administered to the examinee, and are optimal at current trait estimate for the examinee. The optimal item at the trait estimate is selected for administration from the items in the shadow test which have not yet been administered to the examinee. This approach guarantees that the adaptive tests always meet the content specifications and are optimal.

In addition to content constraints, item exposure constraints can also be incorporated in this approach. van der Linden and Veldkamp (2004) proposed a method to control item exposure by imposing item-ineligibility constraints on the shadow-test assembly process. Unlike the SH procedure, this method does not need time-consuming simulations to set values for exposure control parameters prior to the operational use of the test. Instead, exposure rates are automatically set during the adaptive testing. They indicated that the idea of *a*-stratified design may also be incorporated in the shadow-test approach to improve item underexposure.

The Restrictive Maximum Information Method (*Rk*)

Revuelta and Ponsoda (1998) proposed the Restrictive Maximum Information method as a practical alternative to the SH procedure. In this method, item selection is based on the maximum information criteria, but a constraint is added that no item is allowed to be exposed in more than a prespecified percentage of the adaptive tests. Consequently, some items may be available for administration for some tests but become unavailable as their exposure rates approach the prespecified rates. Once their exposure rates drop as more tests are administered, they become available for selection again. This procedure avoids the complexities of deriving item exposure parameters for test items. They pointed out a potential problem with this method is that the ability estimation may be more precise for some examinees but less for others, depending on which tests they receive.

Stratified Strategies

The *a*-Stratified Design (STR)

Chang and Ying (1999) argued that administering highly discriminating items earlier in the test would not be most beneficial because the trait estimate at this point is just a rough guess. Because highly discriminating items tend to discriminate well only over a narrow proficiency range, it would be more advantageous to save these items for the latter stages of testing when the ability estimate is closer to its true value. In their *a*-stratified (STR) method (Chang & Ying, 1999), the item pool is divided into several strata based on item discrimination values. At each point of the test, only one stratum is active and all items are selected from it. Item selection starts from the stratum with the lowest discrimination values, and then gradually moves toward the stratum with the highest discrimination values as the test progresses. Within each stratum, selections are made so that the difficulty values of the items are closest to the estimated ability. Thus, the selection procedure avoids using highly discriminating items during the early stages of the test, leading to a more balanced item-pool usage. Chang and Ying (1999) demonstrated that the STR design can equalize item exposure rates without sacrificing the efficiency and accuracy in ability estimation. Furthermore, this test design can also help reduce test-overlap rates (Hau & Chang, 2001). It has been shown that he STR and SH procedure perform similarly when the size of the item pool is small (Leung et al., 2002).

However, STR needs to overcome the following problems before it can perform well. Firstly, in practice, the item discrimination (*a*) and difficulty (*b*) parameter values often show a positive correlation. If this is the case, this procedure would produce uneven distributions of *b*-values across strata, where some strata may not cover a wide range of *b*-values. As a result, a close match between the examinee's ability and the *b*-value of the item cannot be guaranteed during the item selection process, causing some items to be more frequently selected (i.e., high exposure rates for some items) (Chang & van der Linden, 2003). Secondly, this procedure does not guarantee that the maximum exposure rate will below a target rate, especially when there is a small ratio between the pool size and test length (Leung, Chang & Hua, 2002).

The STR Design with *b*-Blocking (BSTR)

Two modified STR designs, the STR design with *b*-blocking (BSTR) of Chang, Qian, and Ying (2001) and the 0-1 STR design of Chang and van der Linden (2003), have been developed to over the correlation issue. Both methods force more identical distributions of the *b*-values across all strata but based on different principles. In the BSTR design, the item pool is first partitioned into a number of blocks in ascending order of *b*-values, and then each *b*-block is further divided into several strata (*K*) in ascending order of *a*-values. Then, items in the K^{th} *a*-stratum from each *b*-block are grouped together to form a new stratum (i.e., the K^{th} *ab*-stratum). Items are drawn from the less discriminating strata earlier in the test and from more discriminating strata later in the test. Chang, Qian, and Ying (2001) have demonstrated that compared to the original STR design, the BSTR design provides better item exposure control, improves measurement precision, and enhances test reliability.

The 0-1 STR Design

The 0-1 STR design of Chang and van der Linden (2001) takes a different approach to address item pool stratification. Based on the technique of 0-1 linear programming (LP), this method stratifies an item pool optimally with respect to the combination of *a*- and *b*-values. Specifically, target *a*-values for each stratum as well as *b*-values within each stratum are specified first, and then items from the pool are assigned into strata such that their *a*- and *b*-values approximate their target values as closely as possible. A simulation study has shown that the 0-1 STR design outperforms STR, in terms of item exposure control. In addition, it provides improved measurement precision compared to both the original STR and BSTR designs (Chang & van der Linden, 2001). They pointed out that the 0-1 STR can be generalized to incorporate other practical constraints in adaptive testing, such as content balancing. They also addressed that a problem with the 0-1 STR design is that there are no guidelines on how to select target values. If different sets of target values are used, the optimal model could produce quite different stratification results.

Overall, the stratified procedures have the following advantages. Firstly, they automatically provide more balanced pool utilization without sacrificing proficiency and accuracy in trait estimation. Secondly, in comparison to the SH procedure, the stratified method is simpler to implement (Hua & Chang, 2001).

Multistage Adaptive Test Designs

Besides constraining the selection algorithm of CAT, item exposure can also be controlled by implementing a different type of adaptive test designs, the multistage adaptive test (MST) designs. Unlike CAT, MST is partially adaptive where testlets (sets of items), instead of single items, are administered to examinees as intact units. MST provides a compromise between the standard CAT and conventional paper-and-pencil exam and offers the benefits of permitting subtests to be pre-constructed and reviewed by specialists, allowing examinees to review and revise answers, etc (Hambleton, 2002).

Several MST (MST) designs have been developed to take into account the issue of exposure control, including the multiple forms structure (MFS) design (Armstrong & Little, 2003), the computer-adaptive sequential testing (CAST) design (Luecht & Nungester, 1998; Luecht, Nungester & Hadadi, 1996), and the adaptive multi-stage item bundles (BMAT) (Luecht, 2003). In BMAT, the item exposure mechanisms are built in the testlet pre-construction process, thereby excluding the need of any direct exposure controls (Luecht, 2003).

