Research Report
2018-2

Comparison of Algorithms that Allow Item Review in Computerized Adaptive Testing

ZHONGMIN CUI, PHD. CHUNYAN LIU, PHD, YONG HE, PHD, HANWEI CHEN, PHD

"The practice of item review in CAT is hindered by the potential danger of cheating when examinees use particular cheating strategies. Legitimate item review, however, is desirable. What win-win methods are promising?"

AUTHORS

ZHONGMIN CUI, PHD

Zhongmin Cui is a principal psychometrician in Psychometric Research at ACT specializing in educational measurement and statistics, computer programming, and innovation.

CHUNYAN LIU, PHD

Chunyan Liu is a senior psychometrician in Psychometrics and Data Analysis at the National Board of Medical Examiners specializing in test equating, scaling, and linking.

YONG HE, PHD

Yong He is a psychometrician II in the Research division at ACT specializing in test equating and statistical analysis

HANWEI CHEN, PHD

Hanwei Chen is a principal psychometrician in Psychometric Research at ACT specializing in test equating, statistical analyses, psychometric supports for process improvement and pool managements.

ACKNOWLEDGEMENT

The authors thank Meichu Fan, Xiaohong Gao, Kurt Burkum, and Emily Neff for their review and comments on earlier drafts of this report.

SUMMARY

The practice of item review in CAT is hindered by the potential danger of cheating when examinees use particular cheating strategies. Examinees might purposely answer all items incorrectly when item review was allowed so that they had only easy items administered to them. As a result, they could get significantly inflated scores after changing their answers in review. Examinees may inflate their scores by guessing item difficulty. Using this strategy, an examinee might judge the correctness of an item based on the difficulty of its following item and decide whether to revise the answer to the current item. Even for examinees who do not know either of the aforementioned strategies, they could just randomly guess each answer resulting in underestimated interim abilities and thus easier items being selected by a typical CAT algorithm. After changing answers in review, they could get overestimated ability scores. Due to these dangers, item review was not permitted in most CAT

Item review, however, is desirable. Vispoel, Rocklin, and Wang (1994) reported that examinees tended to view the restriction on item review and answer changes as one of the main disadvantages of CAT. Bowles and Pommerich (2001) also found that examinees preferred to have the option of reviewing and changing their answers on adaptive tests. Because examinees feel that they have little control over the testing environments when no review is allowed, they tend to have elevated test anxiety levels that can increase the error in the examinee's ability estimation on adaptive tests.

SO WHAT?

The results of this study found both the rearrangement method and the block review method were effective in minimizing the illegitimate score gain, while examinees gained very little by using the Kingsbury strategy.

NOW WHAT?

The present study shows that the block review method is promising in a win-win situation - examinees had freedom to review within blocks, and the impacts of cheating strategies were controlled. We have not reached the end of the tunnel, however. Although the score gain using the Kingsbury strategy was small, examinees still could noticeably inflate their scores. And, as long as there are blocks, examinees taking rCAT with the block review method cannot enjoy the same benefit of reviewing items as those taking PnP because, with the block review method, there is no way to turn back to a previous block. A better method in providing review opportunities is still wanted.

Abstract

This paper considers item review algorithms in computerized adaptive testing (CAT). The research literature has shown that allowing item review in an educational test could result in more accurate estimates of examinees' abilities. The practice of item review in CAT, however, is hindered by the potential danger of cheating strategies. To provide review opportunities to examinees while minimizing the effect of cheating strategies, researchers have proposed different algorithms to implement CAT with restricted revision options. In this paper, we conducted a simulation study to evaluate these item review algorithms in two ways: 1) the accuracy on ability estimates and 2) the robustness against the cheating strategies. Ten thousand random samples were simulated from population distributions, and the correlation, average conditional standard error of measurement (CSEM), root mean square error (RMSE), and bias statistics were calculated. The block review method seems to be promising for use in practice, although limitations and cautions are discussed.

