Optimum Number of Strata in the a-Stratified Computerized Adaptive Testing Design

Jina-Bing Wen, The Chinese University of Hong Kong
Hua-Hua Chang, University of Texas, Austin
Kit-Tai Hau, The Chinese University of Hong Kong

With the advancement in computer technology and respective psychometric theories, computerized adaptive testing (CAT) has moved from pure research to large scale implementation during the early 1990s. In the a-stratified method, a popular item exposure control strategy proposed by Chang (Chang & Ying, 1999; Hau & Chang, in press), the item pool and item selection process has been usually divided into four strata and the corresponding four stages. In this study, the optimum number of stages and strata with respective to item pool and testing characteristics were explored.

In CAT, tailoring items to test-takers' ability through the selection of appropriate items would be desirable because an examinee is measured most effectively when the items are neither too difficult nor too easy. The logic behind the most prevalent item selection strategy can be mathematically derived (Hau & Chang, in press). In item selection, aside from non-statistical considerations such as content balancing, the most common strategy in the last three decades has been the maximization of item information. Specifically, an item will be selected if it has the maximum information at the currently estimated θ level, which is calculated from the examinee's available responses at that instant (see also other alternatives, e.g., Chang & Ying, 1996; Owen, 1975).

Item information has been typically defined as Fisher information that varies as a function of the test-taker's ability θ. Consider the simple case when all items follow $c \equiv 0$ (i.e., a two parameter model). Then, Fisher information increases monotonically with a, items with high a's will be preferentially selected (e.g., see Hau & Chang, in press).

Test Security, Exposure Control and a-Stratified Design

Test security has been a serious problem in CAT. In contrast to a paper-and-pencil test where examinees are tested with an identical set of items at the same time, in a CAT examinees are tested individually at different sessions with items which will be reused at a later time. Understandably, test security becomes a problem because examinees can remember and share the item content with others. To avoid item content leakage, it is therefore important to control the frequency with which an item has been administered to test-takers. In other words, monitoring items' exposure rate to prevent overexposure is necessary to enhance test security.

Remedies to restrain the over-exposure of high discrimination items have been proposed by McBride & Martin (1983), Sympson and Hetter (1985), Stocking and Lewis (1995), Davey & Parshall (1995), Thomasson (1995), and others. This issue has drawn particularly great attention from researchers when CAT is implemented in high stake tests like TOEFL and ASVAB-CAT.

Working with a totally different item selection philosophy in that a proactive mechanism should be devised to equalize the exposure of high and low discrimination items, Chang (see review, Chang & Ying, 1999) demonstrated the benefit of using their multi-stage a-stratified design.

Essentially in the a-stratified method, the item pool is divided into several strata in an ascending order of their discrimination parameter (for details see Chang & Ying, 1999 or Hau & Chang, in press). The corresponding CAT is also divided into the same number of stages. In each stage of testing, items with maximum information are selected from the corresponding pool stratum. Thus, items with smaller a-parameters are selected first while larger a-parameter items are left for latter stages. Since the estimates of examinee's ability are not close to the true value during early stages, the use high a-parameter items do not necessarily imply a greater precision in ability estimation. Actually simulation studies showed that this method can equalize item exposure without damaging ability estimation efficiency and accuracy (Chang & Ying,1999).

If test security is the only concern, then all examinees should be given a random sample of items from the pool. The random selection tends to approximately equalize the exposure rates of all items in the pool and consequently will help to minimize the item overlap among examinees. On the other hand, if efficiency in ability estimation is the only concern, then according to Fisher information criterion, the high discrimination items should be used instead. The efficiency gain will be at the expense of the unbalanced item usage and the greater cost in item replenishment. In other words, if the total budget in test maintenance is kept constant, apparently there is a tradeoff between test security and efficiency. If both factors are important as in a high stakes examination, then the testing agency has no choice but to spend more money on test development and maintenance, which subsequently results in a many folds increase in the examination fee. Despite the seeming incompatibility between test security and efficiency, the above tradeoff may be avoidable if a method can be found that has a balanced item usage yet maintains efficiency.