The administration model of the CPA Examination is a variation of the CAST design of Luecht and Nungester (1998). It is a three-stage MST with five equal-length testlets of multiple-choice questions for each panel. Multiple panels can be simultaneously constructed prior to administration to ensure equivalence in psychometric properties and reduce overexposure of highly discriminating items across panels (Melican et al, 2010). A full description of this design is available in Melican, Breithaupt, and Zhang (2010) and Breithaupt, Ariel, and Veldkamp (2004). Breithaupt and Hare (2007) showed that in this design, item exposure can be well controlled by constraining the maximum re-use rate for any subtest using mixed-integer programming (MIP) methods for automated panel assembly. However, they also demonstrated that adding a constraint of minimizing the maximum exposure of any testlet during the panel assembly process, helped little in reducing the maximum testlet exposure rate. Because testlets are not overlapping, the exposure rate of a testlet represents the actual exposure rates of items within the testlet.

One issue that has not drawn much attention in the field of item exposure in MST is the conditional exposures. Future research may explore the performance of various MST designs in controlling exposure rates at different trait levels.

Rotating Item Pools

Ariel, Veldkamp, and van der Linden (2004) pointed out that a problem with probabilistic item-exposure control methods (e.g., the SH procedure) is that they are unable to solve the problem of underexposure. They advocated the idea of designing a system of rotating item pools, which has been addressed by a number of researchers (i.e., Stocking & Swanson, 1998; Way, 1998). This method first divides the master pool into parallel sub-pools, and then rotates sub-pools to equalize item exposure rates.

Motivated by Guliksen's (1950) matched random subtests methods, Ariel, Veldkamp and van der Linden (2004) proposed four methods of constructing rotating item pools for CAT. The objective of the methods is to achieve identical distributions of item parameters across sub-pools. Meanwhile, constraints are imposed so that each sub-pool also has similar content attributes. The procedure of the methods consists of two steps. The first step is to assign items in the master pool to interim sets such that items in the same interim set are as parallel as possible. The second step is to distribute items from interim sets to sub-pools. The four methods differ in the ways that items in the master pool are distributed to different interim sets (sequential vs. simultaneous) and then to various sub-pools (random vs. mathematical programming). Using an item pool from the Law School Admission Test, Ariel, Veldkamp and van der Linden (2004) showed that all the methods succeed in controlling item exposure rates at the cost of slightly higher errors of ability estimates. The four methods are essentially equivalent in terms of their item-exposure control performance.

Two types of rotating pools can be constructed: non-overlapping pools and overlapping pools. In the case of non-overlapping pools, each item in the master pool is assigned to one and only one of the sub-pools. Therefore, different sub-pools do not share common items. In the case of overlapping pools, more popular items are assigned to a smaller number of sub-pools, whereas less popular items are assigned to a larger number of sub-pools. Consequently, some items appear in multiple sub-pools. Ariel, Veldkamp and van der Linden (2004) showed that non-overlapping pools are capable of reducing over-exposure of items but ineffective in improving under-exposure. Whereas overlapping pools are more successful in increasing under-exposure rates of less popular items. If the goal is not only to reduce over-exposure but also increase under-exposure, overlapping pools are preferred.

In Ariel, Veldkamp and van der Linden's (2004) study, only bias and RMSDs of ability estimates were examined. Little is known about the effect of their methods on test information. If the main purpose of an exam is to make pass-fail decisions, as in the case of many credentialing exams, it is critical to ensure high measurement precision near the passing score. Future research should continue to investigate the performance of pool rotation on making pass-fail decisions.

Combined Strategies

Revuelt and Javier (1998) combined the Progressive method and Restricted Maximum Information method to develop a new method named the Progressive Restricted method. As in the Progressive method, a random component is added whose influence on item selection declines as the test progresses. In addition, no item is allowed to be exposed in more than a prespecified percentage of the tests. The simulation results of their study showed that the Restricted method is useful to control overexposure and that the Progressive method is a good choice of reducing the number of unused items, both maintaining measurement precision. The combined method provided better overall results than either method, regarding the maximum exposure rate and number of unused items, as well as test precision. They concluded that the combined method may be a useful method to control item exposure without sacrificing test precision.

Leung, Chang, and Hau (2002) implemented an *a*-stratified design with a modified SH procedure (STR-SH) to deal with the problem of overexposure in stratified procedures. Through four simulated studies, they showed that STR-SH is more effective than either the SH or the STR procedure in equalizing item exposure rates, reducing test overlap rates, and controlling the number of overexposed items. They suggest that more advanced STR-SH procedures should be developed conditional on individual trait levels.

Barrada, Olea, and Abda (2008) combined the SH procedure and strategy of rotating item pools by constraining the maximum item exposure rate in the rotating pools. They demonstrated that that this combined method performed similarly to either the SH procedure alone or the rotating pool strategy, in terms of maximum exposure rate, test overlap rate and RMSEs of ability estimates. The results hold for different trait distributions (i.e., those that match the distribution of b-parameter values in the pool and those that do not match), and for item pools with correlated *a*- and *b*-parameters as well as pools with uncorrelated *a*- and *b*-parameters.

Item Exposure Controls with Content-Balancing

Researchers have integrated item exposure control strategies with certain content-balancing techniques to equalize item usage while balancing content coverage. For example, Yi and Chang (2003) proposed the a-stratified method with content blocking (STR_C) to balance content coverage in each stratum. In this method, first, an item pool is partitioned into a number of blocks according to the content specifications. Within each block, the BSTR procedure is applied and a fixed number (K) of strata are formed according to the *a*- and *b*-values. Finally, all items from the K^{th} stratum are pooled across all content blocks to form new strata. Studies have shown that implementing this procedure can balance content coverage, control item exposure, and improve item pool utilization, while maintaining measurement precision (Leung et al., 2003; Yi & Chang, 2003; see Cheng, Chang & Yi, 2007).

Recently, Doong (2009) modified the SH procedure to meet content-balancing requirements. In this procedure, the number of items to be administered from each content area is specified prior to testing. In his sequential selection example, a content area is first selected, and then items are selected from that content area. Once prespecified items have been selected from this content area, selection will move on to the next content area. According to Doong (2009), instead of sequentially selecting a content area, other selection methods, such as random selection, may also be applied in this procedure.

