A Strategy for Controlling Item Exposure in Multidimensional Computerized Adaptive Testing

Yi-Hsuan Lee

Academia Sinica, Taipei, Taiwan

Edward H. Ip

University of Southern California

Cheng-Der Fuh†

Academia Sinica, Taipei, Taiwan

†Research partially supported by a grant NSC 91-2118-M-001-006 from the National Science Council, Taiwan. The authors thank Dr. Yeong-Nan Yeh for a stimulating discussion.

A Strategy for Controlling Item Exposure in Multidimensional Computerized Adaptive Testing

Abstract

While computerized adaptive tests have enjoyed tremendous growth in the past decade, satisfactory solutions for many important problems are still unavailable. Among these problems, a critical one is the control of item exposure rate. Because adaptive algorithms are designed to select optimal items, they tend to choose items with high discriminating power. Thus, these items are selected far more often than other items, which leads to both over-exposure of some parts of the item pool and under-utilization of others. The risk is that over-used items are often compromised. As a result, they create a security problem that could threaten the validity of a test. Chang and Ying (1999) proposed a stratification scheme to control the exposure rate for one-dimensional tests. In this paper we extend their method to multi-dimensional tests. Specifically, we propose a strategy based on stratification in accordance with a functional of the vector of the discrimination parameter. The proposed strategy is conceptually appealing and can be implemented with minimal computational overhead. In the paper, we provide both theoretical and empirical studies to validate the multi-dimensional stratification method. Our empirical results indicate that the proposed method achieves significant improvement over the optimal method of controlling exposure rate and that it requires only a small sacrifice in efficiency.

keywords: item response theory, computerized adaptive testing, discrimination parameter, recursive maximum likelihood estimate, test security.

Introduction

The advent of modern computer technology has created a vast array of both opportunities and challenges for measurement specialists. One example is computerized adaptive testing (CAT). The crucial idea underlying CAT is its customizing of test items to each individual subjects. Much like mass customization for consumer products, in which products are tailored to the preferences of individual consumers, CAT allows test designers to tailor the set of items presented to individual subjects according to their levels of sophistication as measured along a continuum of cognitive or psychological traits. In the context of educational testing, Lord (1970, 1971) was amongst the first to argue that the set of items should neither be too easy nor too difficult, so that the proficiency of the test taker could be accurately and efficiently determined. Some recent discussions of general issues of CAT have included Sands, Waters, and McBride (1997), Wainer et al. (2000), and a special issue of this journal in September 1999.

While the technology of CAT has made tremendous progress over the past decade, many important challenges remain, and unsuspected problems have emerged as well. One of the primary concerns that is still outstanding in the development and implementation of CAT is item exposure (Wainer 2000a). Without appropriate control, "good" items are often automatically selected by a CAT program far too often than "poor" items. It has been estimated that 15-20 percent of an item pool may account for more than 50 percent of the test items in a typical CAT exam (Wainer, 2000b). Because CAT item pools are almost invariably reused (mainly for economic reasons), the overexposure of certain items could easily lead to information sharing among test takers. As a result, overused items are likely to become compromised and thus adversely affect the validity of the measuring instrument. Therefore, effectively controlling item exposure is a critical issue for reducing security risk in the practical implementation of CAT.

The research on controlling item exposure rate in CAT has so far been concentrated on one-dimensional item response theory (IRT) models. Some examples include Stocking and Lewis (1995), Davey and Parshall (1995), van der Linden and Reese (1998), Chang and Ying (1999), and Chang and Zhang (2002). In this paper, we propose a stratified multi-stage strategy for controlling item exposure for d-dimensional tests, where $d > 1$. Specifically, we developed

related theory for the proposed strategy, and we investigated its practical utility. Although the strategy can be generalized to any dimension, in this paper we only considered $d = 2$.

Our proposed method extends the work of Chang and Ying (1999), in which a stratifying scheme was proposed for unidimensional ($d = 1$) tests. The authors' method for controlling item overexposure was based upon stratifying a unidimensional item pool in accordance with the discrimination parameter, a. In the initial stage of CAT, when little information is available about an examinee's ability, items are administered from a stratum with the lowest a-parameters. As the number of responses increases, more information about the examinee's ability is available. Then the test uses the strata with higher values of a to pinpoint the examinee's ability. For unidimensional tests, the scheme is conceptually appealing (it mimics a human examiner's strategy of asking some general questions first before proceeding to more specific questions) and practically straightforward in terms of implementation – it orders items and stratifies them in advance, so the operational overhead is minimal. However, for multidimensional tests, determining how the item pool can be stratified is not trivial, because a is now a vector of length $d > 1$. One logical solution is to use some functional of a for the stratification. We propose a specific functional form for a and present theoretical and empirical studies to validate the choice of this functional.

