A Continuous α -Stratification Index for Item Exposure Control in Computerized Adaptive Testing

Applied Psychological Measurement
2018, Vol. 42(7) 523–537
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618758289
journals.sagepub.com/home/apm

Alan Huebner¹, Chun Wang², Bridget Daly³,
and Colleen Pinkelman⁴

Abstract

The method of α -stratification aims to reduce item overexposure in computerized adaptive testing, as items that are administered at very high rates may threaten the validity of test scores. In existing methods of α -stratification, the item bank is partitioned into a fixed number of nonoverlapping strata according to the items' α , or discrimination, parameters. This article introduces a continuous α -stratification index which incorporates exposure control into the item selection index itself and thus eliminates the need for fixed discrete strata. The new continuous α -stratification index is compared with existing stratification methods via simulation studies in terms of ability estimation bias, mean squared error, and control of item exposure rates.

Keywords

computerized adaptive testing, item exposure control, item response theory

Introduction

Computerized adaptive tests (CATs) have become increasingly popular since the 1990s as a means of administering exams for a wide variety of purposes, including for elementary and secondary schools, government agencies, professional organizations, and graduate and professional schools. In many of these applications, preventing items from being overexposed is an important practical consideration needed to ensure the validity of test scores, as overexposed items have a high risk of being shared among current and future examinees. Thus, it is desirable to have a method of item exposure control to decrease the number of times the most informative items are administered to examinees. While it is possible to construct a CAT under a variety of different measurement models, a voluminous body of research exists on CATs based on item

¹University of Notre Dame, IN, USA

²University of Minnesota, Minneapolis, MN, USA

³84.51°, Cincinnati, OH, USA

⁴Amazon.com, Inc., Seattle, WA, USA

Corresponding Author:

Alan Huebner, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 153 Hurley Hall, Notre Dame, IN 46556, USA.

Email: Alan.Huebner.10@nd.edu

response theory (IRT) models. Georgiadou, Triantafyllou, and Economides (2007) provide summaries and taxonomy of many different methods of item exposure control for IRT-based CATs.

One effective and flexible method of item exposure control is a -stratification, first proposed by Chang and Ying (1999). The method was subsequently expanded in a variety of studies for CATs based on both one- and two-dimensional (hereafter, 1-DIM and 2-DIM) IRT models; for example, see Chang, Qian, and Ying (2001); Leung, Chang, and Hau (2002); Chang and van der Linden (2003); Lee, Ip, and Fuh (2008); and Huebner, Wang, Quinlan, and Seubert (2015). These methods divide the item bank into K strata, with Stratum 1 containing items with the smallest a values, Stratum 2 containing items with the next smallest a values, and so on, up to Stratum K containing items with the largest a values. The method of a -stratification is simple and easy to implement, and it has been shown to reduce exposure of highly informative items while sacrificing measurement accuracy and precision to a degree that many practitioners would likely view as negligible in many testing situations.

In these previous approaches, K has been selected to be a moderate-size fixed constant, with $K=4$ being a common choice for the simulation studies in the articles cited above. The strata are exhaustive and mutually exclusive; that is, each item in the bank must belong to one and only one stratum. In an operational setting, items would likely be assigned to a stratum when the item bank is created, and simulation studies would be required to choose an appropriate K value balancing exposure control and measurement accuracy and precision. In addition, it is not clear how eliminating old items and introducing new ones would be handled. Plausibly, the strata would be recreated every time the item bank is replenished. This would be undesirable from a practical standpoint because each time an item bank is refreshed, a simulation study would need to be conducted to determine the optimal K . Furthermore, it is even more challenging if new items were frequently added and overexposed items frequently eliminated from an item bank. The latter scenario may occur in high-stakes tests when item usage is closely monitored. The current article proposes an alternate approach: a continuous a -stratification index (CAI), which incorporates exposure control as one building block intrinsic to the index itself, therefore displaying greater flexibility when dealing with operational item banks. In this method, each individual item is ranked according to the CAI; there is no need to choose the number of strata.

The remainder of the article describes the proposed CAI method and presents a simulation study in comparison with discrete a -stratification. Specifically, the simulation study examines the performance of the CAI and two static a -stratification methods as evaluated by the ability estimation accuracy and item exposure rate. In particular, it will be investigated whether the proposed CAI method is comparable with or superior to existing stratification methods in terms of controlling item overexposure while maintaining accuracy and precision of examinee ability estimation. Finally, the discussion section presents limitations of the current study and suggests some avenues for further inquiry.

