Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods

Chi-Keung LEUNG
The Hong Kong Institute of Education

Hua-Hua CHANG
University of Texas at Austins

Kit-Tai HAU, Zhonglin WEN
The Chinese University of Hong Kong

Paper for presentation at the NCME Annual Meeting, April 2003. Please send all correspondence to Chi-Keung Leung at Department of Mathematics, The Hong Kong Institute of Education, Tai Po, Hong Kong (Email: ckleung@ied.edu.hk).

Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods

Chi-Keung LEUNG
The Hong Kong Institute of Education

Hua-Hua CHANG
University of Texas at Austin

Kit-Tai HAU, Zhonglin WEN
The Chinese University of Hong Kong

Abstract

Content balancing is often a practical consideration in the design of computerized adaptive testing (CAT). This study compared three content balancing methods, namely, the constrained CAT (CCAT), the modified constrained CAT (MCCAT), and the modified multinomial model (MMM), under various conditions of test length and target maximum exposure rate. Results indicate that there is no systematic effect in measurement efficiency and pool utilization due to content balancing methods. Nevertheless, the MMM appears to consistently over-expose less number of items.

Introduction

In the last two decades, the advancement in computer technology and psychometric theories has accelerated the change of test format from conventional paper-and-pencil (P&P) tests to computerized adaptive testing (CAT) which was first developed under the item response theory models (Lord, 1970). In CAT, an examinee is presented with tailor-made tests in which one item is adaptively selected at a time on the basis of the currently available estimate of the examinee's ability

(Lord, 1980; Weiss, 1982). One of the main advantages of CAT over P&P is that the former enables more efficient and precise trait estimation (Owen, 1975; Wainer, 1990; Weiss, 1982).

To attain high efficiency in CAT, many item selection algorithms adopt the information approach in which an item is selected if it has the maximum Fisher information at the current ability estimate based on the responses to previously administered items. It has been noted that this information criterion would cause unbalanced item exposure distribution (Davey & Parshall, 1995; McBride & Martin, 1983; Sympson & Hetter, 1985; van der Linden, 1998). In particular, highly discriminating items may be overly exposed while some less discriminating items may never be used. This undesirable outcome may eventually damage test security and increase the cost in developing and maintaining item pools.

To remedy the shortcoming of high exposure in maximum information item selection, Sympson and Hetter (SH, 1985) proposed a probabilistic method to directly control exposure rate of active items that are frequently selected. Nevertheless, the SH method cannot directly uplift the usage of those items that are rarely selected.

On a different line of thought, Chang and Ying (1999) have proposed the multi-stage a-stratified design (ASTR) that partitions items into several strata in an ascending order of the item discrimination parameter. Each test then consists of matching numbers of stages and strata, with items of the first stage being selected from the first stratum that mainly contains less discriminating items, and so on. One major rationale for such a design is that in early stages, the gain in information by using the most informative item may not be realized because the ability estimation is still relatively inaccurate. Thus, items with high discrimination values should be saved for later stages. The ASTR has been shown through simulation studies to be effective in both reducing item-overlap rate and the enhancing pool utilization.

Besides estimation efficiency and item exposure control, content balancing, in having the targeted distribution of items from different content domains, is another common practical issue that a CAT design has to take into consideration. Kingsbury and Zara (1989) have proposed the popular constrained CAT (CCAT) method.

Basically, this content-balancing algorithm selects the most optimal item from the content area with the current exposure rate farthest below its target administration percentage. Chen and Ankenmann (1999) have argued that the CCAT may yield undesirable order effects as the sequence of content areas resulted is highly predictable. Instead, they have developed a modified multinomial model (MMM) to meet the balanced content requirement. Subsequently, Leung, Chang and Hau (2000) have proposed a modified CCAT (MCCAT) that can eliminate the predictability of the sequence of content areas of CCAT and satisfy the practical constraint of content balancing as well.

Previous research on content balancing in stratification designs indicate that the MMM, MCCAT, and CCAT have similar effect on measurement efficiency but the CCAT is consistently less effective than the other two methods in terms of pool utilization and control of item overlap rate (Leung, Chang, & Hau, in press). As these findings were observed in stratified CAT designs, they may not be necessarily applicable to maximum information item selections. This study aimed to compare the three content balancing methods, under the information-based selection criterion, at different conditions of test length and target exposure rate.

Content Balancing Methods

(1) The Constrained CAT (CCAT): The selection of an optimal item is restricted to the content area with an exposure rate farthest below its target percentage for the test.

(2) The Modified Multinomial Model (MMM): A cumulative distribution is first formed based on the target percentages of the content areas that sum to 1.0. Then, a random number from the uniform distribution U(0,1) is used to determine the corresponding content area in the cumulative distribution where the next optimal item will be selected. When a content area has reached its target percentage, a new multinomial distribution is formed by adjusting the unfulfilled percentages of the remaining content areas. As random mechanism is incorporated in this method, the sequence of content areas varies.

(3) The Modified Constrained CAT (MCCAT): Instead of being restricted to the content area that has current exposure rate farthest below its target percentage, an optimal item can be chosen from all the content areas that still have quota not fully used up. As a result, the undesirable order effect of CCAT is eliminated.

