# New Item Selection Designs for Computerized Classification Test


## Abstract

This study proposed the novel idea of "stage adaptive" to tailor the item selection process with the decision-making requirement in each step and generated fresh insight into the existing response time selection method. Results indicate that a balanced item usage and stable test-taking times can be achieved.


## Summary

### Objectives

In computerized classification testing (CCT), the main interest is to divide examinees into different proficiency groups. One of the long-term dilemmas faced by CCT is test security. As CCT trends to select items based on maximum Fisher information at the fixed cut scores (MFC) to make decisions reliable (Spray & Reckase, 1996; Thompson, 2009), all examinees will be presented the same set of items. Another increasing concern is test equity. Given the consideration of being adaptive and efficient, examinees will finish the test at different time points as long as they satisfy the specific stopping rule, thereby it is hard to formulate the perceptions of fairness with such an uneven test-taking time (Choe et al., 2018).

Several attempts have been made to address these gaps, such as the Sympson and Hetter (1985) procedure (SH) and the "MFC per unit of time" framework (timed-MFC; Sie et al., 2015). The SH procedure, however, only focuses on preventing items from being overexposed but fails to increase the exposure of underexposed items. Also, though the timed-MFC efficiently reduces testing times, the testing times among examinees remain varied. Thus, the objectives of this research are to: (1) propose the stage adaptive item selection design (SAI) that makes the current need for decision making compatible with the percentile rank of item information; (2) optimize the timed-MFC and put forward the timed-SAI method.

### Study 1

To display the nature of the SAI design and investigate its performance, four factors were considered: item selection designs (random selection method, MFC, and SAI), examinees' abilities (29 levels, evenly spaced from -3.5 to 3.5 by 0.25), and test length (10 and 20 items, representing an extremely short test length and a common minimum test length in variable-length tests, respectively), and weighting parameter ($w$ = 0.50, 0.75, 1.00, and 1.25). Then, a sample of $N = 2{,}900$ was generated, with 100 examinees at each ability level. The item pool consisted of 500 items and the cut score was set to 0. The entire simulation process was replicated 100 times. Selected priority index, item exposure rate, and percentage of correct classifications (PCC) were investigated.

Due to space limitations, only representative results are shown here. From Figure 1, there was a clear trend of decreasing item exposure rate for informative items under the SAI design.

## Study 2

Study 2 focused on ascertaining the potential of the two new timed designs (i.e., the modified timed-MFC and the timed-SAI designs) on test time controlling compared with the traditional MFC method. $N$ (0, 1) rather than a discrete distribution was used to generate 1000 examinees to evaluate the performance of three item selection methods in a more realistic case. The hierarchical model was used to model the response and response time (van der Linden, 2007). Following Fan et al. (2012), medium correlations were set for ability $\theta$ and speed $\tau$ ($\rho_{\theta\tau} = 0.50$) as well as item difficulty $b$ and time density $\beta$ ($\rho_{b\beta} = 0.25$). The weighting parameter $w$ was set to 1 for SAI to balance the classification accuracy and equalization of exposure rates. The stopping rule was the Sequential Probability Ratio Test (SPRT) and the Type I and Type II error were set at 0.05. The width of the indifference region was 0.4. The centering parameter $v$ was created by sequencing from 0 to $e^{\mu_\beta - \mu_\tau} \approx 54.6$ with given length 30 and appending five points with the same interval to expand the tail, which produced 35 evenly spaced levels in total. Three levels of cut score were investigated (-1, 0, and 1). The remaining conditions were identical in study 1.

The test-taking time, PCC, average test length (ATL) and the $\chi^2$ for variable-length testing (Chang & Ying, 1999; Wen et al., 2000) were calculated over 100 repetitions. Also, the sensitivity and specificity were computed for cut score at -1 and 1. As shown in Figures 2-4, the timed-SAI method effectively reduced the mean and variance of test time.

## Practical Implications

Test security and test equity are as longstanding as testing itself. As the item selection method determines which item will be administered and which one will not, it is imperative for the test developers to incorporate the practical needs into the item selection process. In this spirit, the present research enhances the item selection method in two points, that is, assigning items adaptive to the need at the current stage and controlling the response time across all examinees. The promising simulation results show that the new item selection designs can counterbalance the item usage and shrink the deviation and cost of test-taking time while negligibly sacrificing the classification accuracy. Moreover, it is encouraging that the new methods lessen the necessity that an ideal item bank for CCT should consist of items closest to the cut-point.

