To Weight or Not to Weight – Balancing Influence of Initial and Later Items in Adaptive Testing[1]

Hua-Hua Chang
University of Texas at Austin

Zhiliang Ying
Columbia University

[1] Paper presented at the Annual Meeting of National Council on Measurement in Education, New Orleans, LA, April 2, 2002.

To Weight or Not to Weight – Balancing Influence of Initial and Later Items in Adaptive Testing

1. Objective of the Study

In computerized adaptive testing (CAT), such as the Graduate Record Examination (GRE), some examinees may get much lower scores than they would normally get if an alternative paper and pencil (P&P) version were given. This has become an arising issue in online implementation of CAT, and this phenomenon is unique to CAT because of the nature of its sequential selection rule. However, without effective remedial measures, it could significantly undermine the credibility of CAT. In this paper, we would like to address some of the issues, find causes, and propose possible solutions.

2. Problems and Background

The original motivation for adaptive (tailored) testing is to match items with the examinee's trait level (Lord, 1971). Under the usual Item Response Theory (IRT) models, maximizing Fisher information is intuitively to matching item difficulty parameter values with the latent trait level of an examinee. Since the latent trait is unknown, the optimal item selection rule could not be implemented, but may be approximated using the updated estimate each time when a new item is to be selected. This is essentially the basic design behind Lord's original proposal of tailored testing.

The adaptive item selection adds a new dimension of uncertainty to the outcome and evaluation of CAT. One major factor of uncertainty comes from inaccurate estimation of the trait parameter in the initial stages when the number of administered items is small. This could result in grossly underestimating the trait level at early stages. In consequence, easy items are likely to be administered due to the rule of *matching "difficulty" to "ability"*. As to be demonstrated subsequently, such items are ineffective to bring the estimate close to the true trait level, unless the test is sufficiently long or a variable-length test is used.

The main purpose of this investigation is to quantitatively reveal some of the causes that account for underestimation of latent trait θ in CAT. The logistic models, particularly the Rasch model and the 2PL model, are used to demonstrate some key observations. They point to weighting likelihood score as a possibility to alleviate the problem of underestimation. In this connection, the stratified approach by Chang and Ying (1999) is shown to improve the adaptive testing in a natural and automatic way. The main analytic results are presented in the next section. Section 4 summarizes findings from numerical studies. The last section gives some additional remarks.

3. Main Results

The classical IRT assumes that the probability of an examinee given the correct answer to item *i* (Y_i=1) takes form

$$P\ (Y_i = 1 \mid \theta) = P_i(\theta),$$

where θ is the examinee's latent trait and $P_i(\theta)$ is monotone increasing in θ (Lord, 1980). Suppose that an examinee with a fixed θ is given *n* items $Y_1, Y_2, ..., Y_n$. Then θ can be estimated by maximizing the likelihood function

$$L_n(\theta) = \prod_{i=1}^{n} P_i(\theta)^{Y_i} Q_i(\theta)^{1-Y_i}, \tag{1}$$

where $Q_i(\theta) = 1 - P_i(\theta)$. Let $\hat{\theta}_n$ denote the resulting estimator. It is clear that $\hat{\theta}_n$ also solves the following maximum likelihood estimating equation

$$U_n(\theta) = \frac{\partial}{\partial\theta} \log L_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log \frac{P_i(\theta)}{Q_i(\theta)} [Y_i - P_i(\theta)] = 0. \tag{2}$$

Note that (2) has an interesting form and we can think of $Y_i - P_i(\theta)$ as

"observed" – "expected",

which therefore has mean 0. Thus $U_n(\theta)$ is a weighted sum of *"observed" – "expected"*.

It is well known that, under suitable regularity conditions, $\hat{\theta}_n$ is asymptotically normal, centered at the true θ and with variance approximated by $I_n^{-1}(\hat{\theta}_n)$, where $I_n(\theta)$ is the Fisher information function. An original motivation for CAT is to maximize the Fisher information so as to make $\hat{\theta}_n$ most accurate. This can be achieved by recursively estimating θ with current available data and assign further items adaptively. The asymptotic normality and validity of using Fisher information to estimate variance continue to hold under the adaptive item allocation of CAT.