Item Exposure Controls for Polytomous and Testlets data

A majority of literature on item exposure controls focus on dichotomous data. Several researchers, however, have developed strategies to deal with item exposure for other item types. Davis and Dodd (2001) proposed a modified within .10 logits (modified-within-.10-logits) procedure for polytomous items. This procedure selects six items at each point of the test, two with the difficulty level matching the estimated trait of the examinee, two at the estimated difficulty level plus .10, and two at the estimated difficulty minus .10. The next item to be administered is randomly selected among them. Davis and Dodd (2005) compared this method with the randomesque procedure, SH, SLC, and a baseline condition with no exposure control. Their results showed that the modified-within-.10-logits and the randomesque procedure performed as well as the SLC procedure with Masters' (1982) partial credit model, with regard to reducing item exposure and test overlap rates, and promoting item pool utilization (David & Dodd, 2005). This is different from the results obtained using dichotomous data, which indicated better performance of the conditional selection procedures than the randomized procedures. Because the two randomized procedures are easier to implement than the conditional selection procedures, they recommended using the former two methods when dealing with data fitting a partial credit model.

Boyd, Dodd, and Fitzpatrick (2003) applied the modified-within-.10-logits procedure to handle testlet data. They compared this method with the SH procedure and a baseline condition with no exposure control using a three-parameter logistic testlet response theory model and Master's partial credit model. The results indicated that under both IRT models, the modified-within-.10-logits procedure is more able to improve item pool utilization, reduce the maximum item exposure rate and test overlap rate, while maintaining measurement precision.

Comparison Studies

Studies have been conducted to thoroughly investigate the properties of different procedures in controlling item exposure in CAT. Revuelta and Ponsoda (1998) compared eight item selection methods (the Progressive method, the Restricted Maximum Information method, Maximum Information method, One Parameter, MM, Randomesque, SH, and Random Item Selection), in both fixed-test length and variable-test length conditions. The comparisons were based on test precision, maximum exposure rates, number of unused items, and ability estimate precision. The results indicated that none of the methods provide a fully satisfactory solution to the exposure control problem. To reduce item overlap at the beginning of the test, the methods MM and Radomesque are preferred. To reduce maximum exposure rates, the SH and Restrictive Maximum Information methods are good candidate. Combining the Progressive Restricted and Maximum Information methods yielded the best overall results.

Chang and Ansley (2003) compared the performance of five item exposure methods (MM, SH, DP, the Stocking and Lewis unconditional multinomial (SL) procedure, and SLC), in regard to maximum exposure rates, test overlap rates, and conditional standard errors of measurement of trait estimates. This study also investigated the effects of item pool sizes and target maximum exposure rates on the performance of these methods. The results showed that the MM procedure did not improve test security to a noticeable extent, compared to the baseline condition with no item exposure control. The SLC procedure was more able to keep the item exposure and test overlap rates under the target values. However, measurement precision was compromised, particularly for extreme examinees. The DP procedure provided the best overall results. An interesting finding is that for the SH and SL methods, utilizing a larger item pool actually lead to higher conditional maximum exposure rates, which suggests that a large item pool itself cannot guarantee test security.

Barrada, Olea, and Abda (2008) compared the strategy of item pool rotation with the SH procedure in terms of controlling maximum exposure rates, test overlap rates, and RMSEs of ability estimates. They found that the strategy of item pool rotation slightly outperformed the SH procedure. They concluded that the method of rotating item pools should be preferred over the SH procedure, even though the latter is easier to implement. They listed the following reasons of choosing the strategy of item pool rotation over the SH procedure: (1) item selection is faster; (2) it is easier to determine the matrix of item enemies; and (3) it is easier to preserve item pool security.

Summary

This section provides a review of various item exposure control strategies for CAT. In summary, item exposure can be controlled at little cost to measurement precision. The randomized item-selection procedures and the conditional selection strategies can be used to control item overexposure. The stratification strategy is useful to improve the usage of underexposed items. To date, the SH procedure and its variants are the primary procedures adopted by testing programs or research studies (Doong, 2009). However, alternatives such as the stratified methods and rotating item pools have also shown to perform as well as or slightly better than the SH procedure in certain conditions. Future research may consider modifying existing exposure control strategies for item-level adaptive tests for practical use in MST.

Section II: Potential Applications in MST

Exposure control strategies for CAT may be modified and extended to become suitable for practical use in MST. This section attempts to propose two designs that incorporate exposure control procedures for CAT into MST for the three-stage CAST model of the CPA exams. In the first design, the STR_C procedure was combined with a MST design to equalize item exposure rates while taking content specifications into consideration. In the second design, Ariel, Veldkamp and van der Linden's (2004) strategy is proposed to use to develop overlapping pools for a master pool from the CPA exams.

The STR_C procedure and rotating item pool are chosen for this study because they have the following advantages. Firstly, both methods have limited impact on the general design of the current test assembly and delivery of the CPA exams. Secondly, they share the potential to address both overexposure and underexposure of test items, with minimum loss of measurement precision.

Design 1: STR with Content-Balancing for CAST

As mentioned earlier, the administration model of the CPA exams is a CAST design with three stages and five testlets within a panel. The difficulty levels of testlets are classified as moderate and difficult. For each panel, the first stage has only one testlet of moderate difficulty, whereas both the second and third stages have one testlet of moderate difficulty and one difficult testlet. All examinees are administered a moderate (M)-difficulty testlet at Stage 1. Next, they will be routed to take either a M or a difficult (D) testlet at Stage 2, depending on the number of questions answered correctly. At the completion of the second stage, the routing process will continue to select a M or D testlet for administration at Stage 3. Thus, each examinee is administered with three testlets in total. Adaptation occurs after completion of items within each of the first two stages.

In operational administration of the CPA exams, the test assembly method tends to select items with high *a*-values, causing an uneven distribution of item exposure rates. Because the trait estimate generally becomes more accurate as the test progresses, the a-stratified strategy can be used to control exposure rates by grouping items with similar *a* values together and then assembling testlets within the *a*-group at each stage. That is, the item pool can be stratified into three levels according to item *a*-values. Items from the lowest *a* level would be used to assemble testlets at Stage 1, those from the moderate level used at Stage 2, and those from the highest level would be used at Stage 3. Therefore, items with low, medium, and high *a*-values will be selected with equal frequency, leading to increased exposure rates of low *a* items and decreased exposure rates of high *a* items.