## References


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

Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*, 135-158.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655-670.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement*, *39*, 389-405.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedure for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405-414.

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.

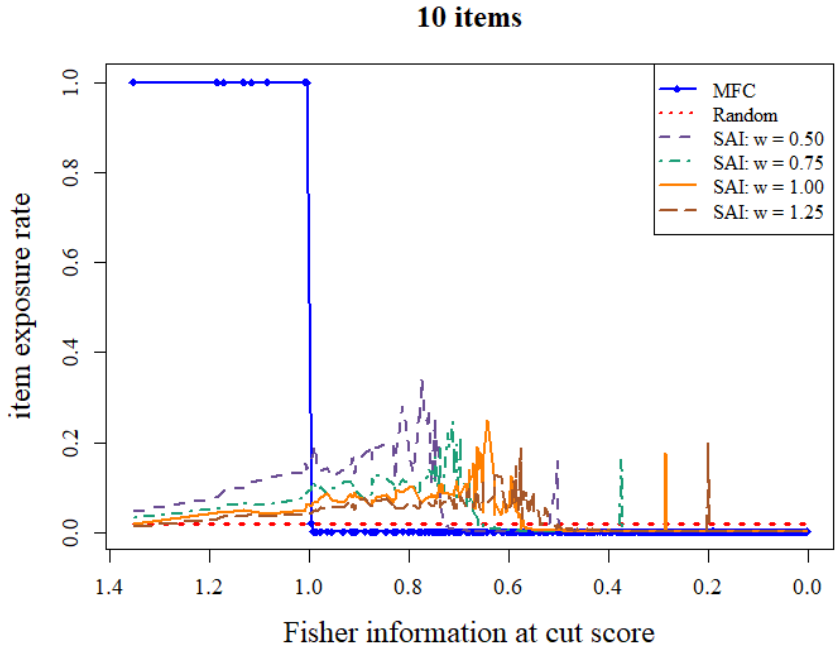
Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*, 778-793.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

Wen, J. B., Chang, H. H., & Hau, K. T. (2000, April). *Adaptation of a-stratified method in variable length computerized adaptive testing*. Paper presented at the American Educational Research Association Annual Meeting, Seattle, WA.

## Figures

(a)




(b)

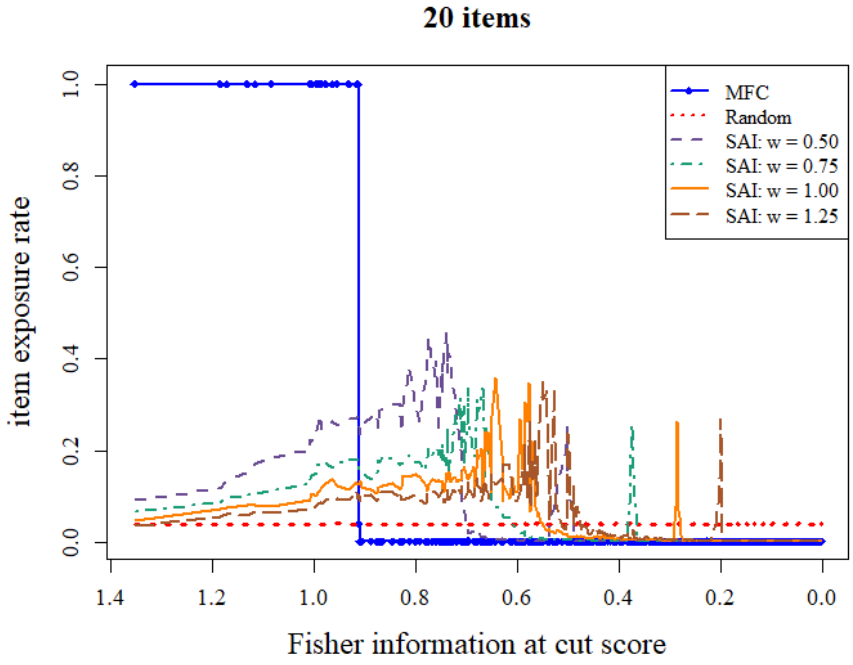



Figure 1. The item exposure rate comparison between three item selection designs.

