An Examination of Item Selection Rules by Stratified CAT Designs Integrated with Content Balancing Methods

Chi-Keung LEUNG
The Hong Kong Institute of Education

Hua-Hua CHANG
National Board of Medical Examiners

Kit-Tai HAU
The Chinese University of Hong Kong

Note. Paper presented at the AERA Annual Meeting, Seattle, April 10-14, Seattle. Address all correspondence to Chi-Keung Leung, Dept. of Math., Hong Kong Institute of Education, Taipo, N.T., Hong Kong. Email: ckleung@ied.edu.hk.

Abstract

The multistage *a*-stratified computerized adaptive testing design advocates a new philosophy on pool management and item selection of using low discriminating items first. It has been demonstrated through simulation studies to be effective both in reducing item overlap rate and enhancing pool utilization with certain pool types. Based on this design, two extended stratification methods have been proposed to deal with practical issues of content constraints and correlation between difficulty and discrimination parameters respectively. These stratification designs on their own do not automatically meet content requirements. Instead, they need to be modified or have to work together with some content balancing methods in order to satisfy all content constraints. This study aimed to investigate whether there is any effect due to the factors of content balancing method and stratification design. Specifically, the three stratification designs were examined together with three well developed content balancing methods, under the practical constraint of strict content specifications. The performance of each of the nine combinations was evaluated in light of item security, measurement efficiency, and pool utilization. Results indicate that there is no interaction effect between the two factors, but main effect does exist in content balancing method on item security and pool utilization, as well as stratification design on item usage.

An Examination of Item Selection Rules by Stratified CAT Designs Integrated with Content Balancing Methods

With the increasing availability of powerful microcomputers and the advances in psychometrics, many large-scale testing programs have been partly or completely converted into computerized adaptive testing (CAT) format, which was first developed under the item response theory framework (IRT; Lord, 1970). In CAT, examinees are presented with tailor-made tests. One item is 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 it enables more efficient and precise trait estimation (Owen, 1975; Wainer, 1990). A key issue in CAT is how to adaptively select the best test items from the item pool. The traditional item selection algorithms rely on local item information. This means that 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 skewed item exposure (Davey & Parshall, 1995; Sympson & Hetter, 1985; van der Linden, 1998). In particular, items with large value of discrimination parameter may be overly exposed while some others are never used. Over-exposure would eventually threaten in item security whilst under-utilization would reduce the cost effectiveness of developing and maintaining those inactive items in the pool. It is understandable, therefore, the control of item exposure and the enhancement of pool efficiency are important issues in CAT designs (Mills & Stocking, 1996; Stocking & Swanson, 1998; Way, 1998).

Methods that attempt to *simultaneously* control maximum item exposure rate to improve item security and uplift exposure of under-utilized items to enhance item pool efficiency have been proposed by Chang and Ying (1999), and Stocking and Swanson (1998), among others. In contrast to the traditional approach of looking for the most informative items at every stage of item-selection, Chang and Ying (1999) have proposed a multi-stage *a*-stratified design (ASTR) that partitions items into several strata in the ascending order of the item discrimination parameter. Each test then consists of matching number of stages and strata, with items of the first stage being selected from the first stratum and so on. One major rationale for such a design is that at early stages, the gain in information by using the most informative items may not be realized because the ability estimation is still relatively inaccurate. Thus items with high discrimination values should be used at later stages. The stratified design has been shown through simulation studies to be effective in the reduction of test-overlap rate and the enhancement of pool utilization when it is used with certain types of item pools. The findings of Hau and Chang (in press) also support that ASTR offers comparable efficiency as MI but has the potential advantages of achieving a more balanced item usage and stable resultant pool structure in testing with continuous item replenishments.

For operational item pools, it is common to find substantial correlation between the *a*- and *b*- parameters of the items. Thus, at the later stages of testing with the high *a* strata, there may not be sufficient low *b* items. Consequently, the ASTR would result in an overexposure of these low *b* items. To tackle such problem, Chang, Qian and Ying (in press) modified the ASTR strategy and developed the *a*-stratified with *b*-blocking method (BASTR). In BASTR, item pool is first divided into many small levels based on *b* parameters. Within each level, items are further grouped in ascending order of *a* values. The first groups of items with the smallest *a* values from corresponding levels are then merged to form the first stratum, and the second groups of items with the second smallest *a* values merged into the second stratum and so on, until eventually the last stratum is formed from items with the largest *a* values from each level. As a result, the *b* distribution remains steady but the average *a* value increases across strata.

