Georgiadou, E., Triantafillou, E. & Economides A.A. : A review of item exposure control strategies for Computerised Adaptive Testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment.* Vol. 5, No 8, May 2007.

J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 5, Number 8 · May 2007

A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005

Elissavet Georgiadou, Evangelos Triantafillou, & Anastasios A. Economides

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 5, Number 8

A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005

Elissavet Georgiadou, Evangelos Triantafillou, & Anastasios A. Economides

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved [date] from http://www.jtla.org.

Abstract:

Since researchers acknowledged the several advantages of computerized adaptive testing (CAT) over traditional linear test administration, the issue of item exposure control has received increased attention. Due to CAT's underlying philosophy, particular items in the item pool may be presented too often and become overexposed, while other items are rarely selected by the CAT algorithm and thus become underexposed. Several item exposure control strategies have been presented in the literature aiming to prevent over-exposure of some items and to increase the use rate of rarely or never selected items. This paper reviews such strategies that appeared in the relevant literature from 1983 to 2005. The focus of this paper is on studies that have been conducted in order to evaluate the effectiveness of item exposure control strategies for dichotomous scoring, polytomous scoring and testlet-based CAT systems. In addition, the paper discusses the strengths and weaknesses of each strategy group using examples from simulation studies. No new research is presented but rather a compendium of models is reviewed with an overall objective of providing researchers of this field, especially newcomers, a wide view of item exposure control strategies.

J·T·L·A

A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005

Elissavet Georgiadou
Evangelos Triantafillou
Anastasios A. Economides

Introduction

In linear tests, either paper-and-pencil or computerized, all examinees are presented with all the items of a particular form of the test. In order to avoid the phenomenon of administering the same test over time, test developers need to frequently revise tests, which can be a time and money consuming effort. An alternative approach to minimize familiarity with the test is to select items from a larger supply in order to provide examinees with an almost unique configuration of randomized questions. However, in linear tests, once items have been selected they remain constant for a given form of test. In computerized adaptive testing (CAT) the approach is different since the items are administered in real time and are sequentially selected according to the examinee's ability level (θ). During the testing process each examinee's ability level can be estimated by the computer and one item at a time is adaptively selected and, in effect, tailored to the estimated level. The computer ensures that the items are not too simple or too difficult for each examinee. Each examinee's responses to items are recorded during the test and a regularly updated estimate of the examinee's ability is maintained. As a result examinees are presented with individualized versions of the *same* test. Several advantages are associated with CAT's administration such as enhanced measurement precision, better test security, and shorter test lengths due to administration of more informative items (Wainer, 2000).

The following four components are needed for developing computer adaptive tests: a) a pool of items to select from; b) a criteria for selecting items; c) a method for scoring the test; and d) a decision of when the test is finished or a stopping rule (Green, Bock, Humphries, Linn, & Reckase, 1984). The item pool or item bank is an accumulation of the test items. It consists of all the items that may be administered during the test and the items' parameters. In other words item pools are files of various suitable test items that are "coded by subject area, instructional level, instructional objective measured, and various pertinent item characteristics (e.g., item difficulty and discriminating power)" (Gronlund, 1998, p. 130). Moreover, item pool files usually include details of the history of the items development, use and re-calibration (Linacre, 2000). The item parameters included in the pool are dependent upon the Item Response Theory (IRT) model selected to model the data and to measure the examinees' ability levels. In an ideal item pool there will be enough items to generate multiple test forms for a range of examinee abilities (Davey & Nering, 2002). "The better the quality of the item pool, the better the job the adaptive algorithm can do. The best and most sophisticated adaptive program cannot function if it is held in check by a limited pool of items, or items of poor quality" (Flaugher, 2000, p. 38). Wise (1997) suggests that the quality of the item pool can be conceptualized according to two basic criteria: a) the total number of items in the pool must be sufficient to supply informative items throughout a testing session; and b) the items in the pool must have characteristics that provide adequate information at the proficiency levels that are of greatest interest to the test developer. Moreover, the integrity of the CAT is dependent upon the item parameters remaining unchanged.

A widely used strategy for selecting an examinee's next item from the item pool, given a provisional estimate of θ based on preceding responses, is the maximum information method (Thissen & Mislevy, 2000, p. 109). This method selects the unused item of the pool that provides the most information at the last estimated ability. However, if items are selected only to maximize the information in the ability estimator, test content may easily become unbalanced for some ability levels, or unfair for certain groups of examinees such as minority groups (van der Linden, 2000). van der Linden argues that "adaptive testing will only be accepted if the statistical principal of adapting the item selections to the ability estimates for examinees is implemented in conjunction with serious consideration of many other non statistical test specifications" (van der Linden, 2000, p. 28).

A major consideration is the exposure control of items. It refers to constraining the administration of more popular items that would otherwise become compromised due to repeated administrations. Due to CAT's underlying philosophy particular items in the item pool may be presented too often and become overexposed, while other items are rarely selected by the CAT algorithm and thus become underexposed. In the case of overexposed items examinees may become familiar with them and prepare for them, resulting in a decrease in the items' actual difficulty, which in sequence would positively bias proficiency estimation and thus decrease the test's validity. For the underexposed items, Revuelta and Ponsoda (1998) argue that from the test developer's perspective, it is undesirable to create a large item pool and then use an item selection method that leaves unexploited a large percentage of the items; rather, all the items should be administered sometimes. This also guarantees more variety in the items the examinees receive. In short, all the items from the item pool should be used for economic reasons while no item should be overused for security reasons.

Davey and Parshall (1995) argue that CAT should aim to maximize test efficiency by selecting the most appropriate items for each examinee and to guarantee that the test measures the same composite of multiple traits for each examinee through administration of items with the same content properties. A precondition for the achievement of these goals is to protect the security of the item pool. Test security is very important, especially for large scale, high-stakes tests that are offered on a continuous basis. In such a case there is a risk that many items may become known to examinees before the actual test dates, as examinees who had already taken the test may share information with examinees who will take the tests at a later date.

The efforts that have been made to prevent some popular items from being overly exposed to examinees focus mainly on management of item pools (e.g. expanding the number of test items, rotating item banks etc.) and on the incorporation of exposure control strategies/procedures into the item selection process. However, a very different approach to address the problem is to detect exposed test items through the generation and investigation of item statistics that can reveal whether test items have become known to candidates prior to seeing the items in the test they are administered (Han & Hambleton, 2004).

The present paper focuses on the item exposure control strategies. Several variables are associated with the control of item exposure, such as precision of measurement, exposure rate, pool utilization, and test overlap. Precision of measurement refers to the degree that the CAT system with exposure controls estimates examinees' abilities when compared to the examinees' known abilities; exposure rate refers to the number of times an item is administered to the total number of CATs administered; pool utilization corresponds to the percentage of items not administered throughout any of the CAT administrations; and test overlap refers to the number of common items amongst the examinees. Since CAT became a popular type of assessment, several strategies have been developed to control item exposure. These strategies seek to prevent overexposure of some items and to increase the use rate of seldom or never-selected items (Revuelta & Ponsoda, 1998). This paper attempts to review the item exposure control strategies that appeared in the relevant literature from 1983 to 2005. Moreover, the paper reviews studies that have been conducted in order to evaluate the effectiveness of such strategies for dichotomous scoring, polytomous scoring and testlet-based CAT systems. Rather than presenting new research in the current paper, a compendium of models is reviewed with the objective of providing researchers of this field, especially newcomers, a wide view of the item exposure control strategies landscape.

Item Exposure Control Strategies

Stocking (1993) classifies item exposure control strategies in two groups: a) methods adding a random component to the maximum information item selection method, and b) methods based on assigning a parameter of each item to control its maximum exposure. Similarly, Way (1998) categorizes exposure control strategies into randomization and conditional selection. As the research field progresses, new strategies appear beyond this classification. Chang and Ying (1999) propose a multistage a-stratified design that partitions items into several strata in an ascending order of item discrimination. Items with low discrimination are presented first and while more precise estimations of examinees' ability levels are determined, items with high discrimination are administered. In an attempt to control item exposure, other researchers combine different strategies together, for example, randomization with conditional selection; *a*-stratified design with conditional selection; shadow test approach with *a*-stratified design etc. Furthermore, other strategies aim to control exposure a priori by pre-constructing adaptive test forms such as the *Computerized Adaptive Sequential Testing* (CAST) developed by Luecht and Nungester (1998). Next, the paper examines all these different strategies under the following classification: a) randomization strategies; b) conditional selection strategies; c) stratified strategies; d) combined strategies; and e) multiple stage adaptive test designs. Table 1 (next page) presents all the strategies under each category in order to assist the reader.