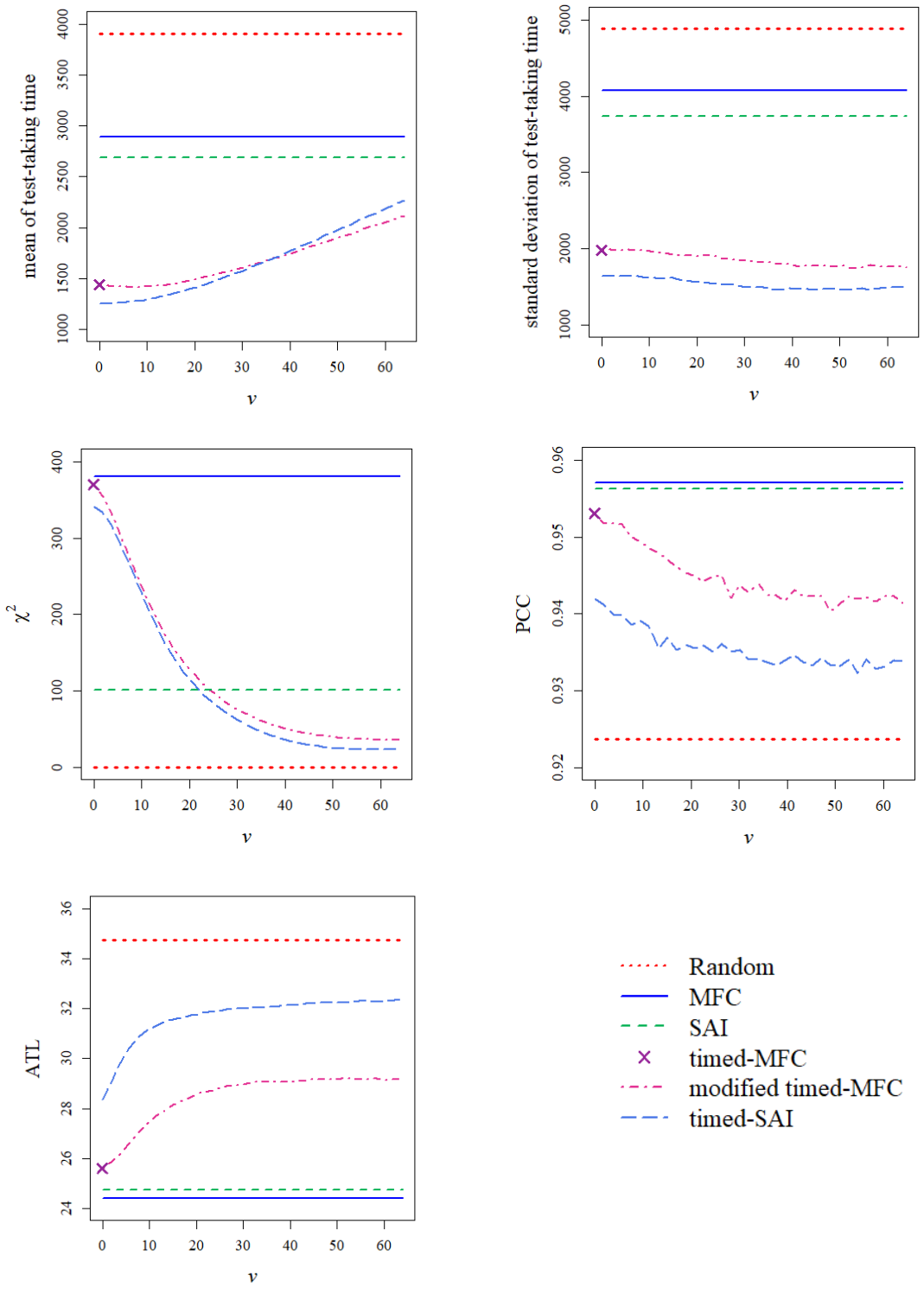



Figure 2. Evaluation indicators for six item selection methods with different centering parameter *v* values (cut score = 0). MFC = maximum Fisher information at cut score, SAI = stage adaptive method, PCC = percentage of correct classifications, ATL = average test length.