An alternative approach has been proposed by Jodoin, Zenisky, and Hambleton (2006), which is to use different target information functions at different stages as proposed. In one of their three-stage MST designs, low target information functions were used for Stage 1 modules while high target information functions were set for Stage 2 and 3 modules. Consequently, less discriminating items were placed at Stage 1 while more discriminating items were saved for Stage 2 and 3. In their study, the purpose of using this design, however, is not to control item exposure but to increase accuracy in ability estimation. According to Jodoin, Zenisky, and Hambleton (2006), improved ability estimation should be obtained by moving more discriminating items to the later stages when ability estimates are more accurate. Nevertheless, no differences were observed between this design and the conventional three-stage MST design, with regard to precision in ability estimation and pass-fail decisions. Based on this finding, it is expected that the use of stratification in the current study will result in more evenly distributed item exposure rates while producing comparable psychometric results.

In operational administration of adaptive tests, an important factor to consider is a balance of content coverage. In order to obtain relatively comparable test scores among examines, it is essential to ensure that MSTs will provide all examinees with required content coverage. One solution is to incorporate content constraints into the testlet assembly process so that testlets are parallel in content coverage (Ariel et al., 2006). Consequently, parallel panels can be easily built based on these content-standardized testlets. To facilitate the content-balancing process, the STR_C method can be applied in MST to ensure content coverage of each stage is similar to that of the master pool. As mentioned earlier, the STR_C takes both content specifications and the distribution of a-parameters into consideration when stratifying the item pool.

The use of STR_C method for the CPA exams can be described as follows:

1. For each content block, partition the item pool into three strata (low, medium, and high) in ascending order of a-values;
2. Assign 20% items (with low levels of discrimination) to stratum 1;
3. Assign 40% items (with medium levels of discrimination) to stratum 2;
4. Assign the rest 40% items (with high levels of discrimination) to stratum 3;
5. Group items in stratum 1 from each content block to form a new stratum 1;
6. Repeat Step 5 for stratum 2 and 3;
7. Set target test information functions (TIFs) for Stage 1 modules (M testlets);
8. Set target TIFs for Stage 2 modules (both M and D testlets);
9. Repeat Step 8 for Stage 3;

10. Assemble testlets within each stage using items from the corresponding stratum;

11. Build parallel panels.

Pilot Study

A pilot study based on the administration model of the CPA exams is offered as one application of the STR_C method for item exposure control in MST. A master pool of 3,265 multiple-choice items from the CPA exams fitting the three-parameter logistic model was used in the study. Figure 1 shows the distribution of the values for the *a*- and *b*- parameters for the items in the pool. As we can see from Figure 1, the item pool lacks of items with low *b*- values and high *a*-values, as well as items with high b-values and high a-values. Overall, the a- and *b*-values show a low correlation of .118. However, the correlation tends to be positive for items with negative *b*-values, but negative for items with positive *b*- values.