Table 1: Item Exposure Control Strategies from 1983 to 2005

Strategy Type	Reference
Randomization	
5-4-3-2-1 strategy	McBride and Martin, 1983
Randomesque strategy	Kingsbury and Zara, 1989
INFO4 Procedure	Thomasson and Drasgow, 1990
Within .10 Logits strategy	Lunz and Stahl, 1988
Progressive strategy	Revuelta and Ponsoda, 1998
Conditional Selection	
Sympson-Hetter (SH) strategy	Sympson and Hetter, 1985
Extended SH strategy	Stocking, 1993
Davey and Parshall strategy	Davey and Parshall, 1995
Stocking and Lewis Multinomial strategy	Stocking and Lewis, 1995
Restricted Maximum Information strategy	Revuelta and Ponsoda, 1998
SH Conditional Procedure strategy	Chang, 1998
Stocking and Lewis Conditioning on Estimated Ability	Stocking and Lewis, 2000
Targeted Exposure Control strategy	Thompson, 2002
Chen and Lei strategy	Chen and Lei, 2005
Shadow Test approach	van der Linden and Veldkamp, 2005
Stratified Strategies	
a-Stratified strategy (*a*-STR)	Chang and Ying, 1999
a-STR with Freezing	Parshall, Harmes and Kromrey, 2000
a-STR with *b*-Blocking	Chang, Qian and Ying, 2001
a-STR CAT with Unequal Item Exposure across Strata	Deng and Chang, 2001
a-STR CAT Design with Content Blocking	Yi and Chang, 2001
Multi-dimensional Stratification	Lee, Ip and Fuh, 2002
0-1 Stratification strategy	Chang and van der Linden, 2003
Combined Strategies	
Progressive Restricted strategy	Revuelta and Ponsoda, 1998
Nering, Davey and Thompson Hybrid strategy	Nering, Davey and Thompson, 1998
Eggen's strategy	Eggen, 2001
Incorporation of the SH into *a*-STR with Content Blocking	Yi, 2002
Combination of the *a*-STR with the SH strategy	Leung, Chang and Hau, 2003
Content Constraints in *a*-STR CAT using a Shadow Test	van der Linden and Chang, 2005
Multiple Stage Adaptive Test Designs	
Computerized Adaptive Sequential Testing	Leucht, Nungester and Hadadi, 1996; Leucht and Nungester, 1998
Adaptive Multi-stage Item Bundles	Leucht, 2003
Multiple Forms Structures	Armstrong and Little, 2003

Randomization Strategies

All strategies under the Randomization category attempt to control the frequency of item administration by randomly selecting an item for administration from a group of several items near the optimal level of maximum information. The assumption underlying these strategies is that after some numbers of initial items, examinees will be sufficiently differentiated so that subsequent items will vary significantly. Randomization Strategies include: a) 5-4-3-2-1 Strategy (McBride & Martin, 1983); b) Randomesque Strategy (Kingsbury & Zara, 1989); c) INFO4 Procedure (Thomasson & Dragsow; see Segall 1994 and Stocking & Lewis 1995a); d) Within .10 Logits Strategy (Lunz & Stahl, 1998); e) Progressive Method (Revuelta & Ponsoda, 1998).

5-4-3-2-1 Strategy

McBride and Martin (1983) attempt to increase item security by indirectly reducing the occurrence of an item. They develop the *5-4-3-2-1* algorithm that uses a randomization scheme to prevent the overexposure of initial items. The 5-4-3-2-1 method they propose selects the first item for administration randomly from the five most informative items. The second item is randomly selected from the four most informative items. This process is continued such that the third and fourth items are randomly selected from the three and two most informative items respectively, until the fifth item. The fifth and subsequent items are as selected to be optimal at the examinee's current updated ability level. The initial selection of five items is arbitrary. As a result the most informative item at a current ability estimate at the early testing process is not always administered.

Randomesque Strategy

Kingsbury and Zara's (1989) propose a strategy similar to the 5-4-3-2-1. The selection of the item is always made at random among the most informative items; However, the *Randomesque* strategy is different from the 5-4-3-2-1 strategy in that it repeatedly selects the same number of the most informative items (e.g. 2, 3, 4, ..., 10) from which one is randomly selected for administration throughout testing and does not switch to maximum information selection at anytime. Kingsbury and Zara (1989) suggest that continuing the randomization technique throughout testing will decrease the overlap in items seen by examinees of similar abilities.

INFO4 Procedure

Thomasson and Dragsow develop the *INFO4 Procedure* (as described in Segall 1994 and referenced in Stocking & Lewis 1995a). In the application described, at every item selection the items in the entire pool are ordered from highest to lowest based on their Fisher information at the current level of estimated ability. These values are then raised to the fourth power. A maximum is placed on these values, the values are then normalized to a sum of one and a cumulative function is formed. A random number is generated and the location of the corresponding item is found for the value of the random number, interpreted as a cumulative probability. This item then becomes the next item to be administered. The INFO4 procedure avoids the problem of determining the best sequence of group sizes that characterizes the simple randomization method. It is similar to the simple randomization approach with randomization at every item selection. Intrinsic to this method is also the implicit dependence of the randomization on the current estimated ability level. However, this procedure depends on the nature of the particular item pool for which it was developed and it may be difficult to generalize to other pools. Nevertheless, this exposure control method was not developed further and no research reports are available (personal communication with F. Drasgow, 4/27/2006).

Within .10 Logits Strategy

Lunz and Stahl (1998) suggest choosing from all items within a certain distance of the target difficulty value rather than choosing from a fixed number of items. They develop the *Within .10 Logits* strategy to examine the number and pattern of items that overlapped across examinees with similar abilities. Their strategy switches the focus of item selection from information to item difficulty because the Rasch model was used and therefore the information and item difficulty yield the same item selection. This procedure randomly selects an item from all items within .10 logits of the desired difficulty level. Therefore all items within the specified range are available for selection rather than an arbitrary number of items. If there are no items available within this range, the item with the closest difficulty level is administered. This procedure is continued throughout testing. Lunz and Stahl (1998) observe a decrease in common items when examinee abilities were different and a decrease in the mean percent of common items across examinees the larger the item pool.

Progressive Strategy

Revuelta and Ponsoda (1998) propose the *Progressive* strategy. It adds to the maximum information method a random component, whose contribution is important at the beginning of the test and gradually less influential as the test progresses. They suggest that when applying the maximum information method, the contribution to the test precision of the initial items is seldom great, since these items are very informative but for ability estimates that very often differ markedly from the final estimates. To remedy this problem they propose the progressive method in order to reduce differences among items in item exposure rate, without producing a serious loss in precision if the random component affected mainly the initial item selections. As the testing session progresses and provisional ability estimates approach final estimates, the information component is gaining the importance that the random component is losing.

Conditional Selection Strategies

Most strategies under this category control the exposure rate of an item through the *exposure control parameter*, which dictates the probability of administering the item, given it is selected. In advance of testing, a desired maximum value (*r*) is specified. Provided that an item has been selected, whether to administer this item to the examinee depends upon the exposure control parameter of the item. The exposure control parameters for the very popular items could be as low as the pre-specified desired exposure rate, indicating that these items cannot be liberally administered when they are selected. For the items that seldom appears, the associated exposure control parameters could be as high as 1.0, meaning that these items are almost always presented once they are selected. The values of these parameters are determined from a series of iterative multifaceted simulations prior to operational use, using all items in the pool that may have to be repeated as test conditions change.