Exposure Control

The foundation of the SH control algorithm rests on the concept of conditional probability: $P(A) = P(A|S)*P(S)$, where $P(S)$ is the probability that an item is selected as the best next item for a randomly sampled examinee from a typical population, and $P(A|S)$ is the probability that the item is administered when selected. The procedure attempts to control $P(A)$, the overall probability that an item is administered, by assigning an exposure control parameter $P(A|S)$ to the item. The exposure control parameters for all items are determined through a series of prior adjustment simulations so that the probability of administration for each item is restricted to about the pre-specified maximum exposure rate (Sympson & Hetter, 1985).

Simulation Design

Item pool: A pool of 700 calibrated mathematics items from four major content areas was used.

Test length: Three test lengths of respectively 16, 28, and 40 items were studied.

Content specifications: For each test length, a fixed proportion of items from the four content areas were applied to all corresponding adaptive tests.

Exposure rate: Two target maximum exposure rates of respectively 0.1 and 0.2 were studied.

Ability traits: A sample of 5000 simulees with abilities randomly generated from N(0,1) was used. Each simulee received an adaptive test from each of the 18 combinations (3 methods x 3 test lengths x 2 exposure rates) of conditions.

Evaluation Criteria

The performances of the content balancing methods were evaluated in terms of

(i) correlation of the true and the estimated theta, (ii) average bias, (iii) mean squared error, (iv) scaled chi-square statistic (Chang & Ying, 1999), (v) number of over-exposed items, and (vi) number of under-utilized items.

Results

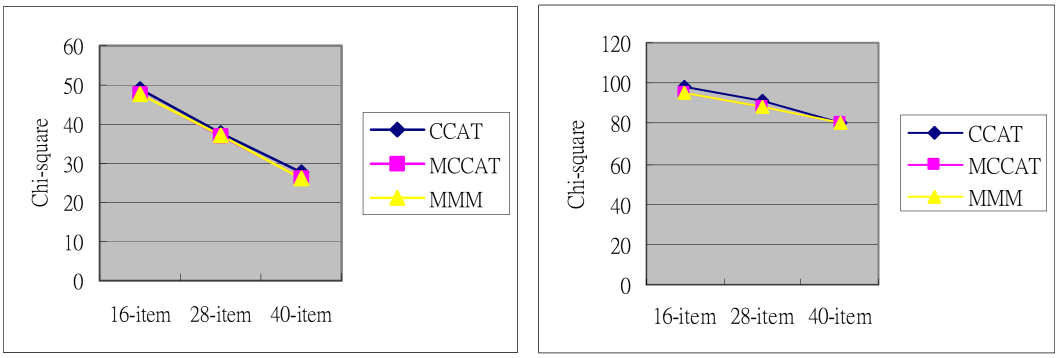
The results of the study are summarized in Table 1. The three content balancing methods appeared virtually unbiased as their estimated bias are all close to zero.

Table 1: Summary Statistics for Three Content Balancing Methods

	CCAT	MCCAT	MMM
16-item test	*r* = .1 (*r* = .2)	*r* = .1 (*r* = .2)	*r* = .1 (*r* = .2)
Correlation	.954 (.961)	.955 (.960)	.954 (.960)
Bias	-.006 (.009)	.007 (.005)	.005 (.008)
MSE	.102 (.089)	.101 (.090)	.101 (.089)
Scaled χ^2	48.9 (98.3)	47.7 (95.5)	47.6 (95.2)
N(exp<.02)	521 (583)	520 (578)	520 (576)
N(exp>*r*)	60 (29)	55 (23)	52 (21)
28-item test			
Correlation	.971 (.975)	.970 (.975)	.970 (.973)
Bias	-.000 (.003)	-.001 (.001)	-.001 (.004)
MSE	.064 (.053)	.064 (.053)	.065 (.054)
Scaled χ^2	37.7 (90.8)	37.0 (88.3)	37.1 (88.7)
N(exp<.02)	393 (501)	386 (498)	380 (499)
N(exp> *r*)	119 (44)	109 (35)	108 (37)
40-item test			
Correlation	.976 (.981)	.976 (.981)	.975 (.980)
Bias	.003 (.002)	-.004 (.000)	.005 (-.002)
MSE	.054 (.040)	.054 (.040)	.054 (.040)
Scaled χ^2	27.7 (80.7)	26.2 (80.2)	26.1 (80.5)
N(exp<.02)	271 (432)	260 (430)	258 (432)
N(exp> *r*)	178 (65)	171 (64)	157 (55)

They offered highly correlated estimates for the corresponding abilities. Their estimated correlation coefficients are comparable under various conditions. In terms of MSE, the three methods also perform similarly. Overall, there is a general trend that when test length increases, the correlation goes up and the MSE goes down. This trend also happens when the target exposure rate increases.