Originally, both ASTR and BASTR were developed without taking into consideration of content requirements. To tackle the situations where the *b* distributions may be very different across content areas, Yi and Chang (2000) have extended BASTR to multiple stratification (denoted by CBASTR hereafter) in which items are divided into strata by three factors, namely the content, *b* parameter, and then *a* parameter.

Very often, however, CAT design has to take into consideration practical constraints such as content balancing. To ensure that each adaptive test has the same mix of contents, some mechanisms are needed to make sure that strict content balancing is incorporated. A method called constrained CAT (CCAT) was proposed by Kingsbury and Zara (1989). Basically, this content-balancing algorithm selects the most optimal item from the content area that is farthest below its ideal administration percentage for each examinee. Chen and Ankenmann (1999) have argued that CCAT may yield undesirable order effects as the sequence of content areas is highly predictable. Instead, they have used a modified multinomial model (MMM) in their research to meet the requirement of strict content balancing. On the other hand, Leung, Chang and Hau (2000) have modified the CCAT to simultaneously satisfy the practical constraint of content balancing and eliminate the predictability of the sequence of content areas.

In this study, the nine integrated item selection rules arising from the three stratification designs and the three content balancing methods described earlier were examined and compared. The objective was to investigate whether there is (i) any difference in the main effect by individual stratification designs and content balancing methods, and (ii) any interaction effect across the stratification designs and the content balancing methods, on measurement accuracy, item security and pool utilization. Every adaptive test administered by each of the nine item selection rules satisfied all the strict content specifications. These selection methods were evaluated in light of (i) reliability, (ii) mean squared error, (iii) item overlap rate, (iv) chi-squared statistic, (v) number of over-exposed items, and (iv) number of under-utilized items.

Method

Item Selection Rules

Nine item selection rules were compared in the study. They were the combinations of the following two factors.

Stratification Design

(i) The *a*-Stratified Design (ASTR): The item pool is partitioned into 4 strata of equal size in an ascending order of the *a*-parameters and each adaptive test is divided into 4 stages accordingly. In each stage, a pre-specified number of items are selected from the corresponding stratum. For each item selection, an optimal unadministered item with difficulty closest to the current ability estimate is chosen for consideration of being administered.

(ii) The *a*-Stratified Design with *b*-Blocking (BASTR): Basically, BASTR is similar to ASTR except the pool stratification method. In BASTR, the correlation of *a*- and *b*-parameters is taken into consideration. Items are first divided into 20 different levels based on the *b*-parameters. Within each level, items are divided into 4 groups in ascending order of the *a*-parameters. Then, the first groups of the corresponding levels are combined to form the first stratum, …, and the last groups combined to form the last stratum. As a result, each stratum has similar *b*-distribution, but the average value of *a*-parameters increases across strata.

(iii) The Multiple Stratification (CBASTR): The CBASTR differs from BASTR only in that the item pool is first divided into groups based on the content areas. Then, in each group, the items are assigned to different strata based on their *b*- and *a*-parameters in a way similar to BASTR. As a result, each stratum has similar content coverage, but the average values of *a*-parameters and *b*-parameters increase across strata.

Content Balancing Method

(i) The Constrained CAT (CCAT): The selection of an optimal item is restricted from the content area that is farthest below its pre-specified percentage in the test.

(ii) 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 and follow a multinomial distribution. 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. However, the target percentage for each content area may not be met exactly due to sampling errors. Thus, whenever a target percentage is reached, a new multinomial distribution needs to be formed by adjusting the rest percentages of the remaining content areas.

(iii) The Modified Constrained CAT (MCCAT): Instead of being restricted to the content area that is farthest below to its pre-specified percentage, an optimal item can be chosen from all the content areas that still have quota not fully used-up. As a result, the undesired order effect of CCAT will be eliminated.

Simulation Design

Item pool: A pool of 700 calibrated mathematics items from four major content areas each having 234, 166, 150, and 150 items respectively were used. Table 1 shows the parameter distributions of the items across the contents. There were significant differences in both *a*- and *b*-parameters across the contents. The correlation of these two parameters was .45. In average, items of Content C and Content D were more difficult whilst those of Content A were relatively easier and less discriminating.