Conditional Selection Strategies include: the Sympson-Hetter strategy (Sympson & Hetter, 1985), Extended Sympson-Hetter Strategy (Stocking, 1993), Davey-Parshall Strategy (Davey & Parshall, 1995), Stocking and Lewis Multinomial Strategy (Stocking & Lewis, 1995a, 1995b), Restricted Maximum Information Strategy (Revuelta & Ponsoda, 1998), Sympson-Hetter Conditional Procedure (Chang, 1998), Stocking and Lewis conditioning on estimated ability (Stocking & Lewis, 2000), Targeted Exposure Control Strategy (Thompson, 2002), Chen and Lei Strategy (Chen & Lei, 2005), and van der Linden and Veldkamp Shadow Test Approach where item-exposure control is implemented by imposing item-ineligibility constraints on the assembly process of the shadow tests (van der Linden & Veldkamp, 2005).

Sympson-Hetter (SH)

Sympson and Hetter (1985) develop a probabilistic method to deal with the issue of controlling item exposure rates. Their strategy distinguishes between the probability P(S) that an item is selected as the most informative item to administer for an examinee randomly sampled from a typical group of examinees, and the probability P(A/S) that an item is administered given that it has been selected. The strategy seeks to control the overall probability that an item is administered P(A), where P(A)=P (A/S)* P(S), and to insure that the maximum value of P(A) for all items in the pool is less than some value r that is the wanted maximum rate of item usage. The exposure control parameters P(A/S)=K are determined over a series of iterative simulations of a test design, with a sample drawn from a typical distribution of abilities. This iterative process results in each item i being assigned an exposure control parameter (k_i) with a value between zero and one. Items with low exposure rate will have k_i values close to one, and those with extremely high exposure rate will have smaller k_i values. These parameters are then used in live testing to limit the probability of administering an item. When an item is selected for administration by an item selection strategy, its exposure control parameter must be compared to a random number between zero and one, drawn from a uniform distribution. If the random number is less than or equal to the exposure control parameter (k_i) for the selected item, the item is administered. If it is not, then the item is blocked from further administration and the next most informative item is selected for consideration. This process continues to the point that an item is administered.

Extended Sympson-Hetter Strategy (ESH)

Stocking (1993) extends the SH strategy to item pools with complex structures and adaptive tests with complex test specifications. The basic procedure is applied to block of items as well as to stimulus material that will have different exposure rates than items associated with stimulus material. The advantage of this approach is that one obtains direct control of probability that an item is administered P(A) in a typical population of examinees.

Davey-Parshall Strategy (DP)

Davey and Parshall (1995) extend the SH strategy to prevent not only individual item overexposure, but also to minimize the extent to which item clusters appear together. Similarly to the SH strategy, the *Davey-Parshall* (DP) strategy requires setting exposure control parameters through simulations prior to live testing. However, each item's exposure control parameter is conditioned on all other items previously administered to the examinee. The DP strategy is based on a table of exposure control parameters where the diagonal elements contain unconditional exposure control parameters, similar to those of the SH strategy. The off-diagonal elements represent the conditional parameters that control the frequency with which pairs or clusters of items appear together given selection. Davey and Parshall (1995) suggest that their strategy reduces the extent to which item overlap across tests administered for examinees with similar abilities and for examinees with differing abilities.

Stocking and Lewis Multinomial Strategy

In order to overcome the practical disadvantages of time-consuming simulations and the dependence of the exposure control parameters on the distribution of examinee ability level used in the simulation of the SH and the *Extended Sympson-Hetter* (ESH) strategies, Stocking and Lewis (1995a) propose a multinomial model for the ESH strategy.

On the one hand the basic model considers, at each phase of testing, the list of items ordered from the most desirable to the least desirable. On the other hand, the model also considers the associated probability that describes the proportion of times each item is selected as the best item in addition to the proportion of times each item is administered, one for each item in the list. The method develops an exposure control parameter for each item using the same adjustment simulations as in the SH algorithm but rather than selecting items based on optimal item selection, this method employs a multinomial model to select the next item for administration. Multinomial probabilities are calculated to determine the probability of selection based on all previous items not being selected.

Stocking and Lewis (1995b) also develop the *Stocking and Lewis Conditional Multinomial* strategy (SLC) to directly control the item exposure to examinees of the same or similar levels of proficiency. In the unconditional method the exposure control parameters of an item is developed to reduce the item's overall appearance in reference to the examinee sample drawn from a target population. Unlike that, the conditional procedure derives for each item in the pool an exposure control parameter with respect to a particular level of examinee ability. This procedure controls against an item being administered to almost all examinees at one particular ability level, even if the item's overall exposure rate is low for examinees across the entire ability range. The advantage of conditional multinomial exposure control is that it allows direct control of item exposure for different levels of ability, while at the same time it may be possible to choose different target maximum exposure rates for different ability levels. Moreover, the exposure control parameters are not dependent on the ability distribution used in the simulation. Thomasson (1995) develops another conditional strategy, similar to the Stocking-Lewis. In this procedure, item exposure control is conditioned on examinee ability while a selection algorithm different than the multinomial method is used. However, according to our knowledge a presentation of this strategy is not currently accessible.

Restricted Maximum Information Strategy

Revuelta and Ponsoda (1998) propose the *Restricted Maximum Information* strategy as a practical alternative to the SH strategy. This strategy suggests that no item is allowed to be exposed in more than a predetermined proportion of tests. It avoids the complexities involved in the assignment of the k_i parameters where k_i is the probability that item i is administered, given that it has been selected. Items are selected by the maximum information method, but none is allowed to be exposed in more than $100k\%$ of the tests. When an item attains this limit it cannot be administered in the current test.

SH Conditional Strategy (SHC)

Chang (1998) proposes the *SH Conditional* (SHC) procedure as a competitive method to the Stocking and Lewis condition on ability procedure (SLC) in controlling the item exposure rate. Rather than deriving the exposure control parameters with respect to an entire examinee distribution representative of the real examinee population, the SHC approach derives the exposure control parameters in reference to a particular ability level.

Stocking and Lewis Conditioning on Estimated Ability

Stocking and Lewis (2000) argue that in both the conditional multinomial approach (Stocking & Lewis, 1995b) and the hybrid conditional approach (Nering, Davey & Thompson, 1998; see combined strategies) the development of the exposure control parameters are conditional on true ability but their use is based on estimated ability. They conduct simulation experiments by setting different targets for different ability levels and conclude that it is not possible to achieve these targets using the conditional multinomial approach to exposure control. They suggest that the problem arises because of the discrepancy between true ability and estimated ability early in the adaptive test, regardless of how the first item is chosen. Based on the simulation conclusions they develop a new approach in which the exposure control parameters are determined by conditioning on estimated ability rather than true ability; a method that partially solves the problem of conditioning on true ability.

Targeted Exposure Control Strategy (TEC)

Thompson (2002) develops the *Targeted Exposure Control* (TEC) strategy that attempts to increase the administration probability of unused items. In contrast to other methods that mainly focus on controlling overexposure, TEC ensures examinees are administered informative items while making good use of the item pool. To select an item with TEC, items must first meet measurement and content constraints. These items form an acceptable set of items, any one of which would be considered appropriate for administration. Once an acceptable set of items is formed, an item's probability of administration is inversely related to its administration rate. Therefore, items that were used less frequently have a higher probability of being administered. Exposure parameters are obtained through simulation as in the SH procedure.

Chen and Lei Strategy

Chen and Lei (2005) modify the SH strategy into a method that can provide item exposure control at both the item and test levels. Their strategy seeks to control the item exposure rate and test the overlap rate simultaneously. The variance and the maximum value of the item exposure rates are controlled simultaneously such that not only can most items be administered with item exposure rates less than a pre-specified value, but the test overlap rate can also be less than a pre-specified value. Thus, based on this approach, item exposure can be controlled at both the item and test levels. Their work is associated with an earlier research conducted by Chen, Ankermann and Spray (1999).

