Solving Complex Constraints in a-Stratified Computerized Adaptive Testing Designs

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

Abstract

The information based item selection methods in administering computerized adaptive tests (CATs) tend to choose the item that provides maximum information at an examinee's estimated trait level. As a result, these methods can yield extremely skewed item exposure distribution in which items with high *a* values may be over-exposed while those with low *a* values may never be selected. Chang and Ying (1999) proposed the *a*-stratified design (ASTR) that attempts to simultaneously control the exposure of high *a* items and improve the utilization of low *a* items. The method has been demonstrated to be effective in improving the utilization of the entire pool without sacrificing the efficiency in ability estimation when it is used with certain types of item pools. Nevertheless, the ASTR may result in a number of items being over-exposed in some pools where the correlation between the *a*- and *b*-parameters is significant. To remedy the over exposure problem in such situations, Chang, Qian and Ying (1999) developed the *a*-stratified with *b*-blocking method (BASTR) based on ASTR. These two stratified methods have not been tested under situations where complex non-statistical constraints are imposed. The Weighted Deviation Model (WDM) was proposed by Stocking and Swanson (1993) to deal with severely constrained item selection in CAT. An adaptation of the general ideas of the WDM to ASTR and BASTR was investigated in this study. Results indicate that both ASTR and BASTR can satisfy most of the non-statistical constraints. The BASTR outperformed the other two methods concerned in that it effectively controlled item exposures, better utilized the entire pool, and substantially reduced the test-overlap rate.

Paper presented at the NCME Annual Meeting 2000, New Orleans, April 25-27, 2000.

Introduction

The advances in modern computer technology and psychometrics have triggered the change of format of conventional paper-and-pencil (P&P) tests to the form of computerized adaptive testing (CAT) which was first developed under the item response theory models (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; Weiss, 1982). It is also better because it allows more flexibility in test schedule and the incorporation of alternate item forms (Straetmans & Eggen, 1998). 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 skew item exposure (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995; van der Linden, 1998). In particular, items with large value of discrimination parameter may be overly exposed while some others are never used. This would eventually damage test security and increase the cost in developing and maintaining item pools.

It is understandable, therefore, the control of item exposure and the enhancement of pool efficiency are important issues in computerized adaptive testing designs (Mills & Stocking, 1996; Stocking & Swanson, 1998; Way, 1998). Methods that 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) 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 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 was shown through simulation studies to be effective in both the reduction of test-overlap rate and the enhancement of pool utilization when used with certain type of item pools.

Nevertheless, the correlation of *a*- and *b*- parameters of the items in some pools may be significant. In such cases, there may not be sufficient items with low *b*s in the last stratum. Consequently, the ASTR would result in quite a number of items being over-exposed. To tackle such problem, Chang, Qian and Ying (1999) developed the *a*-stratified with *b*-blocking method (BASTR) based on ASTR. In BASTR, item pool is first divided into many small levels based on *b* parameters. Within each level, items are sorted in ascending order of *a* values. Then the items with the smallest *a* values from each level are grouped into the first stratum, and the items with the second smallest *a* values grouped into the second stratum,…, and eventually the last stratum contains those items with the largest *a* values from each level.

In many situations, however, CAT design has to take into consideration additional constraints such as content balancing, item type and statistical specifications. Various models have been proposed to solve such complex and sometimes conflicting requirements in test assembly (Armstrong, Jones & Kunce, 1998; Luecht, 1998; Stocking & Swanson, 1993, 1998; van der Linden & Reese, 1998). One strategy, as advocated by Stocking and Swanson (1993) in the Weighted Deviation Model (WDM) is to relax test constraints as desired properties. It can deal with many kinds of constraints and provide acceptable solution to those systems that may not have an optimal solution satisfying all criteria.

This paper reports the findings of a simulation study on an integration of WDM with the ASTR and BASTR. This means that complex test constraints were imposed in all item selection stages of the two stratified designs. As a result, tests administered using the enhanced stratified design had a better face validity. Three item selection methods were compared in the study: (1) WDM with maximum information approach (MI), (2) WDM with ASTR, and (3) WDM with BASTR. The comparison criteria include reliability, average bias, mean squared error, number of over-exposed items, test-overlap rate, chi-squared statistic, number of under-utilized items and deviations from the desired properties.