Table 1: Parameter Distributions across Contents

Content		*Discrimination (a)*	*Difficulty (b)*	*Guessing (c)*
A	Mean	.88	-.24	.18
	S.D.	.25	1.15	.008
B	Mean	1.03	.008	.16
	S.D.	.32	.99	.008
C	Mean	1.15	.48	.18
	S.D.	.34	.89	.008
D	Mean	1.02	.57	.17
	S.D.	.33	1.06	.008

Table 2 shows the item parameter distributions across strata for the three stratification designs. It appears that the *b* distribution was relatively stable across strata for the BASTR but its mean increased in both ASTR and CBASTR. On the other hand, the mean of *a* parameters increased across strata for all stratification methods whilst the *c* distributions were similar for all strata and insensitive to stratification methods.

Test length: The test length was fixed at 35 items: 8 items from the 1st stratum and 9 from each of the next three strata.

Content specifications: The numbers of items from the four content areas for each test were 12, 8, 8, and 7 respectively.

Ability traits: 5000 true theta were generated from N(0,1). Each simulee received nine adaptive tests respectively administered by the nine integrated selection methods.

Table 2: Parameter Distributions across Strata for 3 Stratification Designs

			a-			*b-*			*c-*	
Stratum		*AST*	*BAST*	*CBA*	*AST*	*BAST*	*CBA*	*AST*	*BAST*	*CBA*
1	Mean	.68	.70	.70	.01	.14	.008	.18	.18	.17
	S.D.	.15	.17	.17	1.15	1.11	1.12	.008	.008	.007
2	Mean	.91	.92	.92	.12	.14	.16	.17	.17	.019
	S.D.	.15	.17	.19	1.06	1.04	1.07	.007	.008	.008
3	Mean	1.10	1.10	1.09	.18	.15	.19	.18	.18	.17
	S.D.	.17	.20	.20	1.04	1.07	1.05	.009	.008	.008
4	Mean	1.40	1.38	1.37	.25	.15	.21	.17	.17	.17
	S.D.	.23	.31	.31	1.00	1.04	1.02	.008	.008	.008

Criteria for Evaluation

Reliability

Irrespective of the item selection design, CAT should always provide reasonable reliability. Otherwise, test results cannot be used for inference or decision. Therefore, reliability was an evaluation criterion for the performance of the three selection methods. In this study, 5,000 true abilities were generated and their respective estimates were obtained according to the selection methods employed. Thus reliability here was defined as the correlation between the estimated scores and the true scores (Lord, 1980, p. 52). The higher the reliability, the better the item selection method would be.

Mean Squared Error

Measurement efficiency is another important criterion, which in this study was measured by the mean squared error. Let θ_i, $i = 1,\ldots, 5000$ be the true abilities of the 5000 examinees and $\hat{\theta}_i$ be the respective estimators from the CAT. The estimated mean squared error was computed as

$$\text{MSE} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i)^2 \qquad (1)$$

The smaller the MSE, the more efficient is the item selection method.

Number of Over-exposed Items

The exposure rate of an item is defined as the ratio of the number of times the item is administered to examinees over the total number of examinees taking the test. If an item has a high exposure rate, then it has a greater risk of being known to prospective examinees, which in turn would cause test security and validity problems. Since one of the main concerns of this paper was to compare the three item selection methods in item exposure control, the number of overly exposed items was certainly a key evaluation criterion. Here an item was considered overly exposed if its exposure rate was greater than .2, a commonly used cut-off value. The smaller the number of over-exposed items, the better the item selection method would be.

Number of Under-utilized Items

Items with very low exposure rate are rarely used. If there are too many items with low exposure rates, then the item pool is not well utilized, which challenges directly the cost effectiveness of the item pool and the appropriateness of the item selection method. In this study, an item was considered as under-utilized if its exposure rate was below .02. The smaller the number of under-utilized items, the better the item selection method would be.