The a-stratified strategy has at least three potential advantages. Firstly, it may provide an estimation efficiency comparable to the traditional maximum information approach. Secondly, it automatically leads to a more even item exposure rate control. The major cause for unevenly distributed item exposure and subsequent security problems is that large *a* items are more likely to be selected than the small *a* ones. In the A-STR method, exposure rates will become more evenly distributed because proportionally equal numbers of items are chosen from strata of high, medium and low *a* parameters. Thirdly, in comparison to maximum information integrated with Sympson and Hettermethod, the stratified method is simpler to implement (see Hau & Chang, in press).

Optimum Number of Stratum

In most of the stratified designs (e.g., Chang & Ying, 1999; Hau & Chang, in press), four strata have been used. However, there has not been any attempt to determine how the number of strata would affect the efficiency and item over-exposure. There can be two extremes in the number of strata. On one extreme, if only one stratum, instead of the usual four strata, is used, then all items will be in the same stratum. Within this stratum, items with difficulty nearest to the examinee's current estimated ability will be selected. In that case, such a stratified design will differ from the maximum information approach in that in the former design, the discrimination parameter has not been considered. Thus, such a stratified design with one stratum should have an efficiency lower than that of the maximum information approach. However, if the distribution of item difficulty matches that of the examinees, then item usage will be relatively balanced.

On the other hand, if the number of strata equals to the preset test length, then these strata and hence the items selected will be arranged strictly in the order of ascending discrimination items. That is, item selection will always start from the stratum with the lowest discrimination items and then the items selected will monotonically increase in discrimination. If there are insufficient items of diversified difficulties within each of these strata, then dividing the item pool into many strata may decrease the chance of getting an item close enough to the desired difficulty. In that case, efficiency in ability estimation will suffer, but the impact on item usage may be quite complicated depending on the original pool characteristics.

It can also be speculated that the overall testing performance depends on the number of strata and hence the size of items within each stratum. If there are many items of various levels of discrimination and difficulty within each stratum, then using many strata will lead to a relatively high efficiency, while perhaps at some degree of sacrifice of a more balanced item usage.

The present study will examine the above hypothesis as regards the optimum number of strata through simulation studies with item pool imitating operational conditions as well as other characteristics. The objective is to find the relationship between testing performance (efficiency and item pool usage) the stratification process (number of strata adopted).

Simulated Studies

In a series of simulated studies, we systematically varied the Number of Strata in the stratified approach under a 3 Pool Size (number of items in the pool, 3 levels) X 3 Item Characteristics design.

Pool Size. The three sizes being examined were 200, 400, 800 items.

Item Characteristics. The first set imitated item characteristics of a large scale operational pool. The second and third sets were purposely designed to examine how item characteristics might interact with the number of strata. Both sets contained items with a normal distribution of item difficulty matching students' ability distribution. The second set displayed a hypothetical situation in which item difficulty and discrimination were not related in the sense that within each ability range, there were items with various levels of discrimination (a = 0.4 to 2.0). On the other hand, the third set demonstrated a situation in which difficulty was correlated with discrimination at .5. That means more difficult items were relatively more discriminating while easier items were relatively less discrimination.

Latent trait distribution. Five thousand θ values were generated from a standardized normal distribution $N(0,1)$.

Test length and termination rule. The test length examined was 48 items. Items were selected according to the stratified design as described above.

Number of Strata. The pool was partitioned into 1, 2, 3, 4, 6, 8, 12, 16, 24, 48 strata.

Estimation of θ. The maximum likelihood method was used to estimate the ability for each of the 5,000 examinees.

Evaluation Criterion. Test efficiency and accuracy (bias, mean square error); item pool usage (test-overlap rate, Chi-squared statistic, number of over- and under-exposed items)..

Results and Discussion

The results were in general agreement with our speculation that too few and too many strata might not provide the optimum efficiency and balanced item pool utilization. It was found that the ideal and optimum number of strata to be used in each specific application depended on the item pool structure, test length and other testing conditions. The results showed that test efficiency and the balanced usage of all items might not necessarily increase or decrease monotonically with the number of strata.

Implications for item pool management and future studies are discussed. It is recommended that in an operational CAT design, the optimum number of strata can be determined through simulation studies under conditions specifically chosen for that particular application. Furthermore, future directions of research in which the philosophy of using less discrimination items in the earlier stages of testing without physically partitioned and stratification of the item pools are also elaborated and discussed.

References

Chang, H. H. & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23(3),* 211-222.

Davey, T., & Parshall, C. (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, CA.

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.

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* (p223-226). New York, Academic Press.

Owen, Z. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of 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, October). *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.

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