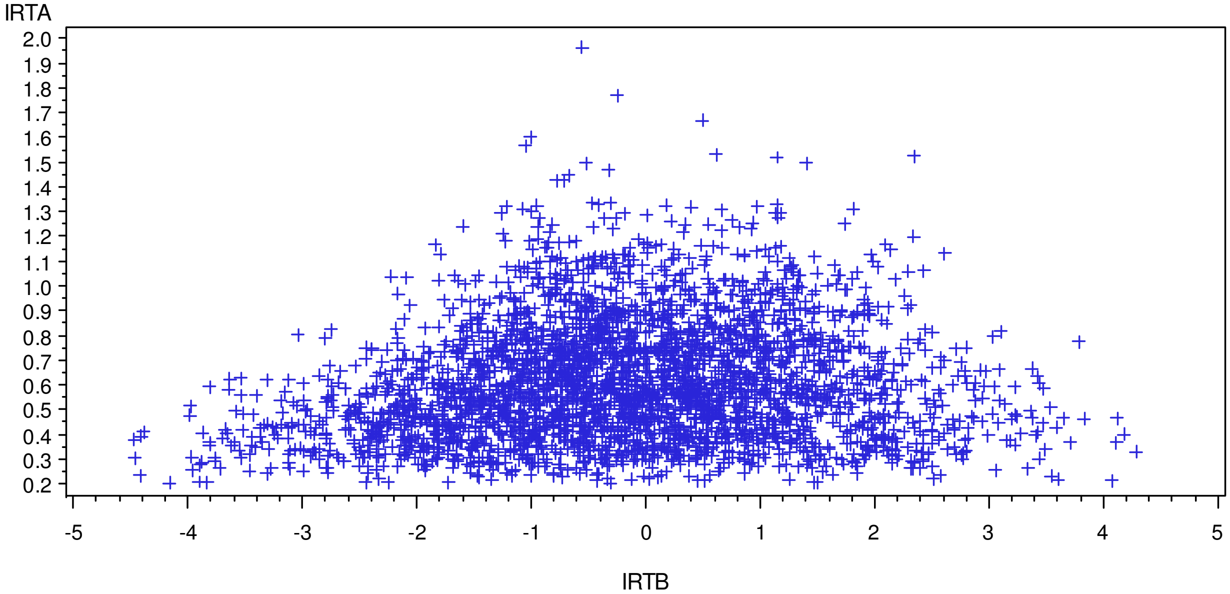
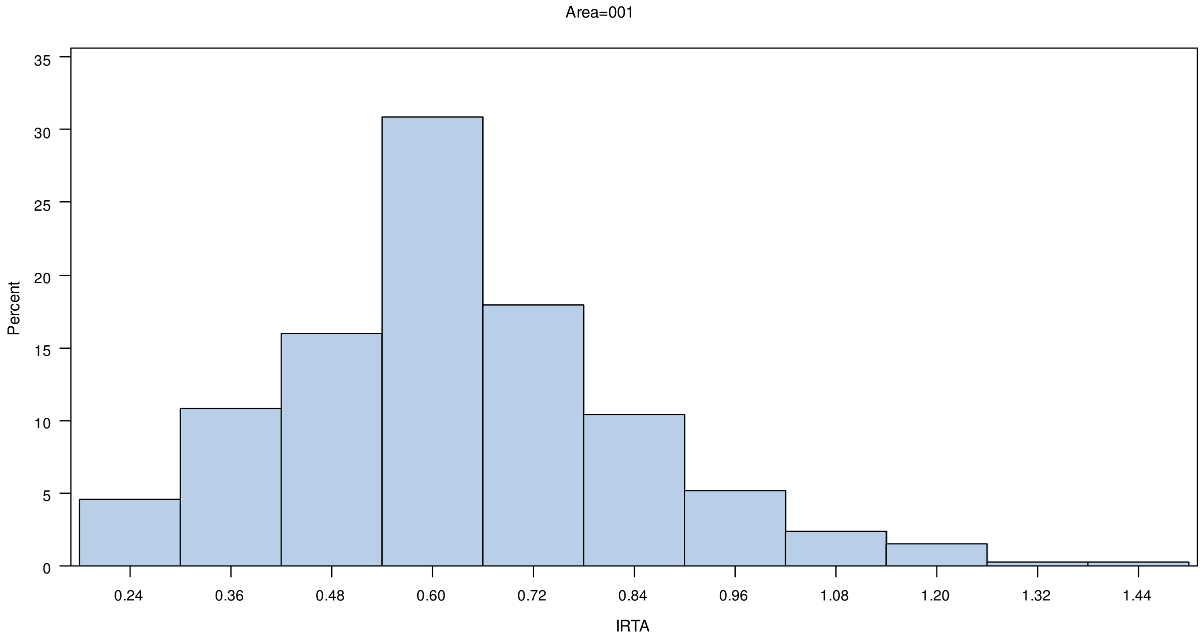
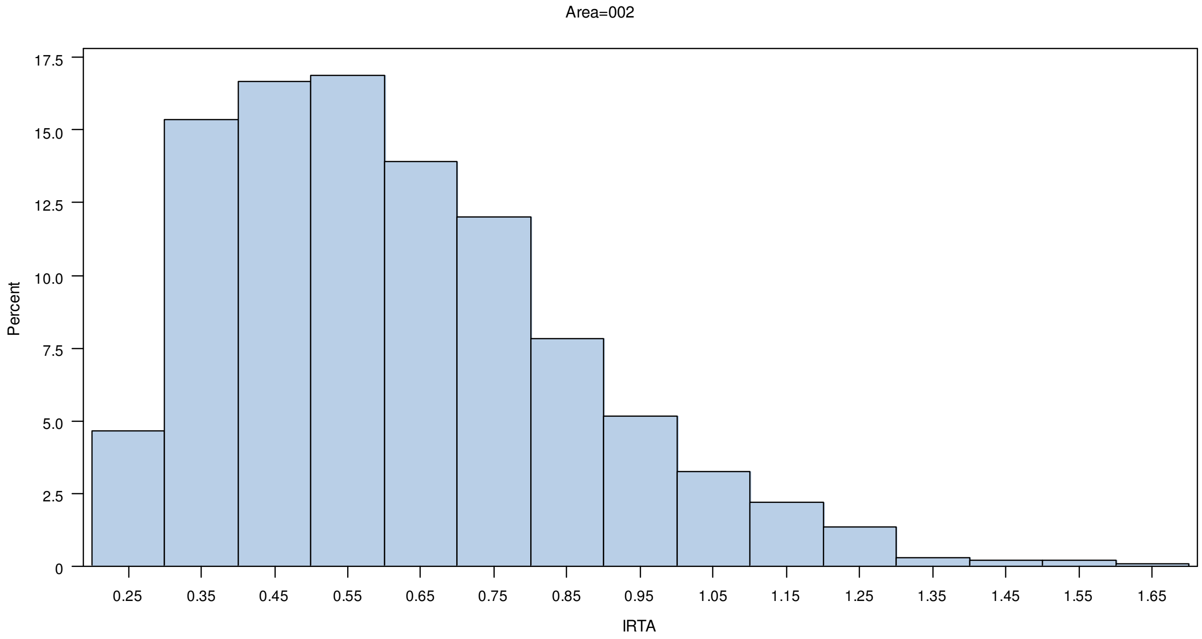
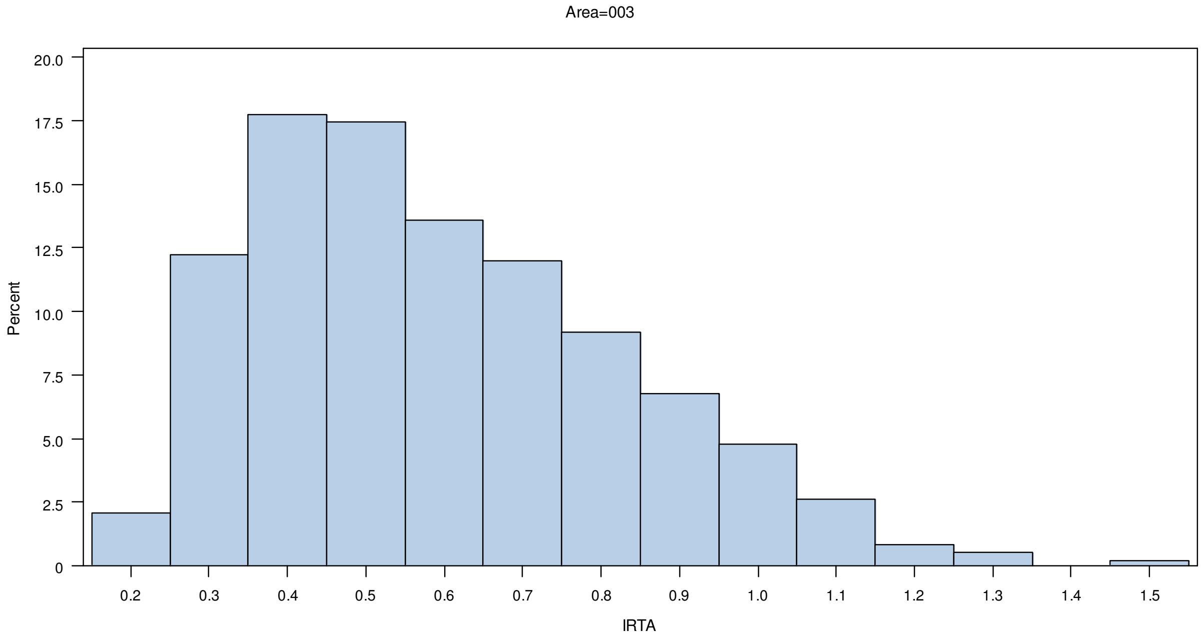


FIGURE 1. *Scatter plot of item parameters (a and b) from a master pool of the CPA exams.*

The items in the pool were of five different content types. Figures 2 illustrate the distributions of the *a*-parameter values for content area 1 through 5, respectively. As we can see from the figures, the shape of the distribution of the *a*-values differs across content areas. If stratification is merely based on the *a*-values, a balance of content across stages will not be ensured. Therefore, for this particular item pool, stratification needs to take both content and the distribution of *a*-values into consideration.