To illustrate possible sensitivity of $\hat{\theta}_n$ in CAT to result on initial items and to motivate remedies, let us first consider the case of the Rasch model. Without loss of generality, we assume the common *a* parameter to be 1. The likelihood estimation function takes the same form, i.e.,

$$U_n(\theta) = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \right), \tag{3}$$

after *n* items were administered. For the MLE $\hat{\theta}_n$, $U_n(\hat{\theta}_n) = 0.$

Let b_{n+1} denote the next item selected. Then $\hat{\theta}_{n+1}$ solves $U_{n+1}(\hat{\theta}_{n+1}) = 0.$ By the mean-value theorem,

$$U_{n+1}(\hat{\theta}_{n+1}) - U_{n+1}(\hat{\theta}_n) = -I_{n+1}(\theta^*_{n+1})(\hat{\theta}_{n+1} - \hat{\theta}_n)$$

where θ^*_{n+1} lies between $\hat{\theta}_{n+1}$ and $\hat{\theta}_n$, and $I_{n+1}(\theta) = \sum_{i=1}^{n+1} e^{\theta - b_i} \Big/ \left(1 + e^{\theta - b_i}\right)^2$ is the Fisher information and is also equal to $-\frac{\partial U_{n+1}(\theta)}{\partial \theta}$. Since $U_n(\hat{\theta}_n) = 0$, it can be shown

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{1}{I_{n+1}(\theta^*_{n+1})}\left(Y_{n+1} - \frac{e^{\hat{\theta}_n - b_{n+1}}}{1 + e^{\hat{\theta}_n - b_{n+1}}}\right). \tag{4}$$

If the item pool is sufficiently rich that allows each given θ to match a difficulty parameter b with the same value, then $b_{n+1} \approx \hat{\theta}_n$ or $e^{\hat{\theta}_n - b_{n+1}} \Big/ \left(1 + e^{\hat{\theta}_n - b_{n+1}}\right) \approx \frac{1}{2}$. This entails that the one-step update from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$ is $\pm\frac{1}{2}$ multiplied by $I^{-1}{}_{n+1}(\theta^*_{n+1})$, which is typical of order 2/*n* for large *n*. Consequently, the larger the *n* is, the smaller the one-step adjustment it gets. It is plausible that if the examinee misses a number of initial items and the test length is shot to moderate, then he/she may not be able to regain a score (estimate) comparable (close) to the true θ, even though he/she responds well to the rest of the items.

The situation becomes even interesting in the case of the 2PL model, which has the estimating function

$$U_n(\theta) = \sum_{i=1}^{n} a_i \left(Y_i - \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}\right), \tag{5}$$

and the Fisher information function becomes

$$I_n(\theta) = \sum_{i=1}^{n} a_i{}^2 \frac{e^{a_i(\theta - b_i)}}{[1 + e^{a_i(\theta - b_i)}]^2}. \tag{6}$$

Similar to recursion (4), we have

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{a_{n+1}}{I_{n+1}(\theta^*_{n+1})}\left(Y_{n+1} - \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1 + e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}\right). \tag{7}$$

Through their analytic derivations and simulation studies, Chang and Ying (1999) argued that, provided necessary constrains are met, the *a*-parameter should be selected in an ascending order. Their motivations come from the considerations of efficiency improvement and item exposure balance. In view of (7), an additional benefit of the *a*-stratified approach of Chang and Ying is that it automatically adjusts step sizes in updating current estimation of θ. Specifically, it shrinks weights at early stages, making it less likely to have extreme values in estimating θ. It also inflates weights at final stages, counteracting the effect of the multiplier $I^{-1}{}_{n+1}(\theta^*{}_{n+1})$ and making it more likely to adjust the final estimator of θ.