Shadow Test Approach

The *Shadow Test* approach is a general scheme for optimizing the selection of items in CAT. The basic idea behind the shadow test approach is that items are not selected directly from the pool but from a shadow test that is a full-length test assembled prior to selecting each item for administration in the adaptive test that has the following properties: (a) They contain all items already administered to the examinee; (b) they are optimal at the current θ estimate of the examinee, for example, in the sense that they maximize test information at the estimate; and (c) they meet all specifications the adaptive test has to meet. The item that is actually administered to the examinee is the one in the shadow test that has not yet been administered and is optimal at the θ estimate. After the item is administered, the unused items in the shadow test are returned to the pool, the θ estimate is updated, and the procedure is repeated (van der Linden & Chang, 2005). Each requirement that an adaptive test has to meet imposes a constraint on the selection of items from the pool. There are several constraints associated with CAT's design: test length, content, response time, item exposure rate, item overlap etc.

van der Linden and Veldkamp (2005) propose a method where item-exposure control is implemented by imposing item-ineligibility constraints on the assembly process of the shadow tests. The method resembles Sympson and Hetter's (1985) method in that the decisions to impose the constraints are probabilistic. However, it does not require time-consuming simulation studies to set values for control parameters prior to the operational use of the test. Instead, the probabilities of item ineligibility can be set on the fly using an adaptive procedure based on the actual item-exposure rates.

Stratified Strategies

The basic idea of the stratified strategies is to limit the exposure on any given item by using it at the most beneficial point in testing. Stratified strategies include: *a*-Stratified Strategy (Chang & Ying, 1999), the *a*-Stratified strategy with freezing (Parshall, Harmes & Kromrey, 2000), the *a*-Stratified strategy with *b*-blocking (Chang, Qian & Ying, 2001), the *a*-stratified CAT with unequal item exposure across strata (Deng & Chang, 2001), *a*-Stratified CAT design with content blocking (Yi & Chang, 2001), the multi-dimensional stratification method (Lee, Ip & Fuh, 2002), and the 0-1 stratification strategy (Chang & van der Linden, 2003).

a-Stratified Strategy (STR)

Chang and Ying (1999) argue that one major cause of unevenly distributed item exposure rates is that when using maximum information item selection, items with large *a* values (discrimination parameter) are more likely to be selected than those with small *a* values. By grouping items with similar a values together and selecting within a group at each stage, exposure rates would be more evenly distributed because items with all *a* values would be selected with equal frequency. Based on this argument Chang and Ying (1999) develop a multistage adaptive testing approach to control item exposure. This approach factors the discrimination parameter (*a*) into the item selection process. In this approach, the items in the item bank are stratified into a number of levels (*K* strata) based on their a values. At the early stages of a test, when little is known about the examinee's ability (θ), items with lower as are administered. As the CAT progresses and the examinee's ability estimate comes closer to approximating the examinee's known ability then items with higher as are administered. At each stage, items are selected according to an optimization criterion from the corresponding level.

a-Stratified Strategy with Freezing

Parshall, Harmes and Kromrey (2000) investigate the effectiveness of item freezing as a means of augmenting the *a-Stratified* strategy. A problem of the a-Stratified strategy is the extreme overuse of some items (Parshall, Kromrey & Hogarty, 2000). A possible solution to this problem is to temporarily render items unavailable for selection when they exceed a target administration rate by "freezing" these items in the selection algorithm until their administration rates drops below the target value. The results of their study suggest that the a-Stratified strategy with freezing perform well at constraining item administration rates to their target maximum goals, without degrading test precision unacceptably.

a-Stratified Strategy with *b*-Blocking (STR_B)

Chang, Qian, and Ying (2001) modify the *a*-Stratified strategy to have *b*-blocking. The basic idea is to force each stratum to have a balanced distribution of *b* values (difficulty level), in order to ensure a good match of θ for different examinees. This is important, because one of the major goals of CAT is to provide such matching. In this method the item bank is divided into *M* blocks in ascending order of *b*-parameter values. Then, each of the *M* blocks is stratified into *K* strata according to their *a*-parameter values. Thus, for the M^{th} block, the first stratum contains those items with the lowest *a* values within the block, and the K^{th} stratum contains items with the highest *a* values. This stratification strategy is basically the same as that of *a*-Stratified strategy (Chang & Ying, 1999), except that it is performed within a *b* block. Then, across the *M* blocks all the items in the K^{th} stratum are recombined to form a single stratum. For *K* strata the test is divided into *K* stages. In the K^{th} stage, select items from the K^{th} stratum based on the closeness of *b* values to the current estimate of θ for an examinee. This process continues for each stage. With this method the *b*-values distribute more evenly within each stratum and the average value of *a* increases across the strata.

a-Stratified CAT with Unequal Item Exposure across Strata (USTR)

Deng and Chang (2001) propose a refined stratification procedure that allows more items to be selected from the high *a* strata and fewer items from the low *a* strata. The *Unequal Item Exposure across Strata* (USTR) procedure was found in a simulation study to effectively improve test efficiency over STR, without unacceptably degrading item usage (Deng & Chang, 2001).

a-Stratified CAT Design with Content Blocking

In operational CATs content balancing is often required in the development and implementation of the tests in order to obtain relatively comparable test scores among examinees as items in one content area may tend to be more difficult than items in another content area (Stocking & Swanson, 1993). Yi and Chang (2001) propose a modified *a*-stratified method referred to as the *a-stratified method with content blocking*. As a further refinement of *a*-stratified CAT designs, this method incorporates content specifications into item pool stratification. The *a*-stratified method with content-blocking (STR_C) takes both the content specifications and the relationship between the *a*- and *b*- parameters into consideration during item pool stratification. The item pool is first stratified into groups according to the content specifications, and then the STR_B procedure is used to obtain all the strata within each content group. Finally, all items with the same stratum number are pooled across all the content groups to form the final strata. The resulting pool has three characteristics: (1) the content coverage of each stratum is similar to that of the full item pool; (2) the distribution of the *b*-parameters in each stratum is as similar as possible to that of the full item pool; and (3) the average value of the *a*-parameters increases across strata. The test is divided into several stages, one per stratum. STR_C then selects items from the corresponding strata based on the match between item difficulty and an examinee's current ability estimate. Items with low *a* values are administered in the early stages of the test and high *a* items are used during the later stages. Note that if there is only one content area, STR_C is equivalent to STR_B.

Multi-dimensional Stratification

Lee, Ip and Fuh (2002) extend Chang and Ying (1999) stratification strategy for one-dimensional tests where one skill is assessed (e.g. computational skills) to multi-dimensional tests where multiple skills are assessed simultaneously (e.g. both analytical and computational skills). They propose a stratified multi-stage strategy for controlling item exposure for d-dimensional tests, where $d > 1$. Their strategy is based on stratification in accordance with a functional of the vector of the discrimination parameter.

0-1 Stratification Strategy

Chang and van der Linden (2003) develop a method based on *0-1* linear programming (LP) to stratify an item pool optimally for use in *a*-stratified adaptive testing. They suggest that the STR strategy was proposed originally to avoid high *a* items to be overly exposed and make more even and efficient use of all items in an item pool. STR performs well for ideal item pools when the *a* and *b* parameters are not correlated, but it could lead to problems when *a* and *b* are correlated. This item pool stratification method (0-1STR) provides a solution for this case with respect to an ideal classification of the pool. It can be thought of as a pre-emptive measure to force balanced distributions of *b* values across strata. As a result, some of the strata formed by the method cover a wider range of *b* values than that for the original STR.

Combined Strategies

In an attempt to control item exposure, several researchers combined different methods to develop more robust strategies that seem to perform better in certain situations than each strategy alone. Combined strategies include: Progressive Restricted strategy (Revuelta & Ponsoda, 1998), Nering, Davey and Thompson's Hybrid strategy (Nering, Davey & Thompson, 1998), Eggen's strategy (Eggen, 2001), incorporation of the Sympson-Hetter Exposure Control Method into the *a*-Stratified method with Content Blocking (Yi, 2002), Combination of the *a*-Stratified strategy and the Sympson-Hetter strategy (Leung, Chang, & Hau, 2002) and Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test approach (van der Linden & Chang, 2005).