Simulation Studies

A series of simulation studies were conducted to investigate the performance of ASTR and BASTR when complex constraints were imposed. We started the investigation with 19 non-statistical constraints come from two common intrinsic item features: content area and cognitive level. The sample consisted of 5,000 examinees with abilities (referred as true ability θ hereafter) randomly generated from N(0,1). Each examinee received three simulated tests administered respectively by the three item selection methods incorporated with WDM: MI, ASTR, and BASTR.

In each test, information obtained by three artificial items in addition to the prescribed length was used to mimic prior information about the examinee. All the three items had the same values for *a* (= 1) and *c* (= .2). The value for b_1 of the first item was randomly selected from N(0, 1). If the first item was answered correctly, the second item was more difficult and the value for b_2 was set at $b_1 + 2$, otherwise $b_2 = b_1 - 2$. The procedure was repeated for the third item. The use of maximum likelihood estimation, as in all the three methods, was possible with the administration of these three artificial items.

Item pool

The item pool contained 700 upper primary mathematics items from four content areas crossed with three cognitive levels. Table 1 shows the distribution of the items. The mean estimated discrimination (*a*) value for the items was 1.02, with SD = .33, and a range of .29 to 2.63. The mean estimated difficulty (*b*) was .16, with SD = 1.06, and a range of –3.44 to 3.40; thus there was a close match between the *b* distribution and ability distribution. The mean estimated pseudo-guessing parameter (*c*) was .17, with SD = .008, and a range of .03 to .50. The correlation of *a*- and *b*-parameters was .45.

Table 1: Item Distribution across Content Areas and Cognitive Levels

	Cog1	Cog2	Cog3
Cnt1	52	142	40
Cnt2	33	104	29
Cnt3	27	93	30
Cnt4	28	81	41

Test constraints

The test length was fixed at 32 items. Based on the pool characteristic and subject concern, the following desired ranges and relative weights were imposed for the various combinations of content area and cognitive level as shown in Table 2.

Table 2: Ranges and Weights for 19 Non-statistical Constraints

	Lower Bound	Upper Bound	Weight
Cnt1	7	13	10
Cnt2	4	11	10
Cnt3	3	10	10
Cnt4	3	10	10
Cog1	3	10	10
Cog2	16	23	10
Cog3	3	10	10
Cnt1, Cog1	1	5	7
Cnt1, Cog2	3	10	7
Cnt1, Cog3	1	4	7
Cnt2, Cog1	0	3	7
Cnt2, Cog2	2	8	7
Cnt2, Cog3	0	3	7
Cnt3, Cog1	0	3	7
Cnt3, Cog2	1	6	7
Cnt3, Cog3	0	3	7
Cnt4, Cog1	0	3	7
Cnt4, Cog2	1	6	7
Cnt4, Cog3	1	5	7

Item selection methods

The performance of the following three item selection methods were compared when WDM was incorporated to deal with complex constraints. In the WDM, test constraints were treated as desired properties. Thus, when there was no feasible solution to the system of the constraints, the objective function could find an item yielding the minimal deviation from the constraints. The model here was:

Minimize

$$\sum w_j d_{L_j} + \sum w_j d_{u_j} + w_\theta d_\theta \qquad \text{(sum of weighted deviations)} \qquad (1)$$

Subject to

$$\sum_{i=1}^{N} a_{ij} x_i + d_{L_j} - e_{L_j} = L_j \quad , j = 1, \ldots, 19 \qquad \text{(for lower bounds of constraints)} \quad (2)$$

$$\sum_{i=1}^{N} a_{ij} x_i + d_{U_j} - e_{U_j} = U_j \quad , j = 1, \ldots, 19 \qquad \text{(for upper bounds of constraints)} \quad (3)$$

$$\sum_{i=1}^{N} I_i(\theta) x_i + d_\theta - e_\theta = \infty \qquad \text{(for max. information criterion)} \ (4)$$

$$d_{L_j}, d_{U_j}, e_{L_j}, e_{U_j} \geq 0 \quad , j = 1, \ldots, 19 \qquad (5)$$

and

$$x_i \in \{0, 1\}, i = 1, \ldots, N, \qquad \text{(Stocking \& Swanson, 1993, p. 280-281)} \ (6)$$

where a_{ij} equaled to 1 if item *i* had property *j* and 0 otherwise; w_j was the weight assigned to constraint *j*; w_θ was the weight assigned to the information constraint; L_j and U_j were the lower bound and upper bound of constraint *j*; d_{L_j} and d_{U_j} represented deficit from the lower bound and surplus from upper bound respectively; e_{L_j} and e_{U_j} represented excess from lower bound and deficit from upper bound respectively; x_i equaled to 1 if *i*th item was included in the test, or 0 otherwise.