For the 3PL model, the recursion becomes

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{\xi_{n+1} a_{n+1}}{I_{n+1}(\theta^*{}_{n+1})}\left(Y_{n+1} - \left(c_{n+1} + (1-c_{n+1}) \frac{e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}}{1+e^{a_{n+1}(\hat{\theta}_n - b_{n+1})}} \right) \right), \tag{8}$$

where ξ_n is a sequence that fluctuates around some certain constant in the design of Chang and Ying (1999). So, again, the ascending order of a_n as advocated in Chang and Ying plays pivotal roles to ensure higher efficiency, more balanced exposure rates as well as to reduce fluctuation due to initial item response irregularity and increase effectiveness in counteracting initial item influence by responses to later items.

4. Simulation Studies

A pilot study was conducted.

Item Pool Structure

Assume the item pool is so sufficiently rich that for every given θ value one can find a corresponding difficulty parameter *b* with the same value. The item pool was partitioned into four strata and the discrimination parameter was identical within each stratum, with *a* equaled to 0.5, 1.0, 1.5 and 2.0 respectively for the four strata. Without jeopardizing the generality of the findings, all discrimination parameters within each stratum were kept constant to maintain a clear change in item parameter distribution as testing progressed. Within each stratum, values of difficulty *b* parameters can be generated with the same value of θ. For simplicity, the guessing *c* parameters for all items were set at zero.

Latent Trait Distribution

One thousand examinees were included with a fixed θ value at 0. Note that the true θ can be fixed at any point.

Test length and termination rule

The test length examined was 40 items. Though a bit longer than most adaptive tests, it was deliberately chosen to show the trend more clearly.

Item selection rules

Hau and Chang (2001) made a comparison of efficiency in terms of MSE among several methods and reported that the ascending *a*- method was better than the descending *a*-method. The focus of this research is to examine whether the use of more discriminating items at the beginning of testing would cause "convergence to a wrong point". Two approaches, namely, the descending *a*- and ascending *a*- methods, were designed in the simulation study. In the descending *a*-method, the use of large *a* parameter items were in the reversed order to that of the ascending *a*-method, that is, large *a* parameter items were used first followed by small *a* ones. The former method is parallel to the maximum Fisher information method whereas the latter is advocated by Chang and Ying (1996, 1999). However, the objective of the research is to defend a general principle -- *low-a items should be used first and high-a items should be used last*, thus it will not entail specific item selection methods, such as the maximum information method, the a-Stratified method, etc. In this regard the simulation design is essentially for all MLE based procedures that can be portrayed by Equation (7).

Step 1. The item pool is partitioned into *4* strata by the *a*-parameter, with the first and last strata containing, respectively, the smallest and the largest *a* items.

Step 2. Accordingly, the testing process is also partitioned into *4* stages to match the *4* item strata.

Step 3. At the *kth* stage, 10 items are selected from the *k*th stratum. The test-taker's ability is updated by Equation (7), which is equivalent to maximizing the likelihood function constructed from the responses to the items already taken. Then items of difficulty parameter equal to the estimated ability are selected and administered as the next item.

Step 4. Step 3 is repeated for $k = 1$ through $k = 4$ stages.

The steps in the descending *a*-method are:

Step 1. The item pool is partitioned into *K* strata by the *a*-parameter as in the ascending *a* method. However, contrary to the ascending *a* method, the earlier and latter strata now contain, respectively, the higher and lower *a* items.

Steps 2 to 4. Identical to the steps in the ascending *a* method, the entire testing process is also partitioned into 4 stages to match the 4 item strata.

Initial Estimators

Let $\hat{\theta}_1$ be the initial estimator of the true θ. Eleven initial estimation points were selected: $\hat{\theta}_1$=-3.0, -2.5, -2.0, -1.5, -1.0, 0, 1.0, 1.5, 2.0, 2.5, and 3.0. Note: the true θ=0.0.

Evaluation Criterion

(i) Average bias, and (ii) Mean squared error (MSE).

Results

Figure 1 shows the mean squared errors for the two methods. For each method, MSEs were calculated based on 1000 replications at each of the eleven starting points. The graph represents the MSEs as a function of the starting values. For the descending *a*-method, the larger the difference between the initial estimator and the true θ is, the higher the value of MSE is. See Figure 1 for details. The same pattern was found in the bias plots in Figure 2. These figures clearly indicate that the ascending *a*-method generated much more consistent results for both bias and MSE. The simulation results may also imply that if the item selection algorithm relies on items with the highest *a*-parameter values at the beginning of the test, it is plausible that if the examinee misses a number of initial items, then he/she may not be able to regain an estimated $\hat{\theta}$ that is comparable to the true θ, even though he/she responds well to the rest of the items. For instance, according to Figure 2, the average bias of the descending method at –3 was –2.8, however, according to Figure 3, the corresponding average number of correct items was 37.