Comparison of Algorithms that Allow Item Review in Computerized Adaptive Testing

Zhongmin Cui, PhD, Chunyan Liu, PhD, Yong He, PhD, and Hanwei Chen, PhD

Introduction

Computerized adaptive testing (CAT) is gaining popularity among test developers. For example, the Smarter Balanced Assessment Consortium (SBAC), one of the two consortia of states commissioned by the US Department of Education, is developing CAT for assessing the Common Core State Standards (SBAC, 2014). Compared to paper and pencil (PnP) testing, CAT is more efficient, secure, and accurate thanks to its adaptive algorithm that tailors test questions according to each examinee's ability (SBAC, 2014; Wainer, 2000). The adaptability of CAT, however, puts CAT in an unfavorable position as far as item review is concerned.

The practice of item review in CAT is hindered by the potential danger of cheating when examinees use particular cheating strategies. For example, Wainer (1993) indicated that examinees might purposely answer all items incorrectly when item review was allowed so that they had only easy items administered to them. As a result, they could get significantly inflated scores after changing their answers in review. Kingsbury (1996) described another strategy that examinees could employ to inflate their scores by guessing item difficulty. Using this strategy, an examinee might judge the correctness of an item based on the difficulty of its following item and decide whether to revise the answer to the current item. Even for examinees who do not know either of the aforementioned strategies, they could just randomly guess each answer resulting in underestimated interim abilities and thus easier items being selected by a typical CAT algorithm. After changing answers in review, they could get overestimated ability scores. Due to these dangers, item review was not permitted in most CAT (Vispoel, Hendrickson, & Bleiler, 2000).

Item review, however, is desirable. Vispoel, Rocklin, and Wang (1994) reported that examinees tended to view the restriction on item review and answer changes as one of the main disadvantages of CAT. Bowles and Pommerich (2001) also found that examinees preferred to have the option of reviewing and changing their answers on adaptive tests. Because examinees feel that they have little control over the testing environments when no review is allowed, they tend to have elevated test anxiety levels that can increase the error in the examinee's ability estimation on adaptive tests (Stocking, 1997; Wise, 1996). The research in the literature has shown that, if answer changes are allowed on a test, the final estimate of an examinee's ability could be more accurate thanks to reduced anxiety level and the opportunity to fix mistakes (Olea, Revuelta, Ximénez, & Abad, 2000; Papanastasiou, 2005; Wise, 1996).

To provide examinees review opportunities while minimizing the effect of aforementioned testing strategies, researchers have studied different means to implement reviewable CAT (rCAT) but with restricted revision options. Stocking (1997) proposed three models allowing limited item review and response change: 1) test takers were allowed to change their responses at the end of the test subject to a maximum number of revisions; 2) test takers were allowed to revise their responses freely, but only within each section; and 3) test takers were allowed to revise responses only within each item set associated with the common stimulus. Later, Papanastasiou (2005) evaluated a ‘‘rearrangement procedure’’ that rearranges and skips certain items with answer changes in order to better estimate the examinees' abilities. More recently, Han (2013) proposed an "Item Pocket" method that examinees could skip answering items by putting them in a parking space and then go back to confirm the answer later.

No research, however, has been found to compare these algorithms in terms of accuracy on ability estimates and robustness against the aforementioned cheating strategies. We attempt to fill this gap in the literature in the present study.

Method

Data

The item pool used in this study consisted of more than 1,000 items calibrated from operational testing data using the three-parameter logistic model (3PL).

The mean and standard deviation of the *b* parameter were rescaled to be zero and one respectively, matching the mean and standard deviation of the examinees' ability distribution. Ten thousand examinees were simulated from a normal distribution to take a fixed-length CAT (30 items) with different cheating strategies (normal test-taking, Wainer strategy, Kingsbury strategy, and random guessing) that are described in detail in a following subsection. Item responses from the simulation were used as the data for analyses.