The heuristic for item selection using WDM was also adopted:

(i) For every item not already in the test, compute the deviation for each of the constraints if the item were added to the test;

(ii) Sum the weighted deviations across all constraints;

(iii) Select the item with the smallest weighted sum of deviations (p. 281).

Method 1 (MI):

1. The content area and cognitive level of each item were indexed.
2. The optimal function of Equation 1 was used.
3. Test constraints and maximum information were treated as desired properties as stated in Equations 2 to 4.
4. The bounds and weights for the 19 non-statistical constraints were set as stated in Table 2. The lower bound for information was set at 100.0 which was practically unreachable in the study. The weight for information was assigned the value of 15.0 that was used by Stocking and Swanson (1993) and was found here to have yielded very small mean squared error.
5. Items were selected by following the three steps of the heuristic mentioned above.

Method 2 (ASTR):

1. The content area and cognitive level of each item were indexed.
2. Test constraints were mathematically formulated as Equations 2 and 3.
3. The item pool was partitioned into 4 strata according to the ascending order of the *a*-parameter values of the items. Each test was divided into 4 stages. In the first stage, items were

administered from the first stratum. In the second stage, the next group of items were administered from the second stratum,…, and the last group of items from the *4*-th stratum.

4. At individual item selection level in each stage, five unadministered items with difficulty parameter closest to the currently estimated theta were chosen. For each of them, the absolute difference between the *b* value and the current ability estimate was computed. The weight for such absolute difference was set at 20.0. Then the total weighted deviations were summed up as if the item had been added to the test. The item with the smallest weighted sum of deviations was administered.

Method 3 (BASTR):

In this method, the procedures for item administration were similar to those in Method 2 (ASTR), but the item pool was partitioned into 4 strata as follows. First the pool was divided into 175 different levels based on the *b*-parameters so that the four items in each level had similar *b* values. Within each level, the items were sorted according to the ascending order of *a*-parameters. Items with the lowest *a* values from each level were assigned to the first stratum whilst items with highest *a* values were assigned to the last stratum. As a result, each stratum then had similar distribution of the *b*-parameters but the average value of the *a*-parameters increased across the strata.

Measure of performance

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 interpreted as the correlation ratio of the estimated scores on the true scores (Lord, 1980, p. 52). The higher the reliability, the better the item selection method would be.

Bias: Accuracy is another important criterion, which in this study was measured by the bias and mean squared error. Let θ_i, i = 1,…, 5000 be the true abilities of the 5000 examinees and $\hat{\theta}_i$ be the respective estimators from the CAT. Then the bias was computed as

$$\text{Bias} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i) \tag{7}$$

The smaller the bias, the better the item selection method would be.

Mean squared error: Using the same notations of true ability and the estimator, mean squared error was computed as

$$\text{MSE} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i)^2 \tag{8}$$

The smaller the MSE, the better the item selection method would be.

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 one of the key evaluation criteria. Here an item was considered as 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. Wightman (1998) explicitly argued that efficient usage of the available items in the pool would be an important evaluation criterion of CAT assemblies. 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:

$$\chi^2 = \sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N} \qquad (9)$$

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

Equation 9 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.

Test-overlap rate: The test-overlap rate is another important summary index in measuring item exposure control (Mills & Stocking, 1996; Way, 1998). Test-overlap rate is indicated by the proportion of items shared by pairs of examinees, averaged across all possible pairwise combinations. Way (1998) 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 test-overlap rate, the bigger the damage to the test validity due to information sharing among examinees who take the test at different times. 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 test-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 test-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 test overlap.

Deviations from desired ranges: As the desired ranges were recommended by the experts of the relevant fields, it was anticipated that the CATs by all selection methods should have fallen within these ranges as far as possible. The less the deviations from the desired ranges of the non-statistical constraints, the higher the face validity and thus the better the item selection method would be.