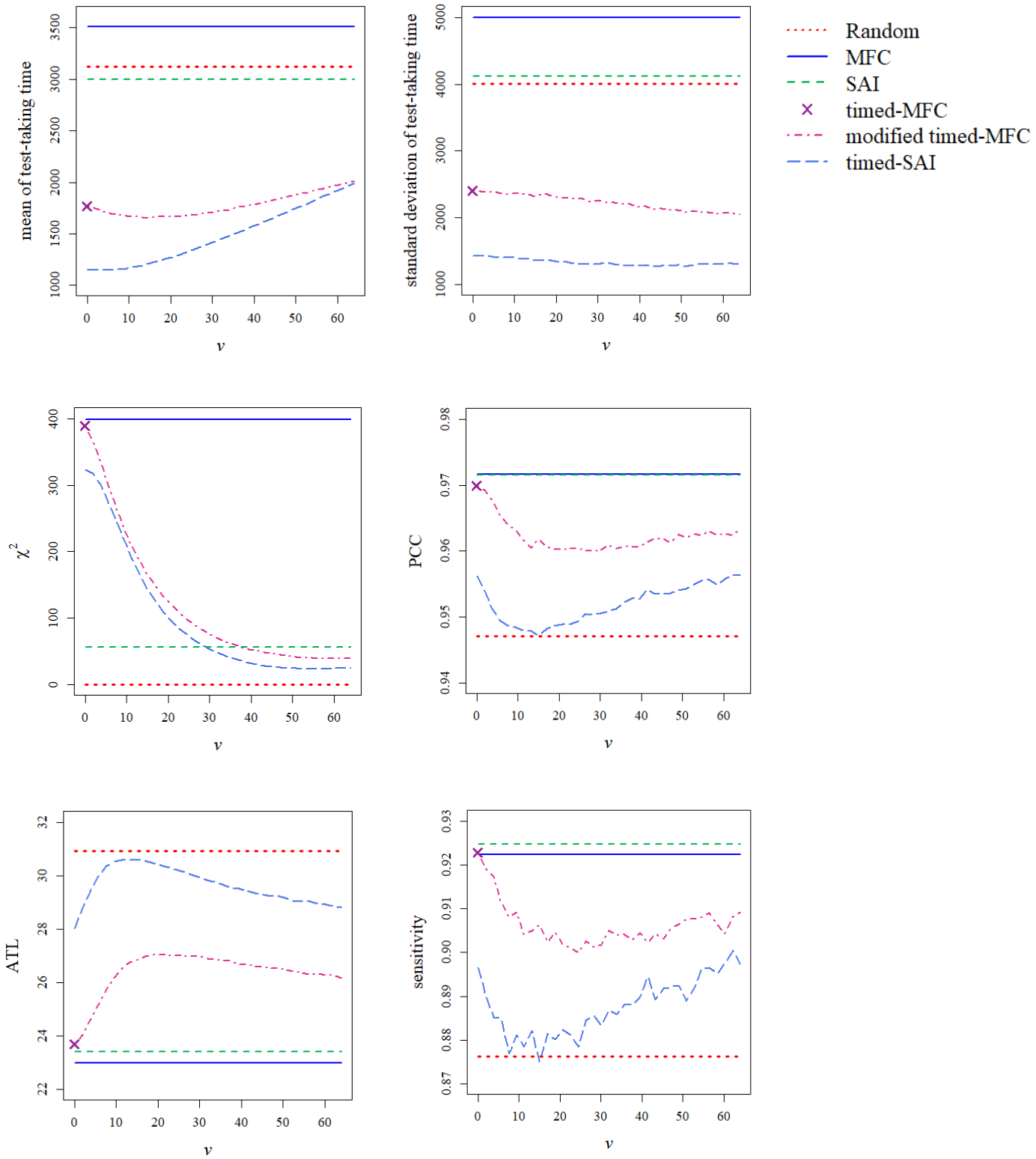



Figure 3. Evaluation indicators for six item selection methods with different centering parameter *v* values (cut score = 1). MFC = maximum Fisher information at cut score, SAI = stage adaptive method, PCC = percentage of correct classifications, ATL = average test length.