5. Conclusions

The analytic derivations in this paper provided theoretical evidence to Chang and Ying (1999) and Hau and Chang (2001) that the ascending *a*-method should be used in computerized adaptive testing. The results of the simulation study provided certain confirmation to the speculation that using high *a*-item first may cause divergence if the student failed a couple of items at the beginning of the test.

The theoretical results presented in this paper may help us in designing more robust item selection algorithm. In view of (7), an additional benefit of the *a*-stratified approach of Chang and Ying is that it automatically adjusts step sizes in updating current estimation of θ. Specifically, it shrinks weights at early stages, making it less likely to have extreme values in estimating θ. It also inflates weights at final stages, counteracting the effect of the multiplier $I^{-1}{}_{n+1}(\theta^*{}_{n+1})$ and making it more likely to adjust the final estimator of θ. The stratified method is just only one of the many possible ways, including incorporation with the maximum Fisher information method, to implement the weighting philosophy.

The current simulation study was oversimplified. Different test lengths, more realistic item pools, and many constrains, such as content balance and item overlap rates control should be included in the future study.

Secondly, we will demonstrate that the adoption of (8), may also alleviate the problem. We intend to show that both the *a*-stratified method and the weighted Fishier information method perform well in practice. It is not so rare that an examinee in a fixed-length CAT may get an unusually poorer score than she/he deserves. A common cause is that the examinee may perform poorly on several initial items encountered. Under certain adaptive item selection rules, it is almost impossible for an examinee to recover with a limited number of later items, since they are likely to have low difficulty parameter values. We observe that weighting the likelihood score suitably may alleviate such problem. As a result, the estimator of examinee's latent trait will be less sensitive to a possible selection of extreme items in initial stages of a CAT. We show that this can be achieved either automatically with the *a*-stratified design or manually by inserting certain desirable weights into the likelihood equation. Many issues will be discussed after the 2002 NCME.

References

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*.

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

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. (2001). Item selection in computerized adaptive testing: should more discriminating items be used first? *Journal of Educational Measurement*,

Leung, C.K., Chang, H., & Hau, K.T. (2001, April). *Integrating stratification and information approaches for multiple constrained CAT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, 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.

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.

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.