Results

The results of the study are summarized in Table 3. The three item selection methods were virtually unbiased as they all yielded biases close to zero. They had high and comparable reliabilities from .93 to .96. In terms of measurement efficiency, the MI was the best as it provided the smallest MSE of .42, and BASTR was relatively less efficient whilst ASTR was in between these two methods. However, the efficiency of MI in trait estimation was at the expense of item and test security. It over-exposed 63 items and yielded unacceptably high exposure rates up to .63. Besides, its test overlap-rate of .265 was much higher than the safe standard of .15 prescribed by Way (1998), meaning that in average about 8 to 9 items in a 32-item test were identical among any two randomly selected examinees. On the contrary, the BASTR offered the best control on item and test security. It over-exposed no item and it yielded the maximum item exposure rate of .16 that was below the cut-off value of .2 and far below the .63 in MI. In addition, the BASTR provided the smallest test-overlap rate of .057, meaning that in average about 1 to 2 items were identical in any two 32-item tests. The performance of the ASTR was in between that of the MI and BASTR. It over-exposed 15 items with a maximum exposure rate of .33 and yielded a small test-overlap rate of .081.

In terms of pool utilization, the BASTR was the best as it yielded the smallest Scaled χ^2 of 9.9. The F ratios of $F_{BASTR,MI}$ and $F_{ASTR,MI}$ were 9.9/155.2 (i.e. .06) and 31.6/155.2 (.20), meaning that there was 94% reduction of skewness of exposure distribution in BASTR relative to MI and 80% reduction in ASTR relative to MI respectively. The BASTR provided the smallest number of under-utilized item of 98 that was far below than the figures of 513 and 201 offered by MI and ASTR respectively. In fact, it utilized every single item in the pool as reflected by the smallest exposure rate of .0004.

The MI offered a better face validity as it had the smallest number (192) out of 5,000 CATs with only a single deviation beyond the 38 bounds associated with the 19 non-statistical constraints. There were quite a number of CATs from BASTR and ASTR with one to three deviations beyond the bounds. Although the maximum total deviation of 3 may be acceptable by the general public, the ASTR and BASTR need to be improved in order to provide a better face validity.

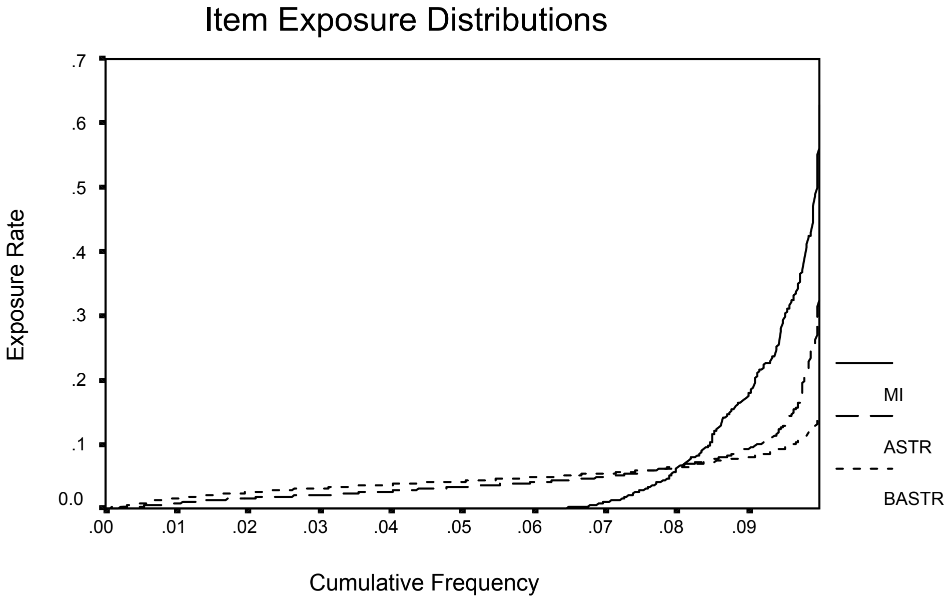
Table 3: Summary Statistics for Three CAT Methods