There is a pressing need for developing practical procedures for controlling item exposure for multidimensional CAT. The past decade has witnessed exponential growth in the use of CAT, which is partly sustained by the adoption of CAT in highly visible, large-scale assessments such as the Graduate Record Exam (GRE) and the Armed Services Vocational Aptitude Battery (ASVAB). The number of CAT measuring instruments has increased from a few hundred in the early 1990s to more than a million now (Wainer 2000b). Along with the phenomenal growth in the number of CAT instruments, the complexity of and scope of CAT applications have also greatly expanded. As a result, methods that are only designed for unidimensional tests are not sufficient to cover the broad array of CAT exams that have been developed. For example, Bock, Thissen, and Zimowski (1997) provided empirical evidence that revealed the multidimensional nature of ASVAB items. In the Graduate Management Admission Test (GMAT), an item that involves comparing production costs in two different countries may require both analytical and computation skills. In a CAT instrument for architects, a test

taker may be asked to watch a clip from a videodisk and then to perform a task. The items may therefore require both reading and spatial abilities. For such examples of CAT items, it is unlikely that student achievement can be captured by a single dimension of proficiency, and multidimensional models (e.g., Reckase, 1997) should be employed. Given the rather sparse literature on multidimensional IRT (MIRT) CAT, our proposed procedure should be of interest to both CAT researchers and practitioners.

The remainder of this paper is organized as follows: First, some background on the multidimensional IRT-based stratification method are provided. Next, we describe the proposed method and discuss some of its theoretical properties. Then, two empirical studies concerning the exposure rate and efficiency of the method are reported. Finally, we present some concluding remarks.

Background

As a prelude to describing the multi-stage stratified procedure for multidimensional items, we would like to briefly provide some background on a basic MIRT model and a commonly used criterion for item selection.

Multidimensional item response models

A MIRT model posits that a subject's responses to test items are driven by a latent set of multiple abilities. Let $\theta = \{\theta_1, \theta_2, \ldots, \theta_d\}$, where $d > 1$, denotes the d-dimensional latent ability and Y denotes the dichotomous item score on a test item (1=correct, 0=incorrect). One commonly used form for the item response function is given by the two-parameter logistic (2PL) model:

$$p(\theta) := P(Y = 1|\theta) = \frac{e^{Da'(\theta - b\mathbf{1})}}{1 + e^{Da'(\theta - b\mathbf{1})}}, \tag{1}$$

where $\mathbf{1}$ is a $d \times 1$ vector of $1'$s and $D = 1.7$ is the constant used for aligning the logistic and probit link functions (Lord, 1980). Note that the exponent in (1) can be expressed in scalar notation by

$$Da'(\theta - b\mathbf{1}) = D\sum_{i=1}^{d} a_i(\theta_i - b).$$

Equation (1) is sometimes called a compensatory MIRT model (Hattie, 1981; Reckase, 1997). The compensatory model assumes that a deficit of ability in one dimension can be

compensated for by proficiency in another dimension. It is especially useful when more than one strategy can be applied to arrive at the correct answer to an item (Reckase, 1997). The model is characterized by two parameters: the vector of discrimination parameter a, and the scalar difficulty parameter b. Note that the compensatory model (1) contains a single difficulty parameter b rather than separate difficulty parameters for each dimension. Although separate difficulty parameters are conceptually plausible, they are indeterminate and thus cannot be estimated from observed response data unless constraints are imposed (Segall, 1996).

A stratified recursive maximum likelihood method

In contrast to a paper-and-pencil test, CAT adaptively selects items from an item pool and then sequentially presents them to the test taker. The way in which an item is selected is generally a function of the estimated ability of the examinee at that point of selection and of the characteristics of the items not yet selected in the item pool. In unidimensional tests, a commonly used method for item selection is based on the maximum Fisher information criterion. Suppose that K items have been administered to an examinee. For a one-dimensional 2PL model, the likelihood function for the latent ability θ is given by :

$$L_K(\theta) = \prod_{i=1}^{K} [p_i(\theta)]^{y_i} [q_i(\theta)]^{1-y_i},$$

where $q_i = 1 - p_i$, and

$$p_i(\theta) := P(Y_i = 1|\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}.$$

The Fisher information about θ is given by

$$I := -\mathrm{E}\Big[\frac{\partial^2 \ln L_K}{\partial \theta^2}\Big] = D^2 \sum_{i=1}^{K} a_i^2 p_i(\theta) q_i(\theta).$$

The next, or the $(K+1)$th item chosen will be the one that maximizes the Fisher information $D^2 \sum_{i=1}^{K+1} a_i^2 p_i(\theta) q_i(\theta)$. This criterion for sequentially choosing items is called D-optimality. Because θ is an unknown quantity, we use the maximum likelihood estimate $\hat{\theta}$ to replace θ in computing the Fisher information. After the $(K+1)$th item is administered and a response is solicited from the examinee, the response will be scored and the maximum likelihood estimate for the specific examinee will be updated. This recursively updating procedure is known as the recursive maximum likelihood estimation (R-MLE) method.

When the test is multidimensional and $d = 2$, the R-MLE method can be generalized as follows. Suppose that $\theta = (\theta_1, \theta_2)$, $a = (a_1, a_2)$, and the unidimensional 2PL model is replaced by its multidimensional counterpart (1). In this case, the the information matrix is given by

$$\begin{aligned} I_{2\times 2} &:= -\mathrm{E}\Big[\frac{\partial^2 \ln L_K}{\partial\theta\partial\theta^T}\Big] \\ &= \begin{bmatrix} D^2 \sum_{i=1}^{K} a_{1i}^2 p_i q_i & D^2 \sum_{i=1}^{K} a_{1i}a_{2i} p_i q_i \\ D^2 \sum_{i=1}^{K} a_{2i}a_{1i} p_i q_i & D^2 \sum_{i=1}^{K} a_{2i}^2 p_i q_i \end{bmatrix}. \end{aligned}$$