Figure 1. MSE of the two methods

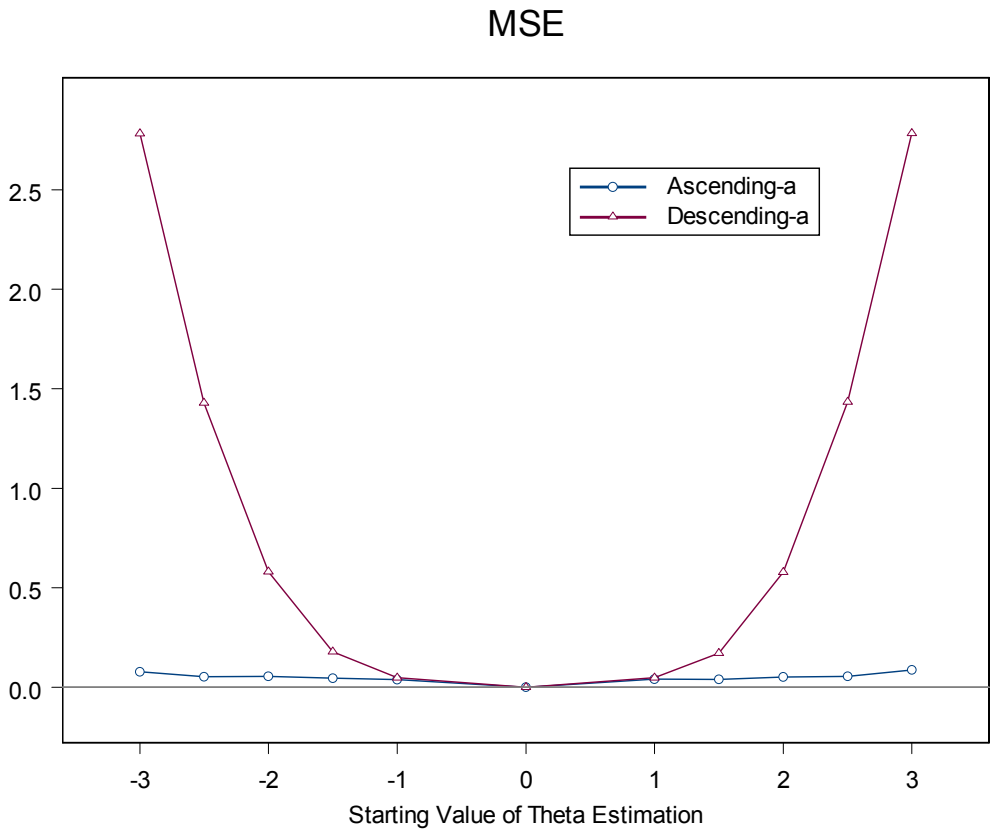

Figure 2. Average bias of the two methods.

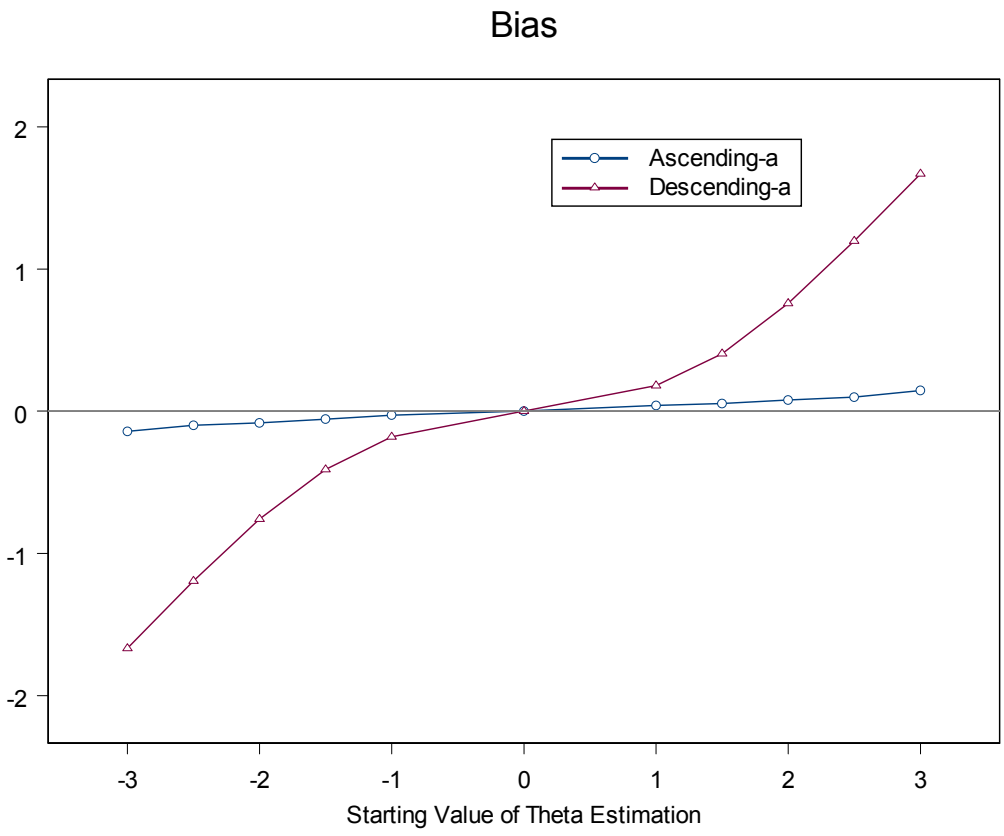

Figure 3. Average Number Correct Items of the Two Methods