Progressive Restricted Strategy

Comparing item exposure control strategies Revuelta and Ponsoda (1998) conclude that the *Restricted Maximum Information* strategy is useful to reduce maximum exposure rates and that the *Progressive* Strategy reduces the number of unused items, while both perform well with regard to precision. Thus, they combine the restricted maximum information strategy and the progressive strategy to create the *progressive restricted* strategy to control item exposure without a serious decrease in test precision. In this procedure, before the administration of a CAT, the available items are determined by the restricted strategy, so that no item will exceed the maximum exposure rate. Once the item pool is determined for a CAT, the progressive strategy is used to select an item for administration.

Nering, Davey and Thompson's Hybrid Strategy

Nering, Davey and Thompson (1998) in comparing a number of different exposure control strategies argue that the control of item security improves with increased conditionality. Based on this conclusion they propose an exposure control strategy that combines elements of the Stocking and Lewis (1995b) conditional on ability level and the Davey and Parshall (1995) strategy conditional on all other items that have already appeared in the test. In this strategy the diagonal elements of the Davey and Parshall table are replaced by separate vectors for each of the ability levels in the stratification on true ability in the Stocking and Lewis conditional on ability strategy. These values limit the frequency with which items can be administered to examinees at each ability level.

Eggen's Strategy

Eggen's research (2001) suggests that the SH strategy presents an effective solution to the overexposure problem while the progressive strategy is effective against underexposure. He proposes a combined application of both strategies to address overexposure and underexposure of items, where item selection is based on a mixture of two criteria: chance and maximum information at the current ability estimate. At the start of a test administration, the weight of the chance criterion is large and that of the maximum information criterion is small, but, as the test administration progresses, the weight of chance decreases and that of maximum information increases.

Incorporation of the Sympson-Hetter Exposure Control Method into the *a*-Stratified Method with Content Blocking (STR_C-SH)

Yi (2002) argues that STR_C can effectively reduce the exposure rates of highly discriminating items; however, it does not have a mechanism to control the maximum observed item exposure rate at a certain level. To overcome this disadvantage Yi incorporates the SH exposure control procedure into STR_C (STR_C-SH) to achieve the goal of limiting the maximum observed item exposure rate at a pre-specified level. STR_C-SH can be implemented in a similar way as STR_C, except that the SH procedure is used within each stratum to control the maximum observed item exposure rates. Similar to MI-SH, exposure control parameters are obtained through a series of simulated CATs and then they are used in the CAT to control the maximum observed item exposure rates. However, STR_C-SH selects items differently from SH. Specifically; items are selected across strata based on the closeness between item difficulty and a current ability estimate. The exposure control parameter of the selected item is compared with a uniform random number to decide if the chosen item should be administered. After obtaining the exposure control parameters, they are used in the CAT to control the frequency with which items are administered.

Combination of the *a*-Stratified Strategy (STR) and the Sympson-Hetter Strategy

Leung, Chang, and Hau (2002) argue that the *a*-Stratified strategy does not automatically guarantee that the exposure for individual items can be kept below a specified rate. This strategy has been demonstrated to be effective in improving the utilization of the entire item pool without sacrificing the efficiency in ability estimation; however, the problem of item overexposure still persists when the ratio of pool size to test length is small. To deal with this problem they combine the *a*-Stratified strategy (STR) and the Sympson-Hetter (SH) strategy to create the enhanced stratified exposure control procedure.

Content Constraints in *a*-Stratified Adaptive Testing Using a Shadow Test Approach

van der Linden and Chang (2005) combine the methods of alpha-stratified adaptive testing and constrained adaptive testing with shadow tests. Their study reveals two main advantages: First, application of the shadow test approach allows the implementation of any type of constraint on item selection in alpha-stratified adaptive testing. Second, the result yields a simple set of constraints that can be used in any application of the shadow test approach to reduce overexposure and underexposure of the items in the pool.

Multiple Stage Adaptive Test Designs

An alternative to implementing an exposure control procedure to modify optimal item selection is to control exposure a priori by pre-constructing adaptive test forms. For this reason multi-stage computer adaptive tests (MST) have been developed from several researchers. A multi-stage computer adaptive test (MST) combines characteristics of both a standard CAT and P&P test because it adapts to the ability of the examinee like CAT and provides P&P benefits such as test specialist review, exposure of pre-selected items and parallel test forms (Armstrong & Edmonds, 2004).

Computerized Adaptive Sequential Testing (CAST) is a case of such test design (Luecht, Nungester & Hadadi, 1996; Luecht & Nungester, 1998). CAST integrates test design, test assembly, test administration, and data management components in a comprehensive manner intended to support the mass production of secure, high quality parallel test forms over time. The basic test design model for CAST is a multistage test, with adaptive testing capabilities at the level of subtests or testlets (a set of items centered on a single stimulus). CAST, involves the pre-construction of modules that contain groups of items and the arrangement of these modules into multi-staged panels. CAST allows security measures to be implemented in several ways: items can be randomly presented within modules; item and module level exposure controls can be implemented as part of the automatic test assembly (ATA) item selection process to reduce item overlap across panels; empirical item overlap can be explicitly constrained as part of the ATA process, when building multiple instances of the panels; panels can be activated on the basis of having minimal item overlap for a particular period of time; and explicit pathways can be periodically evaluated for potential overexposure and the entire panel de-activated when a certain threshold is reached (Luecht & Nungester, 2000).

Luecht (2003) modifies CAST and develops the *Adaptive Multi-Stage Item Bundles* (BMAT). This multistage adaptive testing test development paradigm aims to effectively handle content balancing and other test development needs, psychometric reliability concerns, and item exposure. BMAT involves the construction of banks of parallel testlets to meet various statistical targets and categorical constraints. It also requires automated test assembly (ATA) technology capable of handling multiple, simultaneous objective functions and constraint systems. In addition, BMAT incorporates random selection of the testlets and can allow randomization of the item presentation sequence within modules to prevent attempts at memorization and other forms of collaborative cheating. The net result is a secure method of building high-quality adaptive and mastery tests that have severe constraints on test content.

Similar to BMAT is the *Multiple Forms Structures* (MFS) approach, which was independently developed by Armstrong and Little (2003). An MFS is an ordered collection of testlets. Every test-taker is administered the same testlet(s) early in the test, however, an examinee's progression through the network of testlets is dictated by the correctness of an examinee's answers, thereby adapting the test to his or her trait level. The MFS format is a hybrid between the conventional P&P and CAT formats. The collection of paths through the network yields the set of all possible test forms, allowing test specialists the opportunity to review them before they are administered. The possible paths through the MFS give the possible test forms. Every form must satisfy its own test specifications. Limiting the exposure of an individual MFS to a specific period of time can enhance test security.

Evaluation Studies

Several studies have been conducted in order to evaluate the effectiveness of item exposure control strategies with regard to associated variables of measurement precision, exposure rate, pool utilization, and test overlap. Moreover, ease of implementation is also an important issue and it is usually evaluated with respect to the gains made in each of the previous mentioned variables.

Most of the studies focus on dichotomous scoring (right/wrong). Recently, however, researchers have begun to examine the effects of exposure control when using polytomous scoring (modeling all of the individual possible responses to an item, rather than just whether the response is correct or not) and testlet-based CAT systems (testlets are defined as a group of items that relate to a single stimulus). A number of such evaluation studies are presented below.

Dichotomous Scoring

Chang and Twu (1998) perform a comparative study of item exposure control strategies for CAT. In their study they investigate and compare the properties of five strategies: 5-4-3-2-1 randomization, Sympson and Hetter, Davey and Parshall, Stocking and Lewis unconditional strategy and Stocking and Lewis conditional on ability strategy. They conclude that among the five control algorithms, the Stocking and Lewis conditional on ability strategy best serves the purposes of controlling the observed exposure rates to the desired values, as well as producing the lowest test overlap rates. However, they report that the measurement precision is sacrificed to some extent, particularly at both extreme ability levels.

Revuelta and Ponsoda (1998) compare the Progressive method and the Restricted Maximum Information method with six other item-selection methods (Maximum Information, One Parameter, McBride and Martin, Randomesque, Sympson and Hetter, and Random Item Selection) with regard to test precision and item exposure variables. Results from their study show that the Restricted method is useful to reduce maximum exposure rates and that the Progressive method reduces the number of unused items. Both perform well regarding precision. They conclude that a combined Progressive-Restricted method may be useful to control item exposure without a serious decrease in test precision.