The Fisher information matrix for $d > 2$ is analogous. In general, $I = (I_{rs})$ is given by

$$I_{rs} = -\mathrm{E}\Big[\frac{\partial^2 \ln L_K}{\partial\theta_r\partial\theta_s}\Big] = D^2 \sum_{i=1}^{K} a_{si}a_{ri}p_i q_i.$$

The condition of D-optimality can be extended to higher dimension by using the determinant of the information matrix I. For $d = 2$,

$$|I_{2\times 2}| = D^4\Big\{ \sum_{i=1}^{K} a_{1i}^2 p_i q_i \cdot \sum_{i=1}^{K} a_{2i}^2 p_i q_i - (\sum_{i=1}^{K} a_{1i}a_{2i} p_i q_i)^2 \Big\}. \tag{2}$$

Based on results from simulation studies, Segall (1996) and Miller, Reckase, Spray, Luecht, and Davey (in press) reported that the determinant $|I|$ of the Fisher information matrix was judged to be a more precise measurement of information than the Euclidean distance. Therefore, in the following discussion, we shall use the determinant of the Fisher information as the objective function for item selection.

Method

The proposed method for controlling item exposure uses a functional of a in stratifying the item pool. Without a controlling mechanism, an item selection procedure solely based on the D-optimal criterion tends to only choose "good" items. When $d = 1$, this means that items with a high a parameter are likely to be selected. The Fisher information for the response function is monotone increasing with respect to the a parameter of the selected item (Chang and Ying, 1999). Thus D-optimality could easily lead to overexposure of items with high discrimination parameters. In order to balance item exposure rates, Chang and Ying (1999) proposed partitioning the item pool according to levels of a and then selecting items from each stratum as the test progresses. While it is possible to stratify the item pool according to the

values of the b (difficulty) parameter, such a scheme would likely result in administering items of diverse difficulties to a student regardless of the individual's estimated ability. Such a test would become rather inefficient and defeat the purpose of computerized adaptive testing.

When $d > 1$, the logic for applying a stratification scheme that is based on the a parameter is similar to when $d = 1$. Indeed, it can be seen from (2) that the effect of b is contained in $p_i(\theta)$ and $q_i(\theta)$, where θ is estimated by $\hat{\theta}$. As the test progresses, $\hat{\theta}$ becomes stable, and $|I|$ is primarily a function of the vector a. Therefore, the challenge is to judiciously select an appropriate functional of a that assures that the operational characteristics of the a-stratification scheme for a multidimensional item pool are analogous to its unidimensional counterpart.

We propose an a-stratified multi-stage item selection procedure as follows:

1. Partition the item bank into M levels according to the a-parameter values of items. The first item stratum contains items with the smallest $|a_2 - a_1|$'s, the next stratum contains items with second smallest $|a_2 - a_1|$'s, and so on.

2. Accordingly, partition the test into M stages.

3. In the j-th stage, $j = 1, \cdots, M$, sequentially select from the j-th stratum n_j items using the following criterion. For each point of selection, choose the item with difficulty parameter b with a value that is closest to:

$$\hat{b} = \frac{a_1\hat{\theta}_1 + a_2\hat{\theta}_2}{a_1 + a_2}. \tag{3}$$

   In other words, match the item difficulty parameter b to the recursively updated estimator $\hat{\theta}$.

4. Repeat Step 3 for $j = 1, 2, \ldots, M$.

Operationally, the values of M, n_j, $j = 1, \cdots, M$, are determined prior to the test. The test can also be designed to terminate when $\hat{\theta}$ reaches a prespecified level of accuracy.

A justification for using the functional $|a_2 - a_1|$ in the stratification scheme is in order. The matching of the difficulty parameter to the estimated ability also requires further explanation. We will first provide a heuristic argument for using the functional $|a_2 - a_1|$. Then, in the next two sections we provide both a theoretical and an empirical investigation in order to

substantiate and validate this choice of the functional. Details of matching b to estimated ability will be provided in the next section.

Consider the following two extreme cases: $|a_2 - a_1| \approx 0$, and $a_1 >> a_2$ (or $a_2 >> a_1$). For the sake of convenience, assume that the former case occurs at the first stratum. Then

$$Z \equiv D[a_1(\theta_1 - b) + a_2(\theta_2 - b)] = Da_1[(\theta_1 - b) + (\theta_2 - b)] \approx 2Da_1\left[\frac{\theta_1 + \theta_2}{2} - b\right].$$

This implies that

$$p(\theta) = \frac{e^Z}{1 + e^Z} \approx \frac{e^{2Da_1[(\theta_1+\theta_2)/2-b]}}{1 + e^{2Da_1[(\theta_1+\theta_2)/2-b]}}.$$

Thus, the 2-dimensional ability space has been reduced to a single dimension along the direction $(\theta_1 + \theta_2)/2$. Consequently, only the mean can be estimated, not the individual dimensions. In such a situation, the information $|I|$ is at a minimum. Therefore, when an item comes from the first stratum, the information content lies along the direction of the average of the two dimensions, and one gets the least overall information from the item.