Method

IRT

To introduce notation and facilitate discussion, 1- and 2-DIM IRT models are briefly reviewed. For both cases, let U be the random variable representing the response to a single item, where values 1 and 0 indicate correct (affirmative) and incorrect (negative) responses, respectively. Then, the vector $\mathbf{U} = (U_1, U_2, \dots, U_j)$ denotes the set of responses to j items for a given examinee. In one dimension, the item response function (IRF) for the three-parameter logistic (3PL) model is given by

$$P_j(\theta) = P_j(U = 1|\theta) = c_j + (1 - c_j) \frac{\exp[1.7a_j(\theta - b_j)]}{1 + \exp[1.7a_j(\theta - b_j)]}, \quad (1)$$

where a_j , b_j , and c_j are the discrimination, difficulty, and pseudo-chance parameters of item j , respectively (Birnbaum, 1968), and θ is the latent trait or ability of the examinee.

For the 2-DIM case, the IRF depends on the vector of examinee abilities $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, the vector of item discriminations for item j , $\mathbf{a}_j = \{a_{j1}, a_{j2}\}$, and the single difficulty parameter b_j . Following the formulation used by Lee et al. (2008), the IRF is given by

$$P_j(\boldsymbol{\theta}) = P_j(U = 1|\boldsymbol{\theta}) = \frac{\exp[1.7\mathbf{a}'_j(\boldsymbol{\theta} - b_j\mathbf{1})]}{1 + \exp[1.7\mathbf{a}'_j(\boldsymbol{\theta} - b_j\mathbf{1})]}. \quad (2)$$

Here, $\mathbf{1}$ is a 2×1 dimensional vector of ones, and the c_j parameter is set to 0 for computational simplicity, an approach commonly taken for multidimensional IRT studies. In both Equations 1 and 2, the scaling constant 1.7 is used to ensure a high level of agreement between the logistic and probit versions of the model (Camilli, 1994).

a-Stratification

For CAT based on 1-DIM IRT, the item bank is partitioned into K strata by the items' discrimination parameters; Chang and Ying (1999) discuss how the size and quality of the item bank can guide the choice of K . The method proceeds as follows: (a) The items with the lowest a values are in Stratum 1, the items with the next lowest a values are in Stratum 2, and so on, until Stratum K is reached containing items with the highest a values. (b) The CAT is accordingly divided into K corresponding stages. (c) J_k items are administered from Stratum k , where $J_1 + \dots + J_K$ equals J , the test length. Items within a stratum are selected via some psychometric criterion. (d) Step 3 is repeated for $k = 1, \dots, K$ strata. Note that in Step 3, previous studies have selected items by maximizing Fisher information at the current $\hat{\theta}$. A simpler "match- b " criterion has also been used in which the next item selected is the one with the smallest absolute difference between $\hat{\theta}$ and its difficulty parameter b .

The extension to 2-DIM IRT-based CAT is more complex owing to the fact that the stratification scheme must take into account both discrimination parameters for each item. The method proposed by Lee et al. (2008) is summarized as follows: (a) The item bank is partitioned into K strata, where items with the smallest $|a_1 - a_2|$ are in the first stratum, items with the next smallest $|a_1 - a_2|$ are in the second stratum, and so on, until stratum K is reached containing items with the largest $|a_1 - a_2|$. (b) For every stratum except the first, two substrata are created, sub-1 and sub-2, where sub-1 contains items with $a_1 > a_2$ and sub-2 contains items for which $a_1 \leq a_2$. (c) The CAT is divided into K corresponding stages. (d) Let J_k be the number of items selected from Stratum k , where $J_1 + \dots + J_K$ equals J , the test length. In Stratum 1, J_1 items are selected using some psychometric criterion. (e) After Stratum 1, within a given stratum items are selected from both sub-1 and sub-2, and the test proceeds through Stratum K until J items are administered. Lee et al. (2008) used two different methods of selecting items within a stratum. One is the D-optimality criterion: At the stage of the exam at which j' items have been administered, the next item ($j' + 1$) is chosen to maximize the quantity:

$$(1.7)^4 \left\{ \sum_{j=1}^{j'+1} a_{j1}^2 P_j(\hat{\boldsymbol{\theta}}) (1 - P_j(\hat{\boldsymbol{\theta}})) \times \sum_{j=1}^{j'+1} a_{j2}^2 P_j(\hat{\boldsymbol{\theta}}) (1 - P_j(\hat{\boldsymbol{\theta}})) - \left(\sum_{j=1}^{j'+1} a_{j1} a_{j2} P_j(\hat{\boldsymbol{\theta}}) (1 - P_j(\hat{\boldsymbol{\theta}})) \right)^2 \right\}. \quad (3)$$