Parshall, Davey and Nering (1998) compare empirically the relative effectiveness of exposure control strategies based on simulated CATs. They conclude that the Sympson-Hetter, Stocking and Lewis conditional on ability, and Davey and Parshall strategy outperform a no exposure control method and that the Stocking and Lewis conditional on ability as well as the Davey and Parshall strategy generally outperform the Sympson-Hetter strategy.

French and Thompson (2003) evaluate the following exposure control procedures: the Targeted Exposure Control (TEC), the Sympson and Hetter, and the *a*-stratified with *b*-blocking. The three procedures are applied to a variable length CAT to evaluate their effect on item pool use, test length, and test reliability. Performance is examined, conditionally and unconditionally, on several criteria such as pool utilization, measurement precision, test overlap, and ease of implementation. All procedures perform similarly in terms of reliability, bias, and root mean square error. The Targeted Exposure Control procedure makes better use of the item pool as judged by the percent of zeros, the test overlap, the chi-square statistic, and the maximum exposure rate. TEC is also able to use every item, unlike the other procedures. However, conditional results suggest that none of the procedures performs adequately in the tails of the ability distribution.

Polytomous Scoring

Pastor, Dodd and Chang (2002) investigate the impact of using five different exposure control algorithms in two sized item pools calibrated using the generalized partial credit model. The results of the simulation study indicate that the *a*-Stratified design, in comparison to a no-exposure control condition, could be used to reduce item exposure and overlap, to increase pool utilization, and only minimally degrade measurement precision. Use of the more restrictive exposure control algorithms, such as the Sympson-Hetter and conditional Sympson-Hetter, control exposure to a greater extent but at the cost of measurement precision. Their study recommends use of the more simplistic exposure control procedures, particularly when the test length to item pool size ratio is large, because convergence of the exposure control parameters is problematic for some of the more restrictive exposure control algorithms.

Davis (2002; 2004) investigates the performance of different exposure control mechanisms under three polytomous IRT models in terms of measurement precision, test security, and ease of implementation. Davis's study examines the partial credit, generalized partial credit, and graded response models. In addition to a no exposure control baseline condition, the randomesque, within .10 logits, Sympson-Hetter, conditional Sympson-Hetter, *a*-Stratified, and enhanced *a*-Stratified procedures are implemented to control item exposure rates. The *a*-Stratified and enhanced *a*-Stratified procedures are not evaluated with the partial credit model. Two variations of the randomesque and within .10 logits procedures are also examined that varies the size of the item group from which the next item to be administered is randomly selected. The study concludes with similar results for all three models and indicates that the randomesque and within .10 logits procedures, once implemented with the six item group variation, provide the best option for controlling exposure rates, especially when impact to measurement precision and ease of implementation are considered. The three item group variations of the procedures are, however, ineffective in controlling exposure, overlap, and pool utilization rates to desired levels. The Sympson-Hetter and conditional Sympson-Hetter procedures are difficult and time consuming to implement, and while they achieve control exposure rates to the target level, their performance in terms of item overlap (for the Sympson-Hetter) and pool utilization is disappointing. The *a*-Stratified and enhanced *a*-Stratified procedures both perform poorly across all variables.

Continuing their research, Davis and Dodd (2005) investigate the performance of four procedures for controlling item exposure in a CAT under the partial credit model. In addition to a no exposure control baseline condition, the Kingsbury-Zara, modified-within-.10-logits, Sympson-Hetter, and conditional Sympson-Hetter procedures are implemented to control exposure rates. The Kingsbury-Zara and the modified-within-.10-logits procedures are implemented with 3 and 6 item candidate conditions. The results show that the Kingsbury-Zara and modified-within-.10-logits procedures with 6 item candidates perform as well as the conditional Sympson-Hetter in terms of exposure rates, of overlap rates, and of pool utilization.

Testlet-based CAT Systems

Boyd (2003) examines CAT systems modeled with testlet response theory in order to determine optimal exposure control procedures. Her research examines various exposure control procedures in testlet-based CAT systems using the three-parameter logistic testlet response theory model and the partial credit model. The exposure control procedures are the randomesque procedure, the modified within .10 logits procedure, two levels of the progressive restricted procedure, and two levels of the Sympson-Hetter procedure. Each of these is compared to a baseline no exposure control procedure, maximum information. The testlets are reading passages with six to ten multiple-choice items. The CAT systems consists of maximum information testlet selection contingent on an exposure control procedure and content balancing for passage type and the number of items per passage; expected a posteriori ability estimation; and a fixed length stopping rule of seven testlets totaling fifty multiple-choice items. Measurement precision and exposure rates are examined to evaluate the effectiveness of the exposure control procedures for each measurement model. The exposure control procedures yield similar results for measurement precision within the models. The exposure rates distinguish which exposure control procedures are most effective. The Sympson-Hetter conditions, which are conditional procedures, maintain the pre-specified maximum exposure rate, but perform very poorly in terms of pool utilization. The randomization procedures, randomesque and modified within .10 logits, yield low maximum exposure rates, but use only about 70% of the testlet pool. The progressive restricted procedure, which is a combination of both a conditional and randomization procedure, yield the best results in its ability to maintain and control the maximum exposure rate and it uses the entire testlet pool. The progressive restricted conditions are the optimal procedures for both the partial credit CAT systems and the three parameter logistic testlet response theory CAT systems.

Davis and Dodd (2003) examine item exposure control procedures for testlet scoring of reading passages in the Verbal Reasoning section of the Medical College Admission Test (MCAT) with four computerized adaptive testing (CAT) systems using the partial credit model. The first system uses a traditional CAT using maximum information item selection. The second system uses random item selection to provide a baseline for optimal exposure rates. The third system uses a variation of Lunz and Stahl's randomization procedure. The fourth system uses Luecht and Nungester's computerized adaptive sequential testing (CAST) system. A series of simulated fixed-length CATs are run to determine the optimal item selection procedure. Results indicate that both the randomization procedure and CAST perform well in terms of exposure control and measurement precision, while the CAST system provides the best overall solution when all variables are taken into consideration.

Summary and Discussion

The paper examines five groups of item exposure control strategies (randomization strategies, conditional selection, stratified strategies, combined strategies, and multiple stage adaptive test designs). Each strategy group has its strengths and its weaknesses and the selection always depends on the overall design and use of a single CAT that each strategy employs to control the item exposure. For example, for a small scale school test, randomization strategies appeared to be the best option as they are basically considered to be easy to implement and easily understood. However, Stocking and Lewis (1995a) argue that the success of randomization strategies is difficult to predict with complex but realistic item pool structures and test specifications and may not prevent overuse of some items. Moreover, the best choice of group sizes can only be determined by tedious trial and error approaches with no certainty of success and no easy generalization of different item pools and test structures.

Most strategies under the conditional selection strategies group control the exposure rate of an item through the exposure control parameter, which dictates the probability of administering the item given its selection. The values of these parameters are determined from a series of iterative multifaceted simulations prior to operational use using all items in the pool that may have to be repeated as test conditions change.

The Sympson-Hetter (SH) and the Extended Sympson-Hetter (ESH) strategy (Stocking, 1993) has two main disadvantages: 1) the simulations to obtain the exposure control parameters are time consuming and if the item pool is changed by adding or deleting items—or if the target population changes significantly—the simulations must be rerun; 2) if the structure of the item pool is not a good match with the structure of the specifications it is possible for the ESH strategy to be unable to obtain stable values of the exposure control parameters because they are dependent on the distribution of examinee ability level used in the simulation (Stocking & Lewis,1995a).