When the values of a_1 and a_2 are highly divergent, information about θ is asymmetric. Assume that a_2 is fixed and is taken to be the smallest value in the item pool, and $a_1 >> a_2$. Then

$$Z \equiv D[a_1(\theta_1 - b) + a_2(\theta_2 - b)] \approx D[a_1(\theta_1 - b)].$$

Therefore, choosing an item from the stratified item pool would be analogous to the unidimensional case, with the single dimension being θ_1. Maximum information about θ_1 occurs at the largest value of a_1. Conversely, when $a_2 >> a_1$, maximum information occurs at the largest value of a_2.

The above heuristic suggests that the functional $|a_2 - a_1|$ mimics the behavior of $|a|$ (equivalently a, since $a > 0$) in a unidimensional test. At an early stage of the test, the stratification scheme selects items that try to measure the "average" of the multiple ability dimensions. As the test progresses, items that increasingly pinpoint the location of the true ability along each individual dimension are used. The incremental information content (measured by the determinant of the Fisher information matrix) increases when one transverses across the strata that are ordered according to $|a_2 - a_1|$.

We set up the mathematical framework for selecting the $(K+1)$th item based on the assumption that K items have already been administered. From (2), define the constants A, B and C:

$$A = \sum_{i=1}^{K} a_{1i}^2 p_i q_i,\ B = \sum_{i=1}^{K} a_{2i}^2 p_i q_i,\ C = \sum_{i=1}^{K} a_{1i} a_{2i} p_i q_i,$$

Let (a_1, a_2), b, and p, respectively, denote the discrimination parameter, the difficulty parameter of the $(K+1)$th item, and the probability of correctly responding to the item. The information accumulated up to the $(K+1)$th item is then given by:

$$|I(a_1, a_2, b)| = D^4 \left\{ AB - C^2 + pq(Aa_2^2 + Ba_1^2 - 2Ca_1a_2) \right\}, \tag{4}$$

Note that p and q are both functions of a_1, a_2 and b. Let the domains of variation of a_1,a_2, and b be denoted by D_{a_1}, D_{a_2}, and D_b. For the objective function (4), the optimal (a_1, a_2, b) could either occur at the boundaries or in the interior of the domains. Because the function $|I|$ is smooth in (a_1, a_2, b), it is sufficient for investigating the properties of the partial derivatives of $|I|$ with respect to each of the variables. Specifically, if not all coordinates attain maxima at the boundaries, (a_1, a_2, b) must satisfy some or all of the following equations :

$$\frac{\partial |I|}{\partial a_1} = 0,\ \frac{\partial |I|}{\partial a_2} = 0,\ \text{and}\ \frac{\partial |I|}{\partial b} = 0. \tag{5}$$

By studying the properties of the derivatives, we prove the following results :

1. It is not possible for the optimum to occur in an interior point that simultaneously satisfies all three equations in (5).

2. The trajectory of the optimal value of (a_1, a_2) follows an ellipse or a hyperbola.

3. The values of (a_1, a_2) that maximize the objective function $|I|$ occur at the boundaries of the domain D_{a_1} and D_{a_2}, respectively. Specifically, either the value a_1 is largest in D_{a_1} and a_2 is smallest in D_{a_2}, or vice versa. The outcome depends upon the characteristics of the previously administered items and is a function of the constants A, B, and C.

4. Under the above conditions, the optimal value $\hat{b}$ for the difficulty parameter b is given by (3).

The above results show that for a 2-dimensional test, information increases as $|a_2 - a_1|$ increases. This is analogous to the way information increases as $|a|$ increases in the 1-dimensional case. Therefore, when an item pool is stratified according to the functional $|a_2 - a_1|$, it exhibits properties similar to when a unidimensional item pool is stratified according to $|a|$. Our theoretical investigation provides the basis for the proposed multidimensional stratification strategy. The details of the proof and an illustrative example are developed in the appendix.

Empirical Studies

In this section, we present two empirical studies designed to validate the use of the proposed a-stratified procedure. The first study compares the item exposure rates of the proposed and the traditional methods. In the latter method, the item pool is not stratified, and the selection of items is strictly based on the D-optimal criterion. While the stratified scheme may be able to reduce the problem of item over-exposure, there still exists a trade-off between the extent to which items are evenly chosen and the efficiency in estimating student ability. In general, when the item pool is stratified, the selected item is sub-optimal. Accordingly, a larger number of items is necessary to arrive at the same degree of accuracy for estimating ability than when using optimal items. Therefore, any procedure that uses a sub-optimal item selection procedure should always be benchmarked against the optimal procedure with respect to relative efficiency. After all, one important motivation of CAT is to be able to estimate ability using a small number of items. To this end, the second study empirically compares the efficiency of the proposed method with the optimal procedure and a third commonly used procedure.

In the first study, data were simulated from a multidimensional 2PL model using bivariate normal theta distribution and a set of MIRT ($d = 2$) item parameters from another study (Ackerman, 1988). The values of the MIRT item parameters, which were also reported in Miller (1991), were selected to provide uniform information over the ability continuum. Fifty items were used in the simulation, and the correlation between the two dimensions was set at zero. A total of 120 examinees were simulated.

For the proposed method, we first ordered $|a_1 - a_2|$ and partition them into five equal strata. Then a total of 10 items – two from each stratum – were selected in accordance with the proposed procedure and administered to each subject. Finally, for each item, the exposure

rate, defined as the ratio between the number of times the item was selected and the total number of examinees, was computed.

TABLE 1 Exposure rates of traditional, proposed, and refined methods.