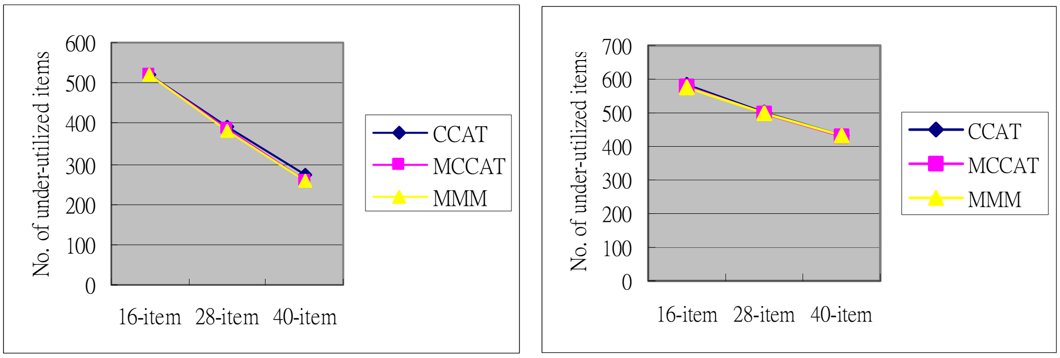
Regarding pool utilization, the CCAT yielded slightly higher values in scaled χ^2 and larger numbers of under-utilized items than the MCCAT and the MMM. Nevertheless, the difference appears not substantial as reflected in Figures 1 and 2. As evidenced by larger χ^2 values and larger numbers of under-utilized items, the item exposure distribution becomes more skewed when the target maximum exposure rate increases from .1 to .2. On the contrary, when the test length increases, the item exposure distribution becomes more even.



(a) $r = .1$

(b) $r = .2$

Figure 1: Chi-square statistics across content balancing method and test length

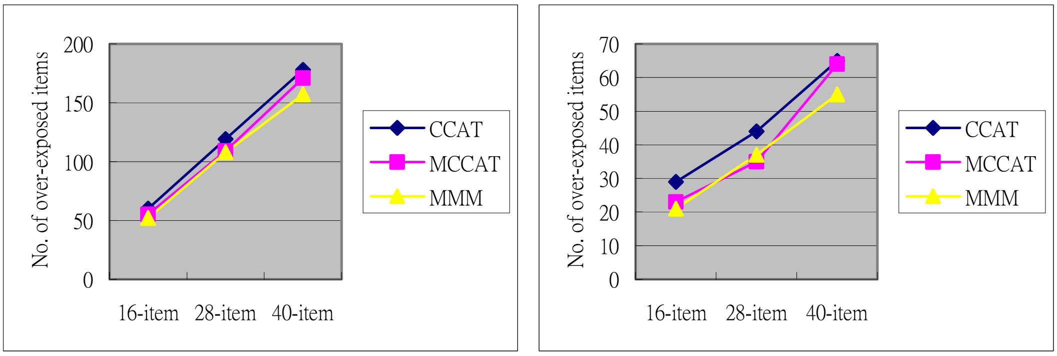


(a) $r = .1$

(b) $r = .2$

Figure 2: No. of under-utilized items across content balancing method and test length

As regard to item security, the MMM appears to be better as it consistently yielded less numbers of over-exposed items. Figure 3 shows that the CCAT tended to over-expose more items for 16- and 28-item tests under target maximum rate of .2 and it performed similarly as the MCCAT at 40-item test.



(a) $r = .1$ (b) $r = .2$

Figure 3: Number of over-exposed items across content balancing method and test length

<u>Summary</u>

As content balancing is a common requirement of many large-scale educational tests, a comprehensive investigation of the two new content balancing methods (MCCAT and MMM) and the conventional one (CCAT) under different conditions of test length and target exposure control provides valuable information for the educational researchers and practitioners in the fields of CAT and measurement.

Results indicate that the three content balancing methods, when working with the maximum information selection approach, offer comparable estimation accuracy and precision in terms of MSE, bias, and correlation coefficient. It is found that the test length and target maximum exposure rate are two significant factors affecting measurement performance: The accuracy and precision increases with the test length and target maximum exposure rate.

The three content balancing methods differ in the number of over-exposed items and the order of item content presented. The MMM tends to over-expose less number of items and thus appears to be better in item security control. Besides, it is found that the order of contents of the presented items in CCAT is highly predictable:

The first few items in each test come from the content with largest pre-specified percentage and the contents of the subsequent test items appear in cycle.

The current findings suggest that among the three methods, the MMM is more desirable for content balancing. It can reduce the predictability of item content and over-expose less number of items. As the present study involves only one item pool, the advantages of the MMM over the other two methods need to be cross-examined with more item pools and testing conditions.

References

Chang, H.H., & Ying, Z. (1999). A-stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 20,* 213-229.

Chen, S., Ankenmann, R.D., & Spray, J.A. (1999, April). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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, San Francisco, USA.

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

Leung, C.K., Chang, H.H., & Hau, K.T. (2000, April). *Content Balancing In Stratified Computerized Adaptive Testing Designs*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, USA.

Leung, C.K., Chang, H.H., & Hau, K.T. (in press). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*.

Lord, M.F. (1970). Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer Assisted Instruction, Testing, and Guidance*. New York: Harper and Row.

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

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*. New York: Academic Press.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

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

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.