Lee et al. (2008) also proposed a simpler criterion analogous to the 1-DIM match- b criterion; within each stratum at a given test stage, the method chooses the item with difficulty parameter b closest to the quantity:

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

Continuous a -Stratification

In general, CAI measures the “similarity” between the current test stage and the percentile rank of the discrimination parameter, and is thus effectively a type of Gaussian generalization gradient used in cognitive modeling (Busemeyer & Diederich, 2010). More specifically, after j' items have been administered, item $(j' + 1)$ is selected to maximize the quantity:

$$\text{CAI} \times \text{Inf}. \quad (4)$$

In Equation 4, Inf is an information criterion evaluated at the current ability estimate, and CAI is the proposed index. For the remainder of this article, Inf will be for the 1- and 2-DIM cases, respectively, the Fisher information and D-optimality criterion noted above, though other information indices could be used (e.g., the Kullback–Leibler). The CAI formula is presented below for each case. The 1-DIM is first described, and then the 2-DIM is discussed.

For the 1-DIM case, CAI is given by

$$\text{CAI} = \exp \left[-\beta \left(\frac{r(a)}{j'/J} - 1 \right)^2 \right]. \quad (5)$$

In the above expression, $r(a)$ is the percentile rank of the a parameter for a particular item in the bank; its value is bounded by 0 and 1. The term $\frac{r(a)}{j'/J} - 1$ in theory is bounded between -1 and 1 . Because the spirit of CAI is to find $r(a)$ as close as possible to j'/J , the term $\frac{r(a)}{j'/J} - 1$ should be close to 0. $\beta > 0$ is a sensitivity parameter that determines the sensitivity of the discrepancy between j'/J and a in the item selection. In both the 1-DIM and 2-DIM cases, the sensitivity parameter β is determined empirically; it is set via trial and error, and may take a different value from one test to another. It can also be regarded as a weight that reflects the importance of CAI relative to the whole item selection criterion. In general, if β is large, then CAI will become more dominant in the item selection process, and hence exposure control will be more stringent. On the contrary, if β is small, the exposure control constraint is less stringent, and the item information will be dominant. When $\beta = 0$, then Equation 4 becomes purely an information-based index with no exposure control. So overall, higher β yields more balanced item exposure but maybe at the sacrifice of precision. CAI will be maximized when the ratio of the percentile rank and j'/J approaches 1. At the beginning of the test when j'/J is small, items with smaller a values tend to be chosen to maximize CAI, and as the test proceeds, more discriminating items are preferred. Therefore, CAI always forces the discrimination parameters of the selected items to be strictly ascending.

The expression for CAI for the 2-DIM case is very similar to that presented above. The 2-DIM version of CAI is given by

$$\text{CAI} = \exp \left[-\beta \left(\frac{r(\mathbf{a}'\mathbf{a})}{j'/J} - 1 \right)^2 \right]. \quad (6)$$

Table 1. Item Selection by Stratum and Test Stage for the Simulation Study.

Case	Length	Test stages for Strata 1/2/3/4
1-DIM	20	(1-5)/(6-10)/(11-15)/(16-20)
	30	(1-7)/(8-14)/(15-22)/(23-30)
2-DIM	40	(1-10)/(11-20)/(21-30)/(31-40)
	60	(1-15)/(16-30)/(31-45)/(46-60)

Note. 1-DIM = one-dimensional; 2-DIM = two-dimensional.

The reasoning and notation is the same as for one dimension, except that in Equation 6, $\mathbf{a}$ refers to the vector of discrimination parameters for a bank item, and $\mathbf{a}'$ is its transpose. This version is a straightforward generalization from the 1-DIM to multidimensional scenario. The square root of the product term, $\sqrt{\mathbf{a}'\mathbf{a}}$, is the so-called multidimensional discrimination parameter (MDISC; Reckase & McKinley, 1991).

It was also noted that a second version of CAI was examined in which the $\left(\frac{r(a)}{j'/j} - 1\right)^2$ term was replaced with $\left|\frac{r(a)}{j'/j} - 1\right|$ for the 1-DIM and similar for the 2-DIM case. However, in preliminary simulations, this index did not outperform the CAI versions described in Equations 5 and 6 in any of the test conditions, and thus was subsequently omitted from the study. The details of the simulation study and evaluation criteria are described in the next section.