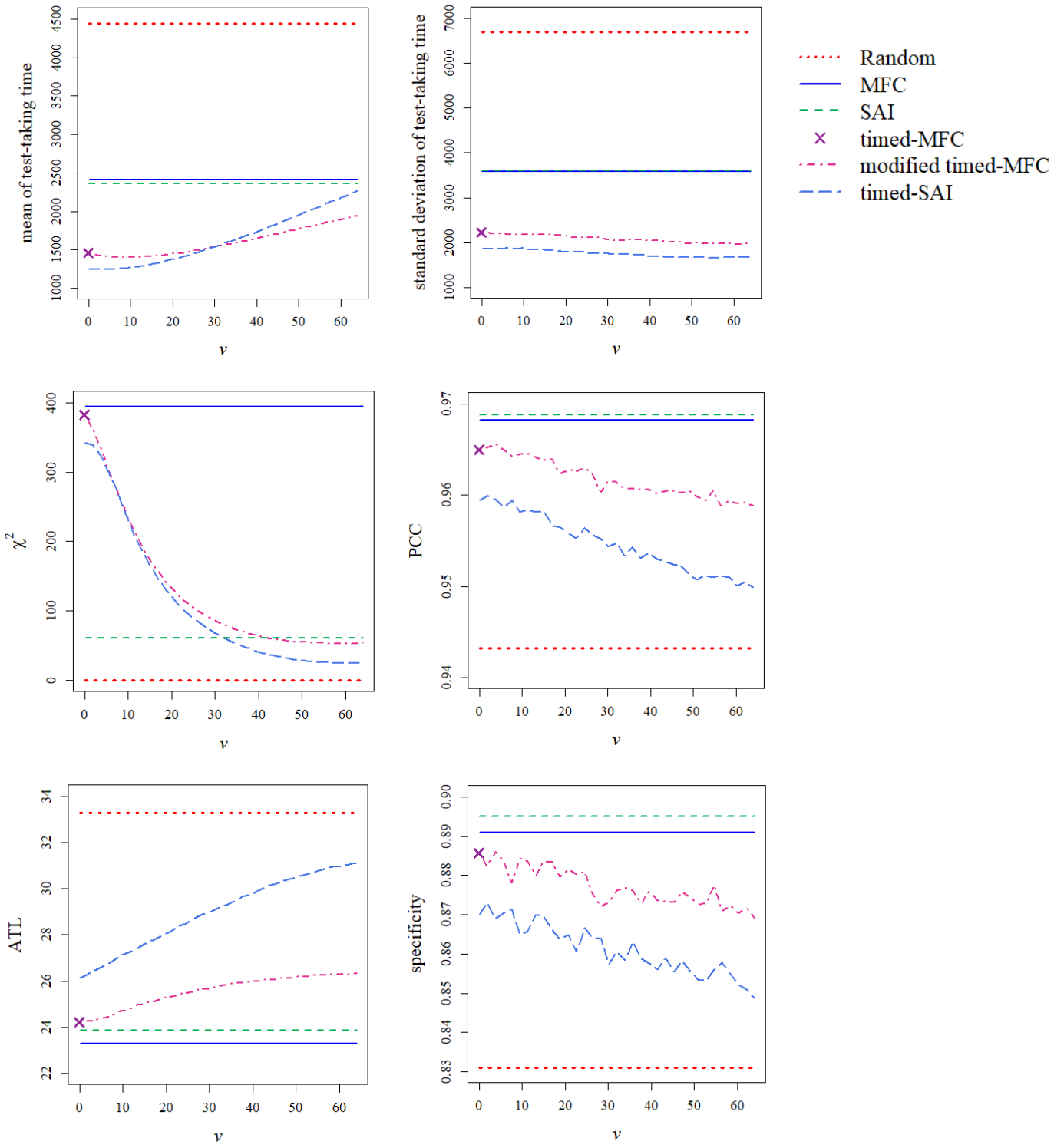



Figure 4. Evaluation indicators for six item selection methods with different centering parameter *v* values (cut score = -1). MFC = maximum Fisher information at cut score, SAI = stage adaptive method, PCC = percentage of correct classifications, ATL = average test length.