	MI	ASTR	BASTR
Bias	.001	.014	-.004

MSE	.042	.067	.076
Reliability	.96	.94	.93
Scaled χ^2	155.2	31.6	9.9
N(exp<.02)	513	201	98
N(exp>.2)	63	15	0
Min exp	.0000	.0000	.0004
Max exp	.63	.33	.16
Overlap Rate	.265	.081	.057
Deviations from Non-statistical Constraints			
N(1 dev)	192	765	1123
N(2 dev)	0	352	323
N(3 dev)	0	196	91
N(4 or above)	0	0	0

Figure 1 shows the item exposure distributions for the three selection methods. The curve for MI is flat over halfway through and then rises rapidly. It is very clear that more than 60% of the pool in MI was never touched while about 10% was over-exposed, showing that this method repetitively selected those more informative items in order to maintain its efficiency in trait measurement. Both the item security and the cost effectiveness of development and maintenance of such a pool in MI are in doubt. In contrast, the curve for BASTR rises steadily all the way through with a minimum of .0004 and a maximum of .016 (refer to Table 3), showing that the entire pool was well utilized and the item exposures were under control. The curve for ASTR is close to but below that of BASTR for most of the way and then rises a bit rapidly near the end, showing that ASTR well utilized most items but some measure on exposure control was needed.

Figure 1: Item Exposure Distributions for Three CAT Methods



Note: the Cumulative Percentage was computed as the proportion of the number of items concerned divided by the pool size.

Discussion

The ASTR was developed in an attempt to remedy the problems associated with the information-based item selection methods: extremely skewed item exposure distribution and high test-overlap rate. Based on this method, BASTR has been proposed in order to tackle the situations in which the correlation between *a*- and *b*- parameters is significant. The main objective of this study was to investigate whether these two stratified methods could be applied to testing programs where complex constraints are imposed. An integration of WDM with these two methods was proposed.

Results indicate that both ASTR and BASTR, when integrated with WDM, can meet most of the test constraints. These two methods were relatively less efficient in trait estimation than MI, but they offered better item and test security measure as well as better utilization of the item pool. The general problems of (i) over exposure, (ii) high test-overlap rate, and (iii) extremely skewed exposure distribution, associated with information-based selection methods were found in MI as it repetitively used some more informative items. As the ASTR and BASTR used different strategy in item selection, the last two problems did not exist with these methods. Besides, the problem of over exposure was much smaller in ASTR and was not seen in BASTR in which each stratum had a wide range of *b*-parameters. Though the face validity of CATs by ASTR and BASTR was much improved by integrating with WDM, some more thoughts on pool partition and item selection are needed for better results.

There are limitations on the generalization of the findings from the present study. Firstly, the 38 bounds associated with the 19 non-statistical constraints in this study were based on two common intrinsic item features: content area and cognitive level. Sometimes, test programs have to deal with very complicated requirements when other constraints such as item overlaps and item sets are added. Further research includes the investigation of the performance of ASTR and BASTR in such complicated situations. Secondly, the correlation between the *a*- and *b*- parameters of the items was quite high. The performances of these three item selection methods may vary with the value of correlation as well as other factors such as test length, pool size and item distributions across intrinsic item features. In fact, the results have shown that each item selection method studied has its own merits and shortcomings. None of them is superior in all aspects. Thus, the choice on item selection method really depends on the pool characteristics and the needs of individual testing programs.

References

Armstrong, R.D., Jones, D.H., & Kunce, C.S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*, 237-247.

Chang, H.H., Qian, J., & Ying, Z. (1999). *a-Stratified Multisage CAT with b-Blocking*. Unpublished manuscript.

Chang, H.H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement, 20*, 213-229.

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

Chang, H.H., & Zhang, J. (1999, June). *Hypergeometric family and test overlap rates in computerized adaptive testing*. Paper presented at the annual meeting of the Psychometric Society, Lawrence, KS.

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.

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.

Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236.

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.

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

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277-292.

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

Straetmans, G.J., & Eggen, T.J. (1998). Computerized adaptive testing: what it is and how it works. *Educational Technology*, *38*, 45-52.

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.

Thomason, G.L. (1995, June). *New item exposure control algorithms for computerized adaptive testing*. Paper presented at the Annual Meeting of Psychometric Society, Minneapolis, MN.

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

van der Linden, W.J., & Reese, L.M. (1998). A model for optimal constrained adaptive testing.

Applied Psychological Measurement, *22*, 259-270.

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.

Wightman, L.F. (1998). Practical issues in computerized test assembly. *Applied Psychological Measurement, 22*, 292-302.