Simulation Study

Simulations were performed to compare the proposed CAI methods with existing discrete a -stratification methods under systematically varied testing conditions. In each testing condition, the CAI was compared with a -stratification selecting items within strata via (a) maximizing Fisher information (1-DIM) or D-optimality (2-DIM) and (b) match- b . To initiate the item selection process, exams were started by selecting $\hat{\theta}$ randomly within the range -3.5 to 3.5 . For the conditions using standard a -stratification, $K=4$ strata were used. Table 1 shows how many items were in the strata for both test lengths in both the 1-DIM and 2-DIM studies. For example, consider the 1-DIM exams with length 30: The first through seventh items were selected from Stratum 1, the eighth through 14th items were selected from Stratum 2, and so on. Also, in all conditions, the CAI sensitivity parameter β was set equal to 2 based upon trial and error in preliminary simulations. The simulation design will be explained in more detail below, as the 1- and 2-DIM cases are considered separately.

1-DIM Study

To investigate the performance of CAI versus standard a -stratification for CAT based on 1-DIM IRT, the test length and examinee ability generation were varied. Specifically, two levels of test length were used (20 and 30 items). Also, following the design of Leung et al. (2002), the true examinee abilities were generated in two different ways: First, a discrete uniform (DU) distribution was used, that is, sequencing from $\theta = -3.2$ to 3.2 by an increment of $.4$, resulting in 17 θ values. Each value was repeated 300 times, resulting in $N=5,100$ examinees total. For the second method, the Normal(0,1) distribution was used to generate $N=5,100$ examinee abilities. Two test lengths and two methods of generating θ s fully crossed with three stratification methods (one CAI and two a -stratification variations) resulted in $2 \times 2 \times 3 = 12$ total conditions

Figure 1. Item pool information curve for one-dimensional study.

Note. The information curve peaks approximately midway between 0 and 1.

for the 1-DIM case. After the administration of each item, $\hat{\theta}$ was computed via brute force; that is, computing the likelihood of the responses at θ values from -3.5 to 3.5 incremented by $.01$ (701 total values) and setting $\hat{\theta}$ as the value maximizing the likelihood. For all conditions, an item bank with $M=500$ items was generated in a manner similar to that of Simulation Study 1 in Wang and Chang (2011); specifically, the a , b , and c parameters were drawn from the Uniform(0, 1.3), Uniform(-1.3 , 1.3), and Uniform(0.20, 0.30) distributions, respectively. The pool information curve resulting from these parameter settings is depicted in Figure 1; the information curve peaks approximately midway between 0 and 1. The performance of each study was evaluated using bias, mean squared error (MSE), and χ^2 statistics indicating the shape of the item exposure rate distribution (Chang et al., 2001; Chang & van der Linden, 2003; Chang & Ying, 1999; Leung et al., 2002). The bias and MSE are given by

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i), \quad (7)$$

and

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2, \quad (8)$$

respectively, where $i=1, \dots, N$ is an index for the examinees. The statistic evaluating item exposure is

$$\chi^2 = \sum_{m=1}^M \frac{(er_m - J/M)^2}{J/M}, \quad (9)$$

where er_m is the observed exposure rate of item m . Larger values indicate more skewed item exposure rates, which means that some items are overexposed, while other items have been rarely administered if at all. On the contrary, lower values signify more even exposure control rates. In particular, the χ^2 statistic depends on M and J . An ideal item exposure scenario is when all items are exposed at the same rate, or in other words each item has the probability J/M of being administered. In theory, this situation results in $\chi^2 = 0$. On the contrary, the maximum value of χ^2 occurs when all examinees see the same set of J items. For example, when $J = 30$ and $M = 500$, er_m would be 1 for the $J = 30$ items administered to all the examinees and 0 for all the other 470 items. Thus, $J/M = 0.06$, and the value of χ^2 would be $470 \times \frac{(0-0.06)^2}{0.06} + 30 \times \frac{(1-0.06)^2}{0.06} = 470$.