Item	Traditional	Proposed	Refined	Item	Traditional	Proposed	Refined
1	0.0000	0.1833	0.0750	26	0.3250	0.1500	0.3000
2	0.0000	0.2917	0.2333	27	0.3500	0.2167	0.0583
3	0.0000	0.2000	0.1583	28	0.3667	0.2083	0.1583
4	0.0000	0.2500	0.2333	29	0.3500	0.1833	0.1500
5	0.0083	0.1583	0.2583	30	0.4417	0.2250	0.2500
6	0.0000	0.1583	0.2917	31	0.3083	0.2583	0.0750
7	0.0083	0.2417	0.0583	32	0.3750	0.2417	0.1750
8	0.0083	0.2167	0.2833	33	0.3917	0.3000	0.0667
9	0.0000	0.1750	0.1667	34	0.3167	0.2000	0.2583
10	0.0000	0.1250	0.2417	35	0.3833	0.0833	0.4250
11	0.0167	0.1583	0.2000	36	0.3500	0.2083	0.1417
12	0.0000	0.1417	0.0500	37	0.4250	0.1417	0.0583
13	0.0167	0.2167	0.1417	38	0.3750	0.2750	0.4500
14	0.0333	0.2917	0.2250	39	0.3333	0.1917	0.0417
15	0.0333	0.1750	0.0750	40	0.4250	0.1000	0.3083
16	0.0250	0.2083	0.4417	41	0.2333	0.2083	0.0667
17	0.0167	0.2167	0.0583	42	0.3583	0.2000	0.2083
18	0.0333	0.2083	0.0417	43	0.4167	0.1333	0.0333
19	0.0167	0.1250	0.2583	44	0.4583	0.1833	0.1417
20	0.0500	0.2583	0.5083	45	0.3417	0.2417	0.0417
21	0.0583	0.1417	0.0917	46	0.4000	0.2333	0.2417
22	0.0500	0.2417	0.3917	47	0.4167	0.1083	0.3750
23	0.0250	0.0833	0.0417	48	0.4250	0.1583	0.2333
24	0.1167	0.2167	0.0833	49	0.3583	0.2417	0.2667
25	0.0750	0.3333	0.4750	50	0.4833	0.2917	0.3917

Table 1 includes the exposure rates of the proposed and traditional methods over the 50 items. For the traditional method, 10 percent of the item pool accounted for 43 percent of all items used. These figures agree with Wainer's observation mentioned earlier. Furthermore, 46 percent of the items were used less than 5 percent of the time. The asymmetry – both the under-utilization and the overexposure of some parts of the item pool – clearly points to the wastage of valuable resources (CAT items are expensive to develop) and the potential security hazard. On the other hand, for the proposed stratified method, exposure rates were rather uniform throughout the item pool.

Table 1 also reports the exposure rate of a "refined" version of the proposed method. In the refined version, given a stratum, we substitute equation (3) into (4) so that $|I(a_1, a_2, b(a_1, a_2)|$ is

only a function of (a_1, a_2). Then we use the D-optimality criterion to select the best item within the stratum. In other words, we select the item with the largest value of $|I|$. Conceptually, the refined version is appealing because it may hold both advantages of being less susceptible to overexposure than the traditional method and more efficient than the proposed method. However, the results in Table 1 shows that the exposure rate of the refined version is not satisfactory – out of 50 items, 17 were used less than 10 percent of the time (compared to 24 and 0 items, respectively for the traditional and the proposed methods.) Computationally, the refined method is also rather expensive. Based on these two reasons, we do not recommend using the refined version.

In the second empirical study, we compared the efficiency of the proposed a-stratified procedure to that of two other commonly used methods – the traditional (optimal) method and the random method. The random method selects an item randomly from the item pool and administers it to the examinee. In other words, at each iteration of the sequential procedure in CAT, every item in the item pool has an equal probability of being selected.

In this study, we designed six strata of items in such a way that different combinations of items within each stratum would satisfy the condition $|a_1 - a_2| = c$, where $c = 0, 0.5, 1.0, 1.5, 2.0, 2.5$, respectively. As a result, the item pool contains a total of $8,296$ items. Table 2 shows the structure of the six strata.

TABLE 2 The structure of item pool in simulation study for efficiency.

Stratum	Combination of values of a_1,a_2,b
1	$(0.5 + 0.1 \times i, 0.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 25$; $j = 0, 1, \ldots, 60$
2	$(0.5 + 0.1 \times i, 1.0 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 20$; $j = 0, 1, \ldots, 60$, and $(1.0 + 0.1 \times i, 0.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 20$; $j = 0, 1, \ldots, 60$
3	$(0.5 + 0.1 \times i, 1.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 15$; $j = 0, 1, \ldots, 60$, and $(1.5 + 0.1 \times i, 0.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 15$; $j = 0, 1, \ldots, 60$
4	$(0.5 + 0.1 \times i, 2.0 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 10$; $j = 0, 1, \ldots, 60$, and $(2.0 + 0.1 \times i, 0.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 10$; $j = 0, 1, \ldots, 60$
5	$(0.5 + 0.1 \times i, 2.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 5$; $j = 0, 1, \ldots, 60$, and $(2.5 + 0.1 \times i, 0.5 + 0.1 \times i, -3.0 + 0.1 \times j)$, for all $i = 0, 1, \ldots, 5$; $j = 0, 1, \ldots, 60$
6	$(0.5, 3.0, -3.0 + 0.1 \times j)$, for all $j = 0, 1, \ldots, 60$, and $(3.0, 0.5, -3.0 + 0.1 \times j)$, for all $j = 0, 1, \ldots, 60$