Scaled Chi-squared Statistic

Chang and Ying (1999) proposed that a uniform exposure rate distribution should be the most desirable in order to have a maximum item pool utilization. If the pool size is *N* and test length is *L*, then the optimum uniform exposure rate is *L*/*N*. They introduced a scaled chi-square to measure the overall item pool usage efficiency:

$$\chi^2 = \sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N} \qquad (2)$$

where er_j represents the observed exposure rate for the *j*th item.

Equation 2 reflects the discrepancy between the observed and the ideal exposure rates. The smaller the χ^2, the better the pool utilization and hence the item selection method would be.

Item Overlap Rate

The item overlap rate (sometimes called the test overlap rate) is another important summary index in measuring item exposure control (Mills & Stocking, 1996; Way, 1998). Item overlap rate is indicated by the proportion of items shared by pairs of examinees, averaged across all possible pair-wise combinations. Way (1998) has argued that such an index, not being calculated on an individual item basis, provides a global picture of how often sets of items are administered. The higher the item overlap rate, the greater the damage to test validity due to information sharing among examinees who take the test at different occasions. He also stressed that this index is critical in determining the size and composition of item pools that are needed for a particular CAT. If the test length is *L* and there are *P* examinees, the item overlap rate here was computed by: (i) first counting the number of common items for each of the *P*(*P*-1)/2 pairs of examinees, (ii) adding up all the counts in the *P*(*P*-1)/2 pairs, and (iii) dividing the total count by *LP*(*P*-1)/2. The smaller the overlap rate, the better the item selection method would be. Chang and Zhang (1999) and Chen, Ankenmann and Spray (1999) separately found that the lower bound for the expected item-overlap rate is *L*/*N*, the ratio of the test length to the pool size. This lower bound serves as a baseline for comparison among item selection methods in controlling item overlap.

Results

The results of the simulation study are summarized in Table 3. All the selection methods yielded similar reliability and MSE, meaning that they offered comparable measurement accuracy and efficiency. Figure 1 and Figure 2 also reflect that there is neither main effect nor interaction due to stratification design or content balancing method on measurement performance.

Table 3: Summary Statistics for Nine Item Selection Rules

	Reliability	MSE	χ^2	N(< .02)	N(> .20)	Overlap
CCAT						
ASTR	.965	.067	37.0	208	11	.096
BASTR	.965	.069	40.0	192	14	.100
CBASTR	.965	.068	34.5	174	12	.093
MCCAT						
ASTR	.972	.069	31.3	179	13	.078
BASTR	.965	.067	29.7	155	12	.074
CBASTR	.964	.070	26.6	177	9	.075
MMM						
ASTR	.965	.067	23.9	153	7	.078
BASTR	.965	.068	23.3	137	10	.077
CBASTR	.965	.068	17.2	110	6	.070