2-DIM Study

The 2-DIM study was similar to but slightly more complex than the 1-DIM. Again, two different test lengths were used (40 and 60 items); the increase in length is necessitated by having to measure two latent traits rather than one. Examinee abilities were again generated in different ways: First, following the designs of Finkelman, Nering, and Roussos (2009) and Wang and Chang (2011), a 2-DIM grid was created by sequencing from -2 to 2 by 0.4 (11 values) for each dimension, resulting in 121 θ values. Each θ value was replicated 50 times, so that the total sample size for each condition was $N = 6,050$. Then, the second method of ability generation was drawing $N = 6,050$ from the bivariate normal distribution:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \text{bivariate normal} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$

Furthermore, under the bivariate normal conditions, ρ was considered an experimental factor with two settings: low and high ($\rho = .3$ and $.7$, respectively). Thus, abilities were generated in three different ways in the 2-DIM case. Two test lengths and three methods of generating θ s fully crossed with three stratification methods resulted in $2 \times 3 \times 3 = 18$ total conditions for the 2-DIM case. As for the 1-DIM, after administering each item $\hat{\theta}$ was computed via brute force; that is, computing the likelihood of the responses at θ vectors with each dimension ranging from -3.5 to 3.5 incremented by $.1$, that is, $\{-3.5, -3.5\}$, $\{-3.5, -3.4\}$... $\{3.5, 3.4\}$, $\{3.5, 3.5\}$ ($61^2 = 3,721$ total values) and setting $\hat{\theta}$ as the pair which maximized the likelihood. The 2-DIM item bank was generated similar to the 1-DIM case; specifically, for $M = 500$ items the a_1 and a_2 parameters were drawn from the Uniform(0, 1.3) distribution, and the b parameters were drawn from Uniform($-1.3, 1.3$). The item pool information 2-DIM plot, or surface, is displayed in Figure 2. The surface peaks at regions where both trait dimensions are positive.

The evaluation criteria are the same as for 1-DIM; however, the formulas for the bias and MSE must be extended for 2-DIM as follows:

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N \{ (\hat{\theta}_{i1} - \theta_{i1}) + (\hat{\theta}_{i2} - \theta_{i2}) \}, \quad (10)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \{ (\hat{\theta}_{i1} - \theta_{i1})^2 + (\hat{\theta}_{i2} - \theta_{i2})^2 \}. \quad (11)$$

These formulations of bias and MSE are the same as those used by Lee et al. (2008). The formula χ^2 statistic is the same as for 1-DIM.

Figure 2. Item pool information surface for two-dimensional study.

Results

1-DIM Study Results

Table 2 displays all results from the 1-DIM simulation study addressing three stratification methods: *a*-stratification with match-*b* item selection (SMB), *a*-stratification with maximum Fisher information item selection (SMI), and CAI as shown in Equation 4. Within a fixed combination of length and ability generation, some trends are readily apparent: First, the SMB method produces by far not only the largest bias and MSE but also the lowest χ^2 . This indicates that the SMB has not only the most even item exposure rate distribution but also the poorest measurement quality. Second, for a given test length and ability, CAI has similar bias and MSE but lower χ^2 , indicating that CAI outperforms SMI in terms of item exposure rates while maintaining measurement performance. Focusing on the conditions in which the abilities are $N(0,1)$ (which would most likely be the case in practice), it is found that CAI yields slightly higher bias but lower MSE when the test length is 20 items. For $N(0,1)$ ability when the test length is 30 items, CAI yields smaller bias, MSE, and χ^2 , indicating better performance than SMI on all criteria.

It may also be noted that the DU method of θ generation generally leads to larger bias and MSE than using $N(0,1)$ for a given stratification method. This is because the DU generates a much larger number of examinees at the extreme ends of the ability scale where the estimates tend to be poorer. While real examinee abilities generally do not follow the DU in practice, it is helpful to examine those conditions for additional insight into the methods. Figures 3 and 4 present the bias and MSE, respectively, plotted against θ for Conditions 4, 5, and 6, which have test length of 20 with abilities generated via DU. The patterns are generally similar for all stratification methods. The measurement quality is poorest near the low end of the θ scale, which is due to the guessing parameter of 3PL. In other words, the guessing parameter adds extra noise to the results by guaranteeing that even a very low-ability examinee will have a 20% to 30%

Table 2. Results From the I-DIM Simulation Study.