Area=002
Percent
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
1.05
1.15
1.25
1.35
1.45
1.55
1.65
IRTA


Area=003
Percent
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
IRTA

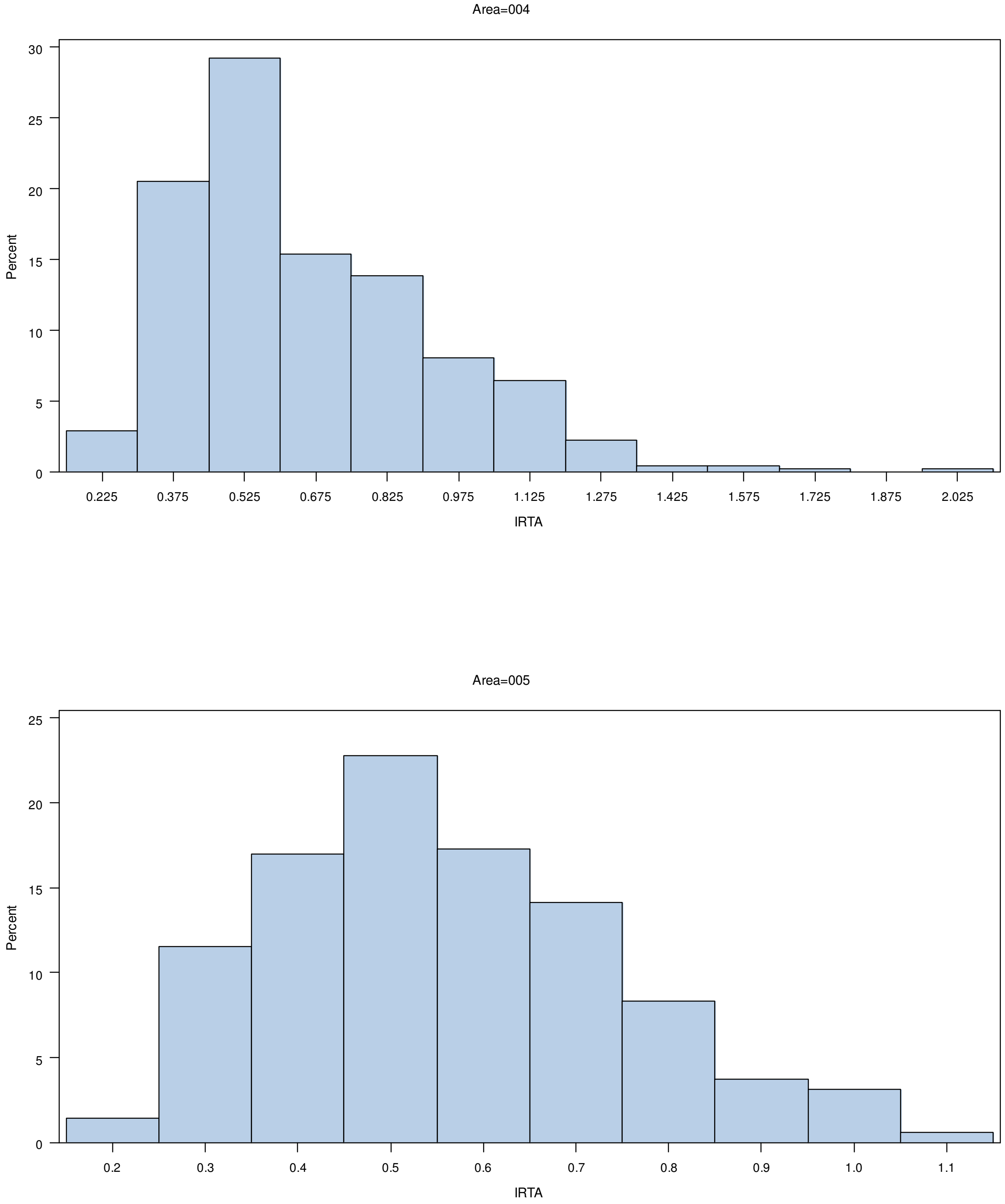


Figure 2. *Distributions of item parameters (a) from a master pool of the CPA exams.*

Figure 3 illustrates the distribution of the number of times an item is administered (cumulative exposure). Apparently, some items from this pool are overexposed whereas a large proportion of the items are never or seldom selected. Hence, to utilize the item pool well, it is necessary to control overexposure and improve underexposure at the same time. The correlation between the number of times an item is cumulative exposure and *a*-values is .378, indicating that items with larger *a*-values tend to be more frequently administered than those with smaller *a*-values. Given that some items from the pool are new and have been used for a short period of time, the true correlation value should be even higher. The stratification strategy, therefore, should be effective in controlling exposure rates for this item pool.