Figure 1: MSE across Stratification Designs and Content Balancing Methods

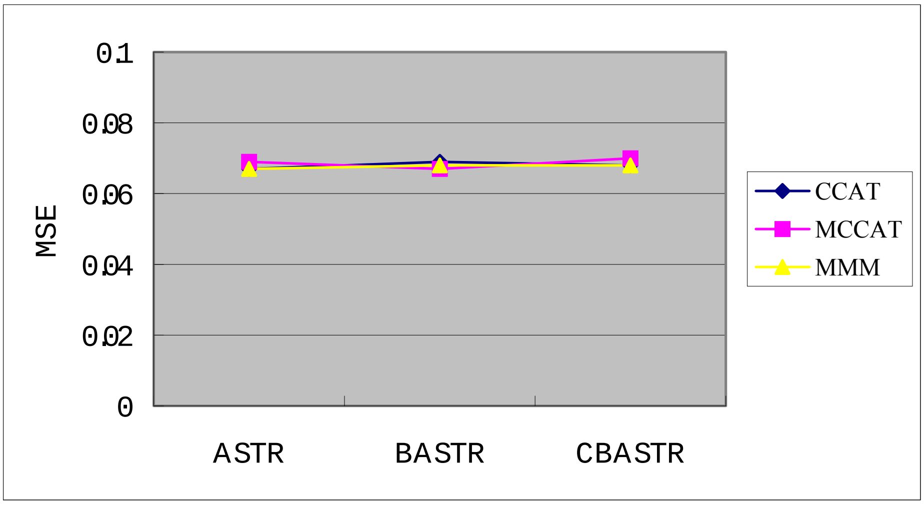
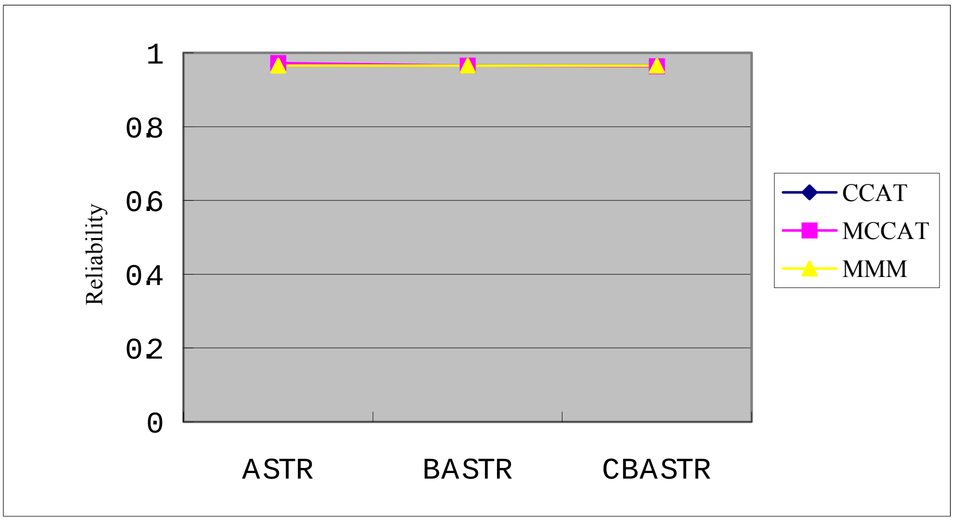


Figure 2: Reliability across Stratification Designs and Content Balancing Methods



Concerning pool utilization, there were significant main effects due to stratification and content balancing method respectively. Figure 3 indicates that MMM made a better use of all items while CCAT was less adequate in this aspect. Among the stratification designs, CBASTR performed the best in pool utilization. These main effects are also evidenced from the much smaller numbers in both the under-utilized and over-exposed items, by MMM and CBASTR within their respective categories.

Figure 3: Chi-squares across Stratification Designs and Content Balancing Methods

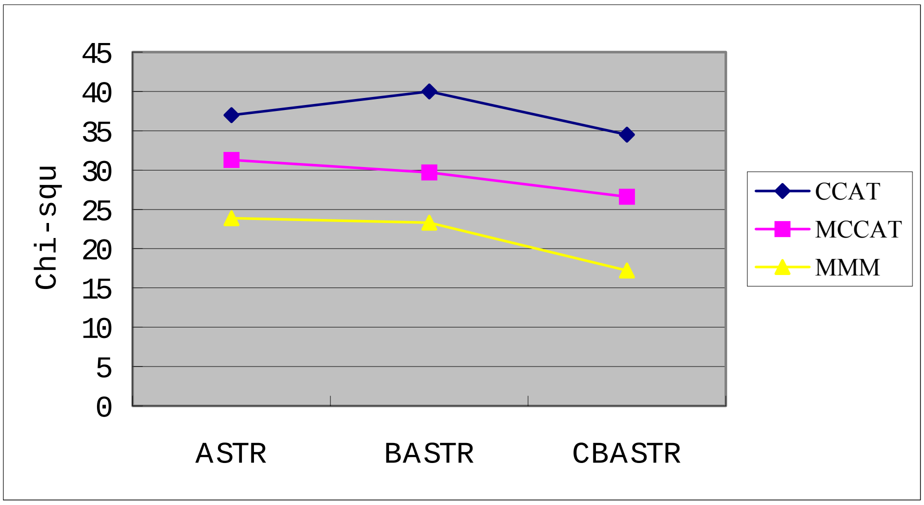
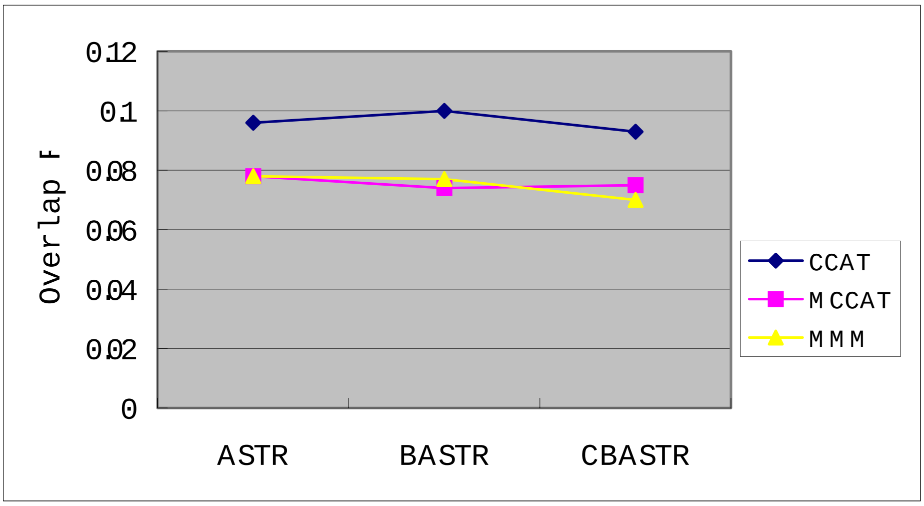