Condition	Length	Ability	Method	Bias	MSE	χ^2
1	20	$N(0,1)$	SMB	0.076	0.31	83.37
2	20	$N(0,1)$	SMI	0.032	0.22	145.03
3	20	$N(0,1)$	CAI	0.037	0.21	101.01
4	20	DU	SMB	0.085	0.44	115.17
5	20	DU	SMI	0.039	0.33	127.85
6	20	DU	CAI	0.040	0.32	103.47
7	30	$N(0,1)$	SMB	0.042	0.19	69.8
8	30	$N(0,1)$	SMI	0.021	0.14	139.57
9	30	$N(0,1)$	CAI	0.019	0.13	98.54
10	30	DU	SMB	0.057	0.28	105.65
11	30	DU	SMI	0.026	0.23	122.37
12	30	DU	CAI	0.031	0.24	99.48

Note. I-DIM = one-dimensional; MSE = mean squared error; DU = discrete uniform distribution; SMB = α -stratification with match- b item selection; SMI = α -stratification with maximum Fisher information item selection; CAI = continuous α -stratification.

Figure 3. Bias for I-DIM study plotted by θ value for Conditions 4, 5, and 6.

Note. I-DIM = one-dimensional.

chance of getting an item correct. However, there is no similar effect for the high-ability examinees, which results in the asymmetric horizontal axis seen in the figures. Conditions 10, 11, and 12 showed similar patterns to those seen in Figures 3 and 4 but with smaller bias and MSE; plots for these conditions are omitted.

Table 3 displays a breakdown of item exposure rates for the simulated test conditions in which examinee θ s were drawn from the $N(0,1)$ distribution (results for the DU conditions were omitted, because, as mentioned above, this case rarely occurs in practice). The second column in the table lists the proportion of items that were administered to less than 0.05 of the examinees, the third column lists the proportion of items that were administered to greater than 0.05 but less than 0.10 of the examinees, and so on. It is seen that CAI yields a lower proportion than SMI of items administered to at least half of the examinees. However, CAI also yields the

Figure 4. MSE for 1-DIM study plotted by θ value for Conditions 4, 5, and 6.

Note. MSE = mean squared error; 1-DIM = one-dimensional.

largest proportion of items administered to 0.20 to 0.30. While Table 2 shows that CAI improved overall exposure rates compared with SMI via the χ^2 statistic, Table 3 shows that there is still some room for improvement. Specifically, for all methods a large portion of the item bank is underutilized, and many other items are exposed too often for the item selection methods to be suitable for some operational testing programs.

2-DIM Study Results

Table 4 reports all results from the 2-DIM simulation study. As in the 1-DIM case, it can be seen immediately that SMB consistently yields a higher MSE and lower χ^2 than the other three stratification methods at given combinations of test length and ability generation method. The correlation ρ between the two latent traits in the bivariate normal conditions is a factor that did not apply to the previous 1-DIM study; here, it is seen that the MSE is generally higher for $\rho = .7$ than for $\rho = .3$ when holding length and stratification method fixed. As would be expected, the differences in MSE across the various stratification methods decrease for longer test lengths. For all testing conditions, CAI yields both smaller MSE and χ^2 than SMI, and the differences in bias between the two methods seemed negligible.

As in the 1-DIM condition, it is of interest to examine the bias and MSE by ability level for the conditions that generated the abilities according to the DU or grid method for 2-DIM. For Conditions 7, 8, and 9, Figures 5 and 6 display the bias and MSE separately for θ_1 and θ_2 ; that is, $(\hat{\theta}_{i1} - \theta_{i1})$ and $(\hat{\theta}_{i1} - \theta_{i1})^2$ for θ_1 , respectively, and likewise for θ_2 . Very similar patterns are seen for all stratification methods. Also, in contrast to Figures 3 and 4, the plots are approximately symmetric about the horizontal axis; the extreme low ends of the ability scale are measured just as precisely as the extreme high ends. This is because the 2-DIM case does not include a guessing parameter in this study.

Table 3. Breakdown of Item Exposure Rates for Simulated Testing Conditions With Abilities Drawn From the $N(0,1)$ Distribution.

Condition	Method	Item exposure rates						
		<0.05	0.05-0.10	0.10-0.20	0.20-0.30	0.30-0.40	0.40-0.50	>0.50
1	SMB	0.825	0.085	0.052	0.015	0.004	0.012	0.006
2	SMI	0.831	0.035	0.048	0.044	0.017	0.012	0.012
3	CAI	0.775	0.062	0.094	0.050	0.010	0.004	0.004
7	SMB	0.640	0.206	0.098	0.027	0.010	0.008	0.010
8	SMI	0.754	0.048	0.071	0.058	0.029	0.015	0.025
9	CAI	0.696	0.050	0.131	0.092	0.017	0.004	0.010

Note. SMB = α -stratification with match- b item selection; SMI = α -stratification with maximum Fisher information item selection; CAI = continuous α -stratification.