Implementation of Reviewable CAT

In addition to non-reviewable CAT which served as the baseline, we implemented three rCAT methods to provide review opportunities to examinees: the block review method, the rearrangement method, and the no restriction method. We did not include the Item Pocket method in the study because this method is more answer delaying than answer changing and does not provide realistic opportunities for examinees to change answers, which is the main purpose for reviewing.

The block review method. This method was also referred to as Model 2 by Stocking (1997). We did not include the other two models proposed by Stocking because Model 1 was found not robust against the Wainer strategy and Model 3 did not allow for reviewing discrete items. With the block review method, we simulated three block conditions: two, three, and six blocks with an equal number of items in each block. For example, in the six-block condition, each block consisted of five items. Examinees were allowed to review and change answers only within a block. Once an examinee started the next block or submitted the whole test, he or she was not allowed to go back to the finished block(s).

The rearrangement method. This method was proposed by Papanastasiou (2005). In this method, examinees were allowed to review and change answers after finishing answering all test items. Examinees could review all test items but could only change the answers of no more than a certain number of items (for example, five). To account for the answer changes, the final ability estimates were computed after skipping and rearranging some items. An item was skipped when the item score on a preceding item was changed because the skipped item might not match the new interim ability estimate in terms of difficulty and information. Papanastasiou (2005) found that removing a small number of less appropriate items yielded better ability estimates than including them. In particular, the following rules were followed in implementing the rearrangement method:

1. When the response to an item is changed from correct to wrong, skip the subsequent item(s) until
    a. an item with a correct response is found,
    b. three items have already been skipped,
    c. the end of test is reached,
    d. the maximum number of skipped items for the whole test is reached (this was set to be five in this study).

    We referred to the item that follows immediately after skipped item(s) as the *resume* item.

2. When the response to an item is changed from wrong to correct, skip the subsequent items until
    a. an item with a wrong response is found,
    b. three items have already been skipped,
    c. the end of test is reached,
    d. the maximum number of skipped items for the whole test is reached (this was set to be five in this study).

    Similarly, the item that follows immediately after skipped item(s) was also referred to as the *resume* item.

3. Interim abilities are re-calculated after each answer change. If information of a previously skipped item is larger than that of the *resume* item, the previously skipped item is inserted before the resume item.

The no restriction method. In this method, examinees were free to review and change the answer to any test item. This method was included to show the impacts of cheating strategies on score accuracy if rCAT is administered without any restriction.

For all these methods, we used the *expect a prior* (EAP) method to estimate interim abilities and the *maximum likelihood estimation* (MLE) method to estimate final abilities. The ability estimates for perfect scores and zero scores were set as 3.5 and -3.5,

respectively. We ignored content balancing during simulation. It is reasonable to believe that the content do not affect cheating strategies and thus the final conclusion of this study. After an interim ability was estimated, the next item was randomly selected from the most informative 25 items at the interim ability. The maximum exposure rate was set to be 0.2.

Cheating Strategies

In addition to the normal test-taking conditions, we simulated examinees using three cheating strategies: the Wainer strategy, the Kingsbury strategy, and the random strategy. Because different rCAT methods had different restrictions on review, examinees needed to adapt their cheating strategies to get the maximum benefit. For example, the rearrangement method used in this study allowed five answer changes. For a test-taker who applies the Wainer strategy, it is more realistic for the examinee to just purposely answer five items incorrectly rather than all items. For this reason, we simulated each testing taking strategy in a way that, we believe, would award test takers the most in each rCAT implementation.

The Wainer strategy. When the rCAT was implemented without any restrictions, examines were simulated to purposely answer each item incorrectly, resulting in easier items being administered. At the end of the test, examinees went back to the beginning of the test and changed answers according to their simulated abilities. We assumed that the probabilities for an examinee to purposely answer an item incorrectly and answer this item correctly were equal. We computed the probabilities using a 3PL model given item parameter values and simulated abilities.