Figure 4 shows that there was main effect on item overlap rate due to content balancing method. In particular, CCAT yielded much higher overlap rates while MCCAT and MMM performed similarly. It appeared that there was no main effect by stratification on this aspect.

Figure 4: Item Overlap across Stratification Designs and Content Balancing Methods



Conclusion

As more and more large-scale educational tests have been fully or partially converted into the form of CAT, item selection rules that can simultaneously (i) improve test security by controlling item overlap rate and number of over-exposed items, (ii) enhance pool utilization, and (iii) satisfy practical constraint of strict content balancing, will definitely receive much attention. The stratified CAT designs have been developed to tackle the first two issues. But they have not been examined when integrated with different content balancing methods to deal with strict content specifications. This study aimed to provide valuable information for researchers and practitioners of the field by investigating the performances of item selection methods formed by various combinations of stratification designs and content balancing methods.

Results indicate that the nine selection methods performed more or less the same regarding reliability and measurement efficiency. There was no evidence of any interaction effect due to stratification design and content balancing method. However, these two factors had main effects on pool utilization or item security, or both. It was found that CBASTR is the best stratification design in terms of pool utilization and control of over-exposed items. The main reason may be that this design keeps sufficient proportions of items of individual content areas in each stratum. Thus, even towards the end of a test, there are still quite a number of candidate items for the desired content.

Among the three content balancing methods, CCAT was the most inadequate in both pool utilization and control of item overlap. Its weakness is due to the associated predictability of content sequence that repeatedly restricts the choice of an appropriate item from a small set of candidates at each point of selection. On the contrary, the randomization mechanism in MMM resulted in a better utilization of the entire pool.

In conclusion, the results of the present study suggest that the combination of CBASTR and MMM is an optimal item selection method for CAT that needs simultaneously to control item overlap, utilize the entire pool, and satisfy all content specifications. The current findings seem applicable to some common pool structure that has two popular features: (i) the *a*- and *b*- parameters are significantly correlated, and (ii) some contents in general are more difficult and more discriminating than others. Future research may investigate how the performance of these integrated methods would vary with other factors such as pool characteristics, practical constraints, termination rules, and ability distribution.

References

Chang, H.H., Qian, J., & Ying, Z. (1999). *A-stratified multistage CAT with b-blocking*. Manuscript accepted for publication.

Chang, H.H., & Ying, Z. (1999). A-stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H.H., Qian, J., & Ying, Z. (in press). A-stratified CAT design with b-blocking. *Applied Psychological Measurement*.

Chen, S.Y., & Ankenmann, R.D. (1999, April). *Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing*. Paper presented at the Annual Meeting of the American Educational Research Association, 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.

Hau, K.T., & Chang, H.H. (in press). Item Selection in Computerized Adaptive Testing: Should More Discriminating Items be Used First?. *Journal of Educational Measurement*.

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.

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.

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

Mills, C.N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*, 287-304.

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., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement*, *22*, 271-279.

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *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.

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

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

Yi, Q., & Chang, H.H. (2000). *Multiple Stratification CAT Designs with Content Control*. Unpublished manuscript.