Table 4. Results From the 2-DIM Simulation Study.

Condition	Length	Ability	Method	Bias	MSE	χ^2
1	40	BN ($\rho = .3$)	SMB	−0.003	0.37	12.87
2	40	BN ($\rho = .3$)	SMI	0.001	0.28	86.55
3	40	BN ($\rho = .3$)	CAI	−0.003	0.26	57.17
4	40	BN ($\rho = .7$)	SMB	−0.007	0.40	14.15
5	40	BN ($\rho = .7$)	SMI	0.000	0.30	81.63
6	40	BN ($\rho = .7$)	CAI	−0.005	0.27	58.61
7	40	Grid	SMB	0.002	0.39	16.77
8	40	Grid	SMI	−0.005	0.31	75.96
9	40	Grid	CAI	−0.008	0.28	51.55
10	60	BN ($\rho = .3$)	SMB	−0.001	0.24	8.88
11	60	BN ($\rho = .3$)	SMI	−0.002	0.19	84.83
12	60	BN ($\rho = .3$)	CAI	0.001	0.17	54.40
13	60	BN ($\rho = .7$)	SMB	−0.003	0.26	9.50
14	60	BN ($\rho = .7$)	SMI	−0.004	0.21	78.47
15	60	BN ($\rho = .7$)	CAI	−0.008	0.19	55.91
16	60	Grid	SMB	−0.004	0.26	13.08
17	60	Grid	SMI	−0.003	0.21	71.60
18	60	Grid	CAI	0.003	0.18	46.56

Note. MSE = mean squared error; BN = bivariate normal ability generating distribution with correlation ρ ; SMB = α -stratification with match- b item selection; SMI = α -stratification with maximum D-optimality item selection; CAI = continuous α -stratification.

Table 5 breaks down the item exposure rates for testing conditions in which abilities were generated according to the bivariate normal ability generating distribution with correlation ρ (BN), revealing patterns similar to Table 3 for the 1-DIM case. Specifically, for the SMI and CAI conditions, the majority of the items in the bank were exposed to less than 0.10 of the examinees, and a subset of items were exposed to a large proportion of examinees.

Discussion

This article has introduced the CAI, a new method for performing α -stratified CAT. Previous α -stratification schemes use a fixed number of strata; in contrast, CAI, as its name suggests, is

Figure 5. Bias for 2-DIM study by θ_1 and θ_2 values for Conditions 7, 8, and 9.
Note. 2-DIM = two-dimensional.

Figure 6. MSE for 2-DIM study by θ_1 and θ_2 values for Conditions 7, 8, and 9.
Note. MSE = mean squared error; 2-DIM = two-dimensional.

a continuous method without fixed, discrete strata. Thus, it is suitable for operational situations in which items are frequently added to or retired from the item bank. In other words, compared with the fixed-strata SMB or SMI methods, in which the optimal number of strata needs to be decided via extensive simulation studies every time the item bank is replenished, CAI is much

Table 5. Breakdown of Item Exposure Rates for Simulated Testing Conditions With Abilities Drawn From the BN Distribution.

Condition	Method	Item exposure rates						
		<0.05	0.05-0.1	0.10-0.20	0.20-0.30	0.30-0.40	0.40-0.50	>0.50
1	SMB	0.173	0.550	0.271	0.002	0.000	0.000	0.004
2	SMI	0.640	0.069	0.106	0.102	0.060	0.019	0.004
3	CAI	0.531	0.129	0.185	0.112	0.035	0.006	0.000
4	SMB	0.196	0.548	0.250	0.002	0.000	0.000	0.004
5	SMI	0.619	0.088	0.112	0.102	0.056	0.021	0.002
6	CAI	0.540	0.115	0.194	0.106	0.038	0.008	0.000
10	SMB	0.010	0.262	0.694	0.029	0.000	0.000	0.004
11	SMI	0.494	0.096	0.133	0.133	0.071	0.058	0.015
12	CAI	0.373	0.135	0.235	0.154	0.075	0.021	0.006
13	SMB	0.015	0.296	0.654	0.031	0.000	0.000	0.004
14	SMI	0.477	0.090	0.152	0.140	0.077	0.050	0.015
15	CAI	0.381	0.131	0.221	0.165	0.077	0.021	0.004