To overcome these practical disadvantages Stocking and Lewis (1995a) propose the unconditional and conditional multinomial exposure control strategies. However, they argue that both the unconditional and conditional multinomial exposure control strategies retain some of the disadvantages of SH and ESH strategies such as the dependence of the conditional exposure control parameters upon the specific item pool and test structure use in the iterative adjustment simulations. Moreover, the conditional control of exposure makes the adjustment process of developing the exposure control parameters even more time-consuming and tedious than when exposure control is unconditional (Stocking & Lewis 1995b). Chang and Harris (2002) explore how the deletion of a single item and the unused items might alter the exposure control parameters of the remaining items derived by the Stocking and Lewis conditional algorithm. Their findings indicate that the original exposure control parameters are no longer appropriate for a modified item pool and the derivation for the exposure control parameters ought to be repeated.

Chang, Ansley and Lin (2000) in comparing the *SH Conditional Strategy* (SHC) (Chang, 1998) with the *Davey and Parshall* and the *Stocking and Lewis Conditional* (SLC) procedures conclude the SHC best serves the purposes of controlling exposure rates to the desired values as well as producing the lowest test overlap rates. Moreover, the SLC procedure is more efficient in preparing the exposure control parameters. Nevertheless, developing these parameters with respect to each ability level is very tedious compared to the SLC method.

In general, the conditioned item selection performs better than the randomized item selection strategies on controlling item exposure. Finding stabilized item exposure parameters through iterative simulation for the conditioned item selection, however, is a very time consuming effort. In addition, the tedious iterative simulations need to be re-conducted whenever there are changes in CAT settings or in the examinee population of interest. In order to ease this problem Chen and Doong (2003) develop a relationship formula between item exposure parameters and item parameters in CATs by using genetic programming (GP), which is a biologically inspired artificial intelligence technique. Based on the relationship formula, item exposure parameters for new parallel item pools can be predicted with moderate errors by using the GP techniques, without conducting any tedious iterative simulations.

With the *Shadow Test* approach van der Linden and Veldkamp (2005) propose an alternative to the Sympson-Hetter item exposure control method, which is based on decisions about the eligibility of the items in the pool before the test taker takes the test. If an item is eligible it remains in the pool; if it is ineligible it is removed from the pool for the test taker. These decisions are based on the outcomes of a probabilistic experiment with probabilities of eligibility that constrain the item-exposure rates to be below the target value. The method resembles Sympson and Hetter's (1985) method of item-exposure control in that the decisions to impose the constraints are probabilistic. However, it does not require time-consuming simulation studies to set values for control parameters prior to the operational use of the test. Instead, it is self-adaptive and can be implemented "on the fly" during operational testing. The method counts certain events during the testing and uses these counts to automatically adapt the probabilities of item eligibility to their optimal level, which is then maintained during the rest of the testing process. An extensive simulation study showed that the probabilities of item eligibility were already stable after 1,000 test takers were tested, and the method produced exposure rates that were below the target for all items.

Stratified strategies aim to limit the exposure on any given item by using that item at the most beneficial point in testing. The results of analytic and simulation studies with regard to the *a*-Stratified strategy (STR), which is the basis of all stratified strategies, indicate that this method has two advantages. First, it increases the utilization of items with low discriminations when they can be most efficiently used. Second, it equalizes item exposure rates by reducing rates for items that would otherwise be overexposed and by increasing rates for those that would otherwise be underexposed (Chang & Ying, 1999). However, Chang and Ying (1999) argue that when compared to the maximum information selection method through CATs simulated using both ideal and operational item pools, STR results in more evenly distributed exposure rates and reduces overlap rates, while achieving somewhat lower test efficiency. The *a-Stratified CAT with Unequal Item Exposure across Strata* (USTR) strategy developed from Deng and Chang (2001) address the issue of efficiency loss.

Chang, Qian, and Ying (2001) report that the *a-Stratified strategy with b-Blocking* improves the control of item exposure rates and yield lower mean squared errors in comparison to the *a-Stratified* strategy. Research results from simulation studies show that *a-Stratified CAT Design with Content Blocking* (STR_C) outperforms the *a-Stratified* strategy, and the *a-Stratified strategy with b-Blocking* and the maximum information selection method with Sympson-Hetter exposure control in a situation where all four procedures are forced to balance content. STR_C lowered exposure rates for highly discriminating items, and increased the usage of less discriminating items without a loss in measurement precision (Yi & Chang, 2001).

Regarding the *Multi-dimensional Stratification* strategy (Lee, Ip & Fuh, 2002) the empirical results from simulation studies indicate that the strategy is conceptually appealing and can be implemented with minimal computational overhead. The exposure rate can substantially improve when stratification is based upon a judiciously chosen function of the discrimination parameter, while the loss of efficiency is relatively mild when the number of test items administered reaches 30.

Simulation results show that the *0-1 Stratification* strategy produces comparable or slightly better statistical features as the *a*-STR strategy but clearly improves item exposure control (Chang & van der Linden, 2003).

The *Combined strategies*, in general, aim to make use of the strengths that the different exposure control strategies. Revuelta and Ponsoda (1998) suggest that the *Progressive Restricted* strategy seems to perform well on precision and exposure control and no parameters have to be determined by previous simulations. Nering, Davey and Thompson (1998) suggest that their hybrid strategy compared to Davey and Parshall (1995) and Stocking and Lewis conditional on ability strategies give better results in terms of minimizing test overlap, controlling exposure rates and force a more balanced use of the item pool. Eggen (2001) combines the SH strategy that presents an effective solution to the overexposure problem with the progressive strategy, which is effective against underexposure in order to address overexposure and underexposure of items.

Simulation studies show that the *Incorporation of the Sympson-Hetter Exposure Control* method into the *a-Stratified Method with Content Blocking* (STR_C-SH) maintains the effectiveness of the *a*-Stratified method with content blocking to produce balanced item usage within a pool, while closely controlling the maximum observed item exposure rate at a pre-specified level. It also results in measurement precision that is comparable to that of the maximum information selection method with Sympson-Hetter exposure control.

Leung, Chang, and Hau (2002) combine the *a-Stratified* strategy (STR) and the *Sympson-Hetter* strategy. The performance of such an enhanced stratified method (STR-SH) is compared with that of STR as well as maximum information-SH. The results indicate the potential advantages of the STR-SH design over the original STR in yielding a more balanced item exposure distribution, further reducing the test-overlap rate (to near the lower bound) and effectively controlling item exposures below target maximum rate.

With regard to the combination of *a*-Stratified adaptive testing and constrained adaptive testing with shadow tests, van der Linden and Chang (2005) argue that a large number of content constraints can easily be implemented in alpha-stratified CAT through a shadow-test approach. For a well-designed item pool (they use LSAT in their empirical study), imposing content constraints on the item selection does not need to have any disadvantageous impact on the statistical properties of the ability estimator. Relative to maximum-information CAT, alpha-stratification tends to result in much more favorable exposure rates for the items. The rates for the popular items are likely to be reduced considerably and, equally important, those for the unpopular items are likely to go up to much more acceptable levels. However, a slight loss in the accuracy of the estimator that is observed can be compensated for by adding a few items to the test, whereas loss due to item compromise or inefficient item use is more difficult to compensate.

With regard to *Multiple Stages Adaptive Test Designs,* CAST, BMAT, and MFS, they are cases of test design that control exposure a priori by pre-constructing adaptive test forms as an alternative of implementing an exposure control procedure to modify optimal item selection. Such designs integrate several test components (i.e. test design, test assembly, test administration, and data management) to support the mass production of secure, high quality parallel test forms over time. However, they are complicated to implement for small scale adaptive tests.

Lastly, it is worth referring to a very recent effort by Yi, Zhang and Chang (2005; 2006) on assessing CAT test security severity. They develop the *AddChart Application* software in order to examine the relationship among item pool size, the number of items each professional test taker (examinees who either are employed by test preparation organizations or have taken the same test several times to boost their test scores) can memorize, and the percentage of the item pool that can be compromised. For example, AddChart Application can demonstrate, for a given item pool, the number of professional test takers that are needed to compromise various percentages of the item pool given that each person can memorize b items (where b is any fixed number less than or equal to the test length). The software can be used to assist practitioners and researchers in designing a more secure CAT based on the information from examining the relationship between the number of professional test takers needed and the percentage of the compromised item pool.