A sample of $1,000$ examinees was generated from a standard bivariate normal distribution. Then a sample of n items was administered to each student using the three methods: stratified, traditional, and random. Subsequently, the maximum likelihood estimate for ability for each examinee at the end of the test of length n was compared to its population value. Finally, we computed the mean square error (MSE), which is given by :

$$\text{MSE} \quad = \quad \frac{1}{1000} \sum_{i=1}^{1000} [(\hat{\theta}_{1i} - \theta_{1i})^2 + (\hat{\theta}_{2i} - \theta_{2i})^2], \tag{6}$$

where $(\theta_{1i}, \theta_{2i})$ is the ability of examinee i.

TABLE 3 Mean square errors by methods, length of test, and examinee ability.

	MSE				
(θ_1, θ_2)	(0.0,0.3)	(1.4,1.7)	(2.7,3.0)	(0.3,2.7)	(2.7,0.3)
n=10					
Traditional	0.821	0.795	0.719	1.021	0.978
Proposed	1.024	2.701	2.952	4.726	2.147
Random	10.274	13.281	14.572	11.071	9.823
n=30					
Traditional	0.552	0.432	0.378	0.624	0.541
Proposed	0.597	0.517	0.407	0.691	0.620
Random	9.204	8.771	14.172	7.210	5.982
n=50					
Traditional	0.148	0.104	0.081	0.274	0.281
Proposed	0.157	0.195	0.087	0.277	0.279
Random	5.842	6.217	8.721	5.716	4.517
n=100					
Traditional	0.064	0.045	0.026	0.068	0.057
Proposed	0.061	0.046	0.029	0.075	0.058
Random	4.718	4.957	5.927	4.182	3.574

Table 3 contains the MSE values cross-classified according to the following factors: (1) method (three levels: traditional, proposed, and random), (2) difference in abilities on the two dimensions (five levels of $|\theta_1 - \theta_2|$), and (3) total number of items administered (four levels: $n = 10, 30, 50, 100$). The result indicates that the proposed method is relatively efficient compared with the traditional method. There is a slight loss in efficiency when $n = 30$ or above, and a modest loss when $n = 10$. For example, when $n = 30$, the loss in inefficiency is approximately 10 percent across ability levels. On the other hand, the loss in efficiency for the random method is quite large, even when $n \geq 30$.

We also examined MSE and separately examine its two components – bias and variance (MSE=bias^2+variance). Our result (not reported here) shows that variance dominates bias.

This implies that if MSE for the proposed method is x percent higher than the traditional method, then approximately x percent more items are required to achieve the same degree of accuracy. For example, suppose MSE for the proposed method is 10 percent higher than the traditional method. Then when a traditional CAT exam uses $n = 30$, the proposed method needs to use $n = 33$ to achieve approximately the same degree of accuracy for estimating student ability.

Conclusion

Several conclusions can be drawn concerning the use of the a-stratified procedure for multidimensional tests. First, the proposed procedure has operational characteristics similar to those in the unidimensional procedure proposed in Chang and Ying (1999). Second, the exposure rate can substantially improve when stratification is based upon a judiciously chosen functional of the discrimination parameter. Finally, the loss of efficiency of the proposed procedure is relatively mild when the number of test items administered reaches 30.

Although the last two conclusions are based upon simulation studies, there is no reason to expect different results for tests with operational characteristics similar to the experimental conditions. For example, when a large item pool is used and the test is not too short, we would expect the loss in efficiency to be minimal. While we believe that our results can be generally applied to multidimensional tests under standard conditions, some open problems still remain. The primary issue is solving the trade-off between the over-exposure of some items and the efficiency of the instrument in multidimensional tests. For example, there are the questions: How many strata should be used? How can the procedure be generalized to $d > 2$? and What will be the pattern of the trade-off? The first question is important because when the number of strata is small some items within a stratum may still be over-used. When the number of strata is too large (the extreme case being one item per stratum), it is possible that some strata are over-used. Some work has been done in this direction (e.g., Parshall, Harmes, and Kromrey, 2000). For the second question, we have conducted some simulation experiments for $d = 3$. Our initial results indicated that the procedure is still satisfactory. The results are reported in a separate working paper and are available from the corresponding author.

Appendix: Proof of theoretical results

The equations of derivative with respect to a_1, a_2 and b in (5), respectively, yield:

$$2(Ba_1 - Ca_2) + D(\theta_1 - b)(1-2p)\left\{Aa_2^2 + Ba_1^2 - 2Ca_1a_2\right\} = 0, \tag{A.1}$$

$$2(Aa_2 - Ca_1) + D(\theta_2 - b)(1-2p)\left\{Aa_2^2 + Ba_1^2 - 2Ca_1a_2\right\} = 0, \tag{A.2}$$

and

$$(a_1 + a_2)(1-2p)\left\{Aa_2^2 + Ba_1^2 - 2Ca_1a_2\right\} = 0. \tag{A.3}$$

In the following, we discuss the several cases when the all or part of the three equations (A.1)-(A.3) hold.