With the block review method, examinees applied this strategy within each block. That is, examines were simulated to purposely answer each item incorrectly in a block and revised the answers according to their true abilities before moving to the next block.

With the rearrangement method, examinees were simulated to answer incorrectly for the first five items, resulting in easier items being administered. Examinees then revised the answers according to their true abilities. The reason for choosing the first five items instead of other places in the test was that more items with incorrect answers in the beginning of a test would result in an easier test which, we assumed, would benefit these examinees the most.

The Kingsbury strategy. When the rCAT was implemented without any restriction, an examinee was simulated to judge the difficulty of the current item and to revise the answer to the previous item if he or she found the current item was easier. We assumed that an examinee would revise the answer to the previous item if the *b*-parameter value for the current item was smaller than that for the previous item by at least 0.5. Because the examinee would answer the previous item with one less option (one option was eliminated due to the clue from the difference in difficulty), we assumed the *c* parameter of this item would be increased by 1/20 (i.e., the difference between the probability of a correct answer through random guessing on a five-option item and on a four-option item).

With the block review method, the preceding procedure was applied within each block. Due to the design of the block review method, examinees were unable to change the last item in each block because the Kingsbury strategy does not allow examinees to go back once a new block is started.

The Kingsbury strategy was not directly applicable when the rCAT was implemented with the rearrangement method because no answer change was allowed until all test items were administered. Examinees, however, could use the information on the difference in difficulties to revise answers during the review process. The same procedure as described in the preceding paragraphs was applied except that the number of item changes was limited to five due to the constraints employed by the rearrangement method.

The random strategy. Not every examinee knows the Wainer strategy or the Kingsbury strategy. For those who do not know either strategy, they could just randomly choose answers, resulting in underestimated abilities that would lead to easy items being administered. Upon reviewing, they can revise the answers according to their true abilities.

In each rCAT implementation, this strategy was simulated in the same way as the Wainer strategy except that examinees chose random answers instead of incorrect answers.

Criteria

We evaluated the final results of each method under each condition in terms of correlation, average conditional standard error of measurement (CSEM), root mean square error (RMSE), and bias. The correlation was computed between estimated and true abilities. The root mean square error was computed by

$$RMSE[\hat{\theta}] = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{N}},$$

and the bias by

$$Bias[\hat{\theta}] = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)}{N},$$

where N is the number of examinees and θ_i and $\hat{\theta}_i$ index each examinee's true and estimated ability, respectively.

The average CSEM was computed by

$$CSEM[\hat{\theta}] = \sqrt{\frac{\sum_{i=1}^{N}\frac{1}{I(\hat{\theta}_i \mid a,b,c)}}{N}},$$

where *a*, *b*, and *c* are item parameters of a 3PL IRT model. Because cheating strategies may not benefit every test taker the same, the results were also evaluated in high, middle, and low ability groups. We arbitrarily chose θ values of -1 and 1 to separate those groups.

Results

Table 1 shows the results of correlation, CSEM, RMSE, and Bias for each test-taking strategy under each rCAT method. When no review was allowed in the test, the correlation, CSEM, RMSE, and Bias were 0.98, 0.21, 0.22, and 0.00, respectively. These values served as the baseline in the comparison.

When review was allowed, the Kingsbury strategy yielded almost the same results as the baseline in terms of correlation, CSEM, and RMSE. The bias yielded by the Kingsbury strategy was slightly larger compared to the baseline, suggesting score inflation. This finding is consistent across all review conditions regardless of different restrictions on review. By contrast, the performance of the random strategy and the Wainer strategy varied with the change of restriction on review.