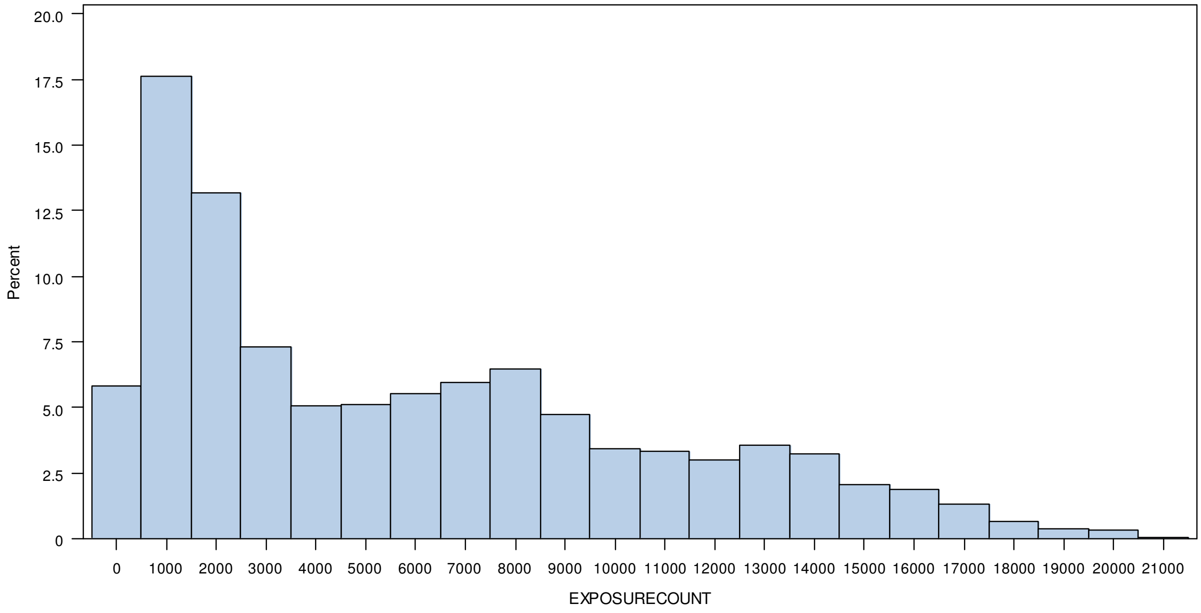


FIGURE 3. *Distributions of cumulative item exposure from a master pool of the CPA exams.*

Table 1 presents descriptive statistics of item parameters by stage obtained from stratification. As expected, the mean discrimination level increases from Stage 1 to 2, and 3. The mean difficulty level also slightly increases from Stage 1 to 3. The range of the difficulty parameter does not differ much between Stage 1 and 2, but becomes slightly narrower at Stage 3. Overall, the stratification results are satisfactory. In the next step, testlets and panels will be assembled using the stratified item pool.

Table 1: Descriptive Statistics of Item Parameters by Stage for Three-Stage Multistage Test

Item Parameter	**Stratum**	**N**	**Mean**	**SD**	**Min**	**Max**
a						
	1	655	.334	.053	.203	.463
	2	1306	.523	.071	.388	.649
	3	1304	.844	.176	.590	1.963
b						
	1	655	-.359	1.716	-4.467	4.289
	2	1306	-.266	1.422	-4.384	4.186
	3	1304	.016	1.128	-3.536	3.791

Issues to Resolve

Several issues need to be addressed before this method can be effectively implemented in the CPA exams. The first is to derive feasible statistical information targets for M and D testlets at each stage. Since in the proposed Design 1, testlets are assembled within each stage using items from the corresponding *a* stratum rather than being constructed simultaneously from the whole item pool, the simultaneous test assembly method currently used for the CPA exams can no longer be applied. Instead, new target TIFs need to be set for each stage. This step is critical because the choice of target TIFs will impact classification decisions and item exposure rates (Ariel, et al., 2006). Next, one may suspect that a higher likelihood of routing errors would occur as a result of using less informative items at Stages 1. However, if an incorrect Stage 2 testlet is chosen, it is still possible that a routine will be recovered at Stage 3. One follow-up study that is planned is to investigate the effect of stratification on routing and classification decisions. Another concern is fatigue as highly discriminating items are all administered at Stage 3.

Design 2: Item Pool Rotations

The second design applies Ariel, Zenisky, and Hambleton's (2004) strategy of constructing and rotating item pools. As mentioned earlier, rotating pools can be developed as overlapping or nonoverlapping. In overlapping pools, some items appear in several pools. While in nonoverlapping pools, each item from the master pool can only be assigned to one of the pools. In this study, overlapping pools will be constructed because the master pool exhibits both underexposure and overexposure problems.

Ariel, Zenisky, and Hambleton (2004) have proposed different methods to ensure that sub-pools constructed show similar distributions of item parameters and content coverage. Their method of constructing overlapping pools consists of two stages. In the first stage, the items of the master pool are assigned to a number of interim sets that are closely parallel. For example, items can be assigned into interim sets such that the differences between their parameter values within each interim set are minimized.

$$\delta_{ij} = \left|a_i - a_j\right| + w\left|b_i - b_j\right| \tag{1}$$

where δ_{ij} is a metric representing the differences between items *i* and *j* in interim sets and *w* is a parameter to adjust for differences between the scales of the two parameters.

A 0-1 programming method can be used to assign items into interim sets. The objective function is to minimize the values of δ_{ij} for all possible combinations of items within the same interim set

$$\min \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \delta_{ij} x_{ij} \tag{2}$$

where x_{ij} are decision variables, which are equal to 1 if item *i* and *j* are chosen in the same set and are equal to 0 otherwise, and *I* represents the number of items in the master pool. A constraint is added to require that every item is assigned to an interim set only once

$$\sum_{i|i<j} x_{ij} + \sum_{i|i<j} x_{ji} = 1, \forall j. \tag{3}$$

In the second stage, items from the interim sets are distributed to sub-pools. The objective function is to minimize the differences between the total information of the sub-pools at certain ability levels (θ_k). The model can be expressed as:

$$min\ Z, \tag{4}$$

subject to

$$\sum_i I_i(\theta_k) y_{is} - \sum_j I_j(\theta_k) y_{jp} \leq z, \forall \theta_k, s, p, (s \neq p, i \neq j) \tag{5}$$

$$\sum_i I_i(\theta_k) y_{is} - \sum_j I_j(\theta_k) y_{jp} \geq -z, \forall \theta_k, s, p, (s \neq p, i \neq j) \tag{6}$$

where y_{ij} and y_{ip} are decision variables, which are equal to 1 if item *i* and *j* are assigned to pool *s* or *p*, respectively, and are equal to 0 otherwise. Again, constraints are added to guarantee items in the same interim set to be assigned into different sub-pools

$$\sum_{i \in Q_r} y_{is} = 1, \forall_S \tag{7}$$

$$\sum_{s_r} y_{is} \leq n^{(r)}, \forall i \tag{8}$$

$$\sum_{s_r} y_{is} \geq n_r, \forall i \tag{9}$$

$$y_{is} \in \{0,1\}, \tag{10}$$

where Q_r is the r_{th} interim set. Equation (8) and (9) set constraints on the maximum ($n^{(r)}$) and minimum number of times (n_r) an item can be assigned to a pool. Popular items should be assigned to a smaller number of pools while unpopular items should be assigned to a larger number of pools.

Additional content constraints can be added into the model to ensure that overlapping pools have comparable distributions of content:

$$\sum_{i \in V_m} y_{is} \geq n_m, \forall s \tag{11}$$

$$\sum_{i \in V_m} y_{is} \leq n^{(m)}, \forall s \tag{12}$$

where V_m represents the item set that belongs to content *m*. Equation (11) and (12) set the lower (n_m) and upper bounds ($n^{(m)}$) of the number of item type *m*.