Case 1. When (A.1)-(A.3) simultaneously hold, first consider the situation in which (A.3) holds. There are two possibilities:

(a) $Aa_2^2 + Ba_1^2 - 2Ca_1a_2 = 0$. Substitute into (A.1) and (A.2), we get

$$\begin{array}{rcl} Ba_1 & = & Ca_2, \\ Aa_2 & = & Ca_1. \end{array} \tag{A.4}$$

The equations in (A.4) imply that

$$a_1a_2(AB - C^2) = 0. \tag{A.5}$$

Because A, B, C are constants that only depend on the previously administered K items, we get $a_1 = 0$ or $a_2 = 0$, which contradicts the assumption that neither a_1 nor a_2 can be zero. Therefore this case is not possible.

(b) If $1 - 2p = 0$, we can immediately obtain equation (A.5) from (A.1) and (A.2). By using the same argument as in part (a), this is not possible.

Case 2. (A.1)-(A.3) do not simultaneously hold. There are two possibilities:

(a) (A.1) and (A.2) hold, but (A.3) does not. We show mathematically that this case cannot occur. Assume a_1 and a_2 are given so that $|I|$ is a function of b only. ¿From (4), the only factor dependent upon b is pq. The maximum of pq occurs at $p = 0.5$. Therefore, there exists some b at which both $p = 0.5$ and the derivative of $|I(b)|$ is zero. This contradicts our assumption.

(b) Suppose (A.3) holds but (A.1) and (A.2) do not. (A.3) holds when $1-2p=0$, and this implies that $p=0.5$. Substitute p into the following equation

$$p=\frac{e^{D[a_1(\theta_1-b)+a_2(\theta_2-b)]}}{1+e^{D[a_1(\theta_1-b)+a_2(\theta_2-b)]}},$$

leads to $\exp\{D[a_1(\theta_1-b)+a_2(\theta_2-b)]\}=1$, or $a_1(\theta_1-b)+a_2(\theta_2-b)=0$, which implies that b is given by (3). That is, if we can obtain optimal values for a_1 and a_2, the optimal value of b can also be determined.

When (A.1) and (A.2) do not hold, we need to directly examine the objective function $|I|$ for a given b. Indeed, we can prove that the maxima of $|I|$ with respect to a_1 and a_2 occur on the boundaries. For the sake of convenience, we assume that the lower (upper) boundaries for both dimensions are L (U).

To maximize $|I|$ is equivalent to maximizing $f(a_1,a_2)=Aa_2^2-2Ca_1a_2+Ba_1^2$. Define $\Delta:=(-2C)^2-4AB=4(C^2-AB)$. Slightly rewriting $f(a_1,a_2)$, we have

$$f(a_1,a_2)=\frac{1}{A}[(Aa_2-Ca_1)^2-(C^2-AB)a_1^2]. \tag{A.6}$$

By Cauchy inequality,

$$\Big(\sum_{i=1}^{K}a_{1i}^2p_iq_i\Big)\Big(\sum_{i=1}^{K}a_{2i}^2p_iq_i\Big)\geq\Big(\sum_{i=1}^{K}a_{1i}a_{2i}p_iq_i\Big)^2 \quad\Rightarrow\quad AB\geq C^2. \tag{A.7}$$

The inequality in (A.7) is strict. Hence, the condition $C^2-AB\leq 0$ automatically holds.

1. If $\Delta>0$, i.e. $C^2-AB>0$, it contradicts equation (A.7), so this case is not possible.

2. If $\Delta=0$, i.e. $C^2-AB=0$,

$$a_1a_2(C^2-AB)=0\ \Rightarrow\ Ba_1-Ca_2=0\ or\ Aa_2-Ca_1=0.$$

Thus either (A.1) or (A.2) holds, which contradicts our assumption.

3. If $\Delta<0$, i.e. $C^2-AB<0$, then for any a_1 and a_2, $f(a_1,a_2)>0$. Therefore, given a positive value R of $f(a_1,a_2)$, the curve $f(a_1,a_2)=Aa_2^2-2Ca_1a_2+Ba_1^2=R$ is an ellipse whose major axis lies on the line with slope $\tan\alpha$, where

$$\cot 2\alpha=\frac{A-B}{-2C},\ \alpha\in(0,\frac{\pi}{2}).$$

Standardizing the ellipse gives

$$\frac{a_2'^2}{R/E}+\frac{a_1'^2}{R/F}=1, \tag{A.8}$$

where (a_1', a_2') is the transformed coordinates, E and F are coefficients after standardization, and $\sqrt{R/E}$ and $\sqrt{R/F}$ are the length of major axis and minor axis, respectively.

(i) If $A > B$, $\cot 2\alpha < 0 \;\Rightarrow\; \alpha \in (\pi/4,\ \pi/2)$. From equation (A.8), the larger R is, the longer the major and minor axis are. Therefore, the ellipse that passes through the point $(a_1, a_2) = (L_+, U_-)$ is the largest optimal ellipse.

(ii) If $B > A$, $\cot 2\alpha > 0 \;\Rightarrow\; \alpha \in (0,\ \pi/4)$. By the same argument in (i), the largest ellipse is optimal, and it passes through the point $(a_1, a_2) = (U_-, L_+)$.

(iii) If $A = B$, $\cot 2\alpha = 0 \;\Rightarrow\; \alpha = \pi/4$. By the same argument in (i), the largest ellipse is optimal, and it passes through either the point $(a_1, a_2) = (L_+, U_-)$ or the point $(a_1, a_2) = (U_-, L_+)$.