Summarizing, since researchers acknowledged the several advantages of computerized adaptive testing over traditional linear test administration the issue of item exposure control has received increased attention. Several item exposure control strategies have been presented in the literature aiming to prevent overexposure of some items and to increase the use rate of rarely or never selected items. The present paper attempted to review such strategies that appeared in the relevant literature from 1983 to 2005 and classified them into five main categories: a) randomization strategies; b) conditional selection; c) stratified strategies; d) combined strategies; and e) multiple stage adaptive test designs. The paper focused on studies that have been conducted in order to evaluate the effectiveness of item exposure control strategies for dichotomous scoring, polytomous scoring and testlet-based CAT systems. The paper also discussed the strengths and the weaknesses of each strategy group using examples from simulation studies.

References

Armstrong, R. & Edmonds, J. (2004, April). *A study of multiple stage adaptive test designs.* Paper presented at the annual meeting of National Council of Measurement in Education, (NCME), San Diego, CA.

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.

Boyd, A.M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems.* Unpublished PhD Thesis, The University of Texas at Austin.

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

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

Chang, H.H. & 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, S.W., Ansley, T.N. & Lin, S.H. (2000, April). *Performance of item exposure control methods in computerised adaptive testing: Further explorations.* Paper presented at the annual meeting of American Educational Research Association (AERA), New Orleans, LA

Chang, S.W. & Twu, B.Y. (September 1998). *A comparative study of item exposure control methods in computerised adaptive testing.* ACT Research Report Series, ACT-RR-98-3.

Chang, S.W. (1998). *A comparative study of item exposure control methods in a computerized setting.* Unpublished PhD Thesis, The University of Iowa, Iowa City.

Chang, S.W. & Harris, D. (2002, April). *Redeveloping the exposure control parameters of CAT items when a pool is modified.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans.

Chen S.Y., Ankermann, R.D. & Spay, J.A. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerised adaptive testing.* ACT Research Report Series, ACT-RR-99-5.

Chen S.Y. & Doong S. H. (2003). *Predicting item exposure parameters in computerized adaptive testing.* Paper presented at the 2003 Annual Meeting of the American Educational Research Association (AERA), Chicago, IL.

Chen, S.Y. & Lei, P.W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*(2), 204–217.

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

Davey, T. & Nering, M. (2002). Controlling item exposure and maintaining item security. In C.N. Mills, M.T. Potenza, & J.J. Fremer (Eds.), *Computer-Based Testing: Building the Foundation for Future Assessments.* (pp. 165–191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Davis, L. & Dodd, B. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*(3), 335–356.

Davis, L. & Dodd, B. (March 2005). *Strategies for controlling item exposure in computerized adaptive testing with the partial credit model.* Pearson Educational Measurement Research Report 05-01.

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

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

Deng, H. & Chang, H.H. (2001). *a-Stratified computerized adaptive testing with unequal item exposure across strata.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle WA.

Eggen, T.J.H.M. (2001). *Overexposure and underexposure of items in computerized adaptive testing.* Measurement and Research Department Reports 2001-1, Arnhem, The Netherlands: CITO Groep.

Flaugher, R. (2000). Item pools. In H. Wainer (Ed), *Computerized adaptive testing: A primer* (2nd ed.) (pp. 37–59). Mahwah, NH: Lawrence Erlbaum Associates.

French, B. & Thompson T. (April, 2003). *The evaluation of exposure control procedures for an operational CAT.* Poster presented at the annual meeting of the American Educational Research Association (AERA), Chicago, IL.

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

Grolund, N.E. (1998). *Assessment of student achievement,* Sixth Edition. Needham Heights, MA: Allyn & Bacon.

Han, N. & Hambleton, R. (2004, April). *Detecting exposed test items in computer-based testing.* Paper presented at the meeting of the National Council of Measurement in Education (NCME), San Diego.

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

Lee, Y.H., Ip, E.H. & Fuh, C.D. (2002). *A strategy for controlling item exposure in multidimensional computerized adaptive testing.* Retrieved May, 05, 2005, from http://www3.stat.sinica.edu.tw/library/c_tec_rep/c-2002-11.pdf

Leung, C.K., Chang, H.H. & Hau, K.T. (2002, December). Item selection in computerized adaptive testing: Improving the *a*-Stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26*(4), 376–392.

Linacre, J.M. (2000). Computer-adaptive testing: a methodology whose time has come. MESA Memorandum No. 69. In S. Chae, U. Kang, E. Jeon & J. M. Linacre (Eds.), *Development of computerised middle school achievement test* (translation from Korean). Seoul, South Korea: Komesa Press.

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.

Luecht, R.M. & Nungester, R.J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden (Ed.), *Computerized Adaptive Testing: Theory and Practice,* (pp. 289–309). Dordrecht, The Netherlands: Kluwer.,

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.

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.

Nering, M.L., Davey, T. & Thompson, T. (1998, June). *A hybrid method for controlling item exposure in computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society. Urbana, IL.

Parshall, C., Davey, T. & Nering, M.L. (1998, April). *Test development exposure control for adaptive testing.* Paper presented at the annual meeting of the National Council of Measurement in Education (NCME), San-Diego, CA.

Parshall, C., Harmes, J.C., & Kromrey, J.D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research, 40*(1), 28–52.

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

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4) 311–327.

Segall, D.O. (1994). *CAT-GATB simulation studies.* San Diego, CA: Navy Personnel Research and Development Centre.

Stocking, M.L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm.* Technical Report RR 3-2. Princeton, NJ: Educational Testing Service.

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

Stocking, M.L. & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing.* Educational Testing Service Research Report 95-24, Princeton, NJ.

Stocking, M.L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice,* (pp. 163–182). Norwell MA: Kluwer.

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.

Thomasson, G.L. (1995). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

Thissen, D. & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer,* (pp. 101–133). Hillsdale NJ: Erlbaum.

Thompson, T. (2002, April). *Employing new ideas in CAT to a simulated reading test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

van der Linden, W.J. & Veldkamp, B.P. (2005, December). *Constraining item exposure in computerized adaptive testing with shadow tests.* Law School Admission Council Computerized Testing Report 02-03.

van der Linden, W.J. & Chang, H.H. (2005, August). *Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach.* Law School Admission Council, Computerized Testing Report 01-09.

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.). Mahwah, NJ: Lawrence Erlbaum Associates.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*(4), 17–27.

Wise, S.L. (1997, March). *Overview of practical issues in a CAT program.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Yi, Q. (2002, April). *Incorporating the Sympson-Hetter exposure control method into the a-stratified method with content blocking.* Paper Presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.

Yi, Q. & Chang, H. (2001, June). *a-Stratified computerized adaptive testing with content blocking.* Paper presented at the Annual Meeting of the Psychometric Society, King of Prussia, PA.

Yi, Q., Zhang, J. & Chang, H. (2005, April). *Identifying practical indices for enhancing item pool security.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada.

Yi, Q., Zhang, J. & Chang, H. (2006). Assessing CAT Test Security Severity. *Applied Psychological Measurement, 30*(1), 62–63.

Acknowledgment

The work presented in this article is partially funded by the General Secretariat for Research and Technology, Hellenic Republic, through the E-Learning, EL-51, FlexLearn project.

Author Biographies

Dr. Elissavet Georgiadou is a lecturer/advisor at the School of Applied Arts (Postgraduate course, Graphic Arts-Multimedia), Hellenic Open University, Greece. Her current research interests include adaptive learning, adaptive assessment, design and evaluation of educational hypermedia systems and typography. Her research was presented at a number of international conferences and scientific journals. She can be contacted at elisag@otenet.gr .

Dr. Evangelos Triantafillou is a researcher at the Computer Science Department, Aristotle University of Thessaloniki, Greece. His research interests include Educational Technology, Multimedia Educational Technology, and Adaptive Hypermedia Systems. He has published several papers in international scientific journals and presented his work at a number of international conferences. He can be contacted at vtrianta@csd.auth.gr .

Dr. Anastasios A. Economides is an Associate Professor and Vice-Chairman in the Information Systems Postgraduate Program at the University of Macedonia, Thessaloniki, Greece. His current research interests include mobile, collaborative and adaptive learning and networks. He has published over one hundred peer-review papers. He can be contacted at economid@uom.gr .

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College