Note. BN = bivariate normal ability generating distribution with correlation ρ ; SMB = α -stratification with match- b item selection; SMI = α -stratification with maximum D-optimality item selection; CAI = continuous α -stratification.

easier to implement. Formulas for CAI were proposed for both 1-DIM and 2-DIM CATs. A simulation study compared the performance of the CAI methods with that of existing α -stratification methods. The SMB method was shown to be the best by far in terms of item exposure control as indicated by the consistently low values of χ^2 , a measure of the evenness of item administration rates. However, the SMB also yielded consistently high MSE. If practitioners find the SMB too imprecise a measurement tool, then SMI and CAI may be viable options. The present study results suggest that the performance of CAI was similar to or better than that of SMI in terms of bias and MSE while producing smaller χ^2 values. In particular, for the 1-DIM case with abilities generated from the $N(0,1)$ distribution and a test length of 30 items, CAI outperformed SMI on all three evaluation criteria. Also, CAI had lower MSE and χ^2 than SMI in the 2-DIM study for all combinations of test length and ability distribution with comparable bias. Thus, depending on the situation, CAI may provide an optimal middle ground between SMB and SMI.

The manner in which the test is started may affect the performance of each method. For example, in preliminary simulations, each test was started by administering to each examinee a set of three items chosen randomly from the bank. Under this scheme, the CAI did not outperform SMI on χ^2 or MSE criteria. The simulation design in the current article started each exam by selecting a value of $\hat{\theta}$ randomly before the administration of the first item. Under this plan, as discussed above, the CAI was shown to outperform SMI in a number of different testing conditions. This emphasizes that while CAI has been shown to be a comparable or better performing alternative to SMI, each testing situation is different, and there is no “one-size-fits-all” answer.

There were several limitations to the current study, as no single study can realistically examine all possible testing scenarios: First, it should be noted that only one value was examined for β , the sensitivity parameter present in all versions of the CAI. In addition, only one item bank was simulated for each study; results may differ when the item bank characteristics are varied. In particular, one may wish to vary the size of the bank and/or generating distributions for the parameters. Practitioners wishing to implement CAI should experiment with different values of β for their particular operational item banks. Also, for the existing α -stratification methods to

which the CAI was compared, the number of strata was fixed at $K=4$ for all conditions. The trends described in the previously discussed tables and figures may change when β and/or K are varied. It may be worthwhile to study the case in which the strata are unbalanced, that is, when the strata sizes are not equal and/or different numbers of items are taken from each stratum.

An advantage of the CAI method that may not be readily apparent is its ability to be extended to more than two dimensions. The extension of the a -stratification method to two dimensions proposed by Lee et al. (2008) does not readily apply to three or more dimensions. However, the fact that the CAI is computed in terms of the vector of discrimination parameters ($\mathbf{a}'\mathbf{a}$) means that this extension would be straightforward. An application to a three-dimensional case could be quite interesting.

Future studies may seek to enhance the proposed CAI methods. For example, CAI could be combined with a method of maximum exposure control to limit the maximum item exposure rate. An on-the-fly method of maximum exposure control could pair nicely with the flexibility of the CAI. Also, it could be advantageous to relax the fact that CAI always forces the discrimination parameters of the selected items to be strictly ascending. Specifically, finding a systematic way to relax this criterion would improve item exposure rates while negligibly sacrificing measurement quality.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.
- Bussemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Thousand Oaks, CA: Sage.
- Camilli, G. (1994). Origin of the scaling constant $d=1.7$ in item response theory. *Journal of Educational and Behavioral Statistics*, 19, 293-295.
- Chang, H., Qian, J., & Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a-stratified computerized adaptive testing. *Applied Psychological Measurement*, 27, 262-274.
- Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Finkelman, M., Nering, L. M., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, 46, 84-103.
- Georgiadou, E., Triantafyllou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed for 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5(8). Available from <https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1647>
- Huebner, A., Wang, C., Quinlan, K., & Seubert, L. (2015). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods*, 48, 1443-1453. doi:10.3758/s13428-015-0659-z
- Lee, Y. H., Ip, E. H., & Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, 68, 215-232.

- Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympon-Hetter algorithm. *Applied Psychological Measurement, 26*, 376-392.
- Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.
- Wang, C., & Chang, H. (2011). Item selection in multidimensional computerized adaptive tests—Gaining information from different angles. *Psychometrika, 76*, 363-384.