As an illustration, Figure A1 shows the surfaces of $|I|$ under two scenarios $A < B$ and $A > B$. The optimal points both occur at the boundaries of (a_1, a_2).

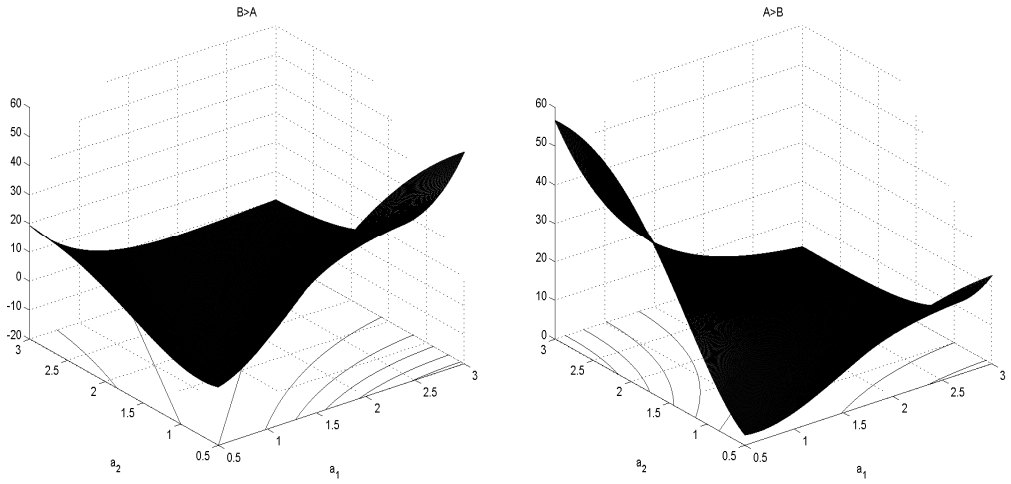
We performed a small-scale experiment to verify the above results. In the experiment, items parameters were generated such that the boundaries of a_1, a_2 were given by $L = 0.5$, $U = 3.0$. The range of b was $(-3, 3)$. In the experiment, we selected several examinees that cover a range of values of θ. We only consider positive value of θ because the distribution of θ is symmetric about 0. We assumed that each examinee was administered K items, and the D-optimal criterion was used to select the $(K+1)$th item.

Table A1 shows the experimental results on the characteristics of the $(K+1)$th item. The results indicates that the optimal a values do occur at the boundaries of the domains. Furthermore, which item was selected is a function of the relative magnitude of A or B. The results support the theory that we developed above.

TABLE A.1: Characteristics of the selected ($(K+1)$th) item.

		$A > B$	$B > A$
$K = 5$	$\theta = (1.7, 2.3)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.1$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.8$
	$\theta = (2.4, 2.4)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.4$	$a_1 = 3,\ a_2 = 0.5,\ b = 2.4$
	$\theta = (1.7, 1.2)$	$a_1 = 0.5,\ a_2 = 3,\ b = 1.4$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.6$
$K = 10$	$\theta = (1.7, 2.3)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.2$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.8$
	$\theta = (2.4, 2.4)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.4$	$a_1 = 3,\ a_2 = 0.5,\ b = 2.4$
	$\theta = (1.7, 1.2)$	$a_1 = 0.5,\ a_2 = 3,\ b = 1.3$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.6$
$K = 20$	$\theta = (1.7, 2.3)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.2$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.8$
	$\theta = (2.4, 2.4)$	$a_1 = 0.5,\ a_2 = 3,\ b = 2.4$	$a_1 = 3,\ a_2 = 0.5,\ b = 2.4$
	$\theta = (1.7, 1.2)$	$a_1 = 0.5,\ a_2 = 3,\ b = 1.3$	$a_1 = 3,\ a_2 = 0.5,\ b = 1.6$

Figure A 1: Graph showing surface of objective function $|I|$ as a function of (a_1, a_2) for two scenarios : $B > A$, $A > B$.



References

Ackerman, T. A. (1988, May). *Comparison of multidimensional IRT estimation procedures using benchmark data*. Paper presented at the ONR Contractor's meeting, Iowa City, IA.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, **34**, 197-211.

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

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67, 387-398.

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

Hattie, J. (1981). *Decision criteria for determining unidimensionality*. Unpublished doctoral dissertation, University of Toronto, Canada.

Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.

Lord, M. F. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, **31**, 3-31.

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

Miller, T. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program* (Research report ONR91-2). The American College Testing Program.

Miller, T., Reckase, M. D., Spray, J. A., Luecht, R., & Davey, T.. *Multidimensional item response theory*. (in press).

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, 28-52.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In van der Linden, W. J., & Hambleton, R. K. (Eds.), *Handbook of Modern item Response Theory*, 271-286. New York: Springer-Verlag.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized Adaptive Testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, **61**, 331-354.

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

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, **22**, 259-270.

Wainer, H. (2000a) Rescuing computerized testing by breaking Zipf's Law. *Journal of Educational and Behavioral Statistics*, **25**, 203-224.

Wainer, H. (2000b). CATs: Whither and whence. *Psicólogica*, 121-133.

Wainer, H., Dorans, D. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized Adaptive Testing: A primer* (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.