Once the overlapping pools have been constructed, the current automated test assembly (ATA) method can be used to build panels. For each testing window, only one sub-pool is used for test assembly. Compared to Design 1, this method does not require modifications to the current ATA method. A disadvantage of this method is that the process of constructing rotating pools can be time-consuming. Moreover, in this design, multiple optimization models need to be built, such as formulating the problem of assigning items to interim sets as a 0-1 mathematical programming problem. These optimization models can be developed in commercial application such as OPL Studio. Open-source software such as R with lp_solve may also be useful for solving these problems.

Section III: Concluding Remarks

Item exposure is an important issue in computer-based testing. The advantages of adaptive testing over the conventional paper-and-pencil testing may not sustain if item exposure is not well controlled. The first part of the study briefly described and evaluated different exposure control methodologies for item-level adaptive tests. The second part of the study proposed two designs that extend exposure control procedures for CAT to MST. In the first design, the *a*-stratified method with content blocking (STR_C) procedure (Yi and Chang, 2003) is combined with the MST model of the CPA exams. A pilot study on the stratification based on a master pool from the CPA exams was presented. The second design proposed to use the method of item pool rotations (Ariel, et al., 2004) to construct nonoverlapping pools for a master pool from the CPA exams.

There are several limitations associated with the proposed designs. For Design 1, fatigue and routing errors are potential concerns. For design 2, the process of constructing rotating item pools can be time-consuming. Furthermore, a remaining issue to resolve for Design 1 is to derive feasible target TIFs for each MST stage.

This is a preliminary study and more work needs to be done to systematically investigate the performance of the proposed methodologies on item exposure control in the MST context. Follow-up studies are being planned for implementation and comparison of proposed methodologies using the CPA exams data.

REFERENCES

Armstrong, R., & Little, J. (2003, April). *The assembly of multiple form structures*. Paper presented at the 2003 annual meeting of National Council of Measurement in Education, (NCME), Chicago, IL.

Ariel, A.,Veldkamp, B. P., & Breithaupt, K. (2006). Optimal Testlet Pool Assembly for Multistage Testing Designs. *Applied Psychological Measurement, 30*(3)*,* 204-215.

Ariel, A.,Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing Rotating Item Pools for Constrained Adaptive Testing. *Journal of Educational Measurement, 41*(4)*,* 345-359.

Barrada, J. R., Olea, J, & Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. The Spanish Journal of Psychology. *11*(2)*,* 618-625.

Boyd, A. M., & Dodd, B. G.; Fitzpatrick, S. J. (2003). A comparison of exposure control procedures in CAT systems based on different measurement models for testlets using the Verbal Reasoning Section of the MCAT.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67(1),* 5-20.

Breithaupt, K., Ariel, A. & veldkamp, B. P. (2004). Balancing item exposure and optimality in automated assembly for multistage testing. AICPA technical report.

Chang, H., Qian, & Ying, Z. (2001). *a*-Stratified multistage adaptive testing with *b*-blocking. *Applied Psychological Measurement, 25*(4), 331-341.

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

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

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

Chen, S., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *4* (2), 149-174.

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

Chen, S., & Lei, P. (2005). Controlling item exposure and test overlap in Computerized Adaptive Testing. *Applied Psychological Measurement, 29*(3)*,* 204-217.

Cheng, Y., Chang, H, & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467-482.

Davis, L. L., & Dodd, B.G. (2005). *Strategies for controlling item exposure in computerized adaptive testing with the partial credit model.* Technical Report. Pearson Educational Measurement.

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

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

Davey, T., & Nering, M. (2002). Controlling item exposure & maintaining item security. In Mills, Potenza, Fremer & Ward (Eds.), *Computer-Based Testing: Building the foundations for future assessments* (pp. 165-192). London: Lawrence Erlbaum Associates.

Doong, S. H. (2009). A knowledge-based approach for item exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 34*(4), 530-558.

Fork, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In Mills, Potenza, Fremer & Ward (Eds.), *Computer-Based Testing: Building the foundations for future assessments* (pp. 41-66). London: Lawrence Erlbaum Associates.

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

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K. (2002). New CBT technical issues: Developing items, pretesting, test security, and item exposure. In Mills, Potenza, Fremer & Ward (Eds.), *Computer-Based Testing: Building the foundations for future assessments* (pp. 193-204). Mahawah, NJ: Lawrence Erlbaum Associates.

Hua, K., & Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Hua, K, Wen, J, & Chang, H. (2002). Optimal number of strata in the a-stratified computerized adaptive testing. *Paper presented at the American Educational Research Association Conference, New Orleans.*

Jodoin, M. G, Zenisky, A., & Habmbleton, R. K. (2006). Comparison of the Psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220.

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

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

Leung, C, Chang, H., & Hau, K. (2002). Item selection in Computerized Adaptive testing: Improving the a-stratified design with the Sympson-Hetter Algorithm. *Applied Psychological Measurement, 26(4),* 376-392.

Luecht, R. M. (2003, April). Exposure control using adaptive multi-stage item bundles. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). Heuristics based CAT: Balancing item information, content, and exposure. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

Lunz, M. E., & Stahl, J. A. (1998). *Patterns of item exposure using a randomized CAT algorithm*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Master, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Melican, G. J., Breithaupt, K. & Zhang, Y. (2010). Designing and implementing an multistage adaptive test: The Uniform CPA Exam. In van der Linden et al. (Eds.) Elements of Adaptive Testing, (pp. 167-189). New York: Springer.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: latent trait test theory and computerized adaptive testing,* (pp. 224-236). New York: Academic Press.

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

Revuelta, J., & Ponsoda, V.. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35* (4), 331-327.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, *23*(1), 57-75.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks 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.* In Proceedings of the 27th annual meeting of the Military Testing Association, (pp. 973-977). San Diego CA: Navy Personnel Research and Development Centre.

van der Linden, Wim J. (2006). Assembling a computerized adaptive testing item pools as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31*(1), 81-99.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*(3), 249-265.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-292.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and Practice,* (pp. 27-52). Norwell MA: Kluwer.

Wainer, H. (2000). Computerized adaptive testing: A primer (2nd ed.) Mahawah, NJ: Lawrence Erlbaum Associates.

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

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