When review was allowed without any restriction, the correlation values yielded by the random strategy and the Wainer strategy were much smaller than the baseline as well as the Kingsbury strategy. At the same time, CSEM, RMSE, and bias all increased. The bias values yielded by the three strategies were mostly positive, indicating score inflation, although the magnitude of inflation yielded by the Kingsbury strategy was much smaller compared to the other two strategies. This finding is consistent with the literature in two ways: (1) cheating strategies could help examinees to inflate test scores, and (2) the Kingsbury strategy was not very effective in increasing examinee test scores (Vispoel, et al., 2002; Wise et al., 1999). A seemingly surprising finding shown in Table 1 is that the random strategy yielded larger bias than the Wainer strategy. This seems to suggest that there was no advantage for examinees to answer incorrectly on purpose than to just answer randomly. Randomly answering seemed to inflate scores more than the Wainer strategy when review was allowed without any restriction.

.

Table 1. Correlation, CSEM, RMSE, and Bias Yielded by Cheating Strategies

Review Method	Test-taking Strategy	Correlation	CSEM	RMSE	Bias
Not Allowed	Normal	0.98	0.21	0.22	0.00
	Kingsbury	0.98	0.21	0.23	0.05
No Restriction	Random	0.85	1.31	0.82	0.21
	Wainer	0.83	2.75	0.87	0.13
	Kingsbury	0.98	0.21	0.22	0.05
Rearrangement	Random	0.96	0.48	0.32	0.02
	Wainer	0.95	0.68	0.37	0.04
Block review					
	Kingsbury	0.98	0.21	0.22	0.05
2 Blocks	Random	0.94	0.39	0.36	0.01
	Wainer	0.94	0.40	0.36	-0.01
	Kingsbury	0.98	0.21	0.22	0.05
3 Blocks	Random	0.97	0.26	0.27	-0.01
	Wainer	0.96	0.28	0.28	-0.01
	Kingsbury	0.98	0.21	0.22	0.05
6 Blocks	Random	0.98	0.22	0.23	0.00
	Wainer	0.97	0.23	0.23	0.00

When the rearrangement method was employed to provide restricted review opportunities, all statistics yielded by the random strategy and the Wainer strategy tended to suggest the effectiveness of the rearrangement method in controlling score inflation, as shown in Table 1. In particular, the rearrangement method yielded larger correlation, and smaller CSEM, RMSE, and Bias values, compared to the condition when the review was provided without any restriction. Similarly, the block review method was also effective in controlling score inflation compared to the no restriction condition. Table 1 also shows that the block review method tended to perform better than the rearrangement method by yielding larger correlation values and smaller CSEM, RMSE, and Bias values, especially when the number of blocks was more than two. For the block review method with six blocks, all statistics yielded by the random strategy and the Wainer strategy were very similar to the baseline. This finding suggests that the block review method has the potential to provide review opportunities to examinees without sacrificing accuracy of ability estimation, although there is a danger that this method tended to perform poorly if the number of blocks was small.

Figure 1 shows the distribution of the difference between estimated and true abilities. Because the results for the random method and the Wainer method were similar, we only included results from the Wainer method. As can be seen from this figure, most difference values were within a range of (-1, 1) unit in ability scale for the baseline. When review was allowed without any restriction, the range of most difference values grew to (-4, 4), indicating a loss of score accuracy. There were more score inflations than score deflations under the condition of no restriction. When the rCAT was administered with the rearrangement method to control reviewing, the range of most difference values shrank to (-1, 2) compared to the no restriction condition. In addition, the rearrangement method yielded considerably less score inflations and deflations than the no restriction method.

The bottom right panel in Figure 1 shows the results for the rCAT with the block review method to control reviewing. As can be seen in this figure, it is hard to tell the difference between the block review method and the baseline in terms of the distribution of difference in abilities, indicating superb performance of the block review method. These findings are consistent with the results in Table 1 that (1) both the rearrangement method and the block review method were effective in controlling score inflation, and (2) the block review method performed better than the rearrangement method.

Figure 1. Distribution of the difference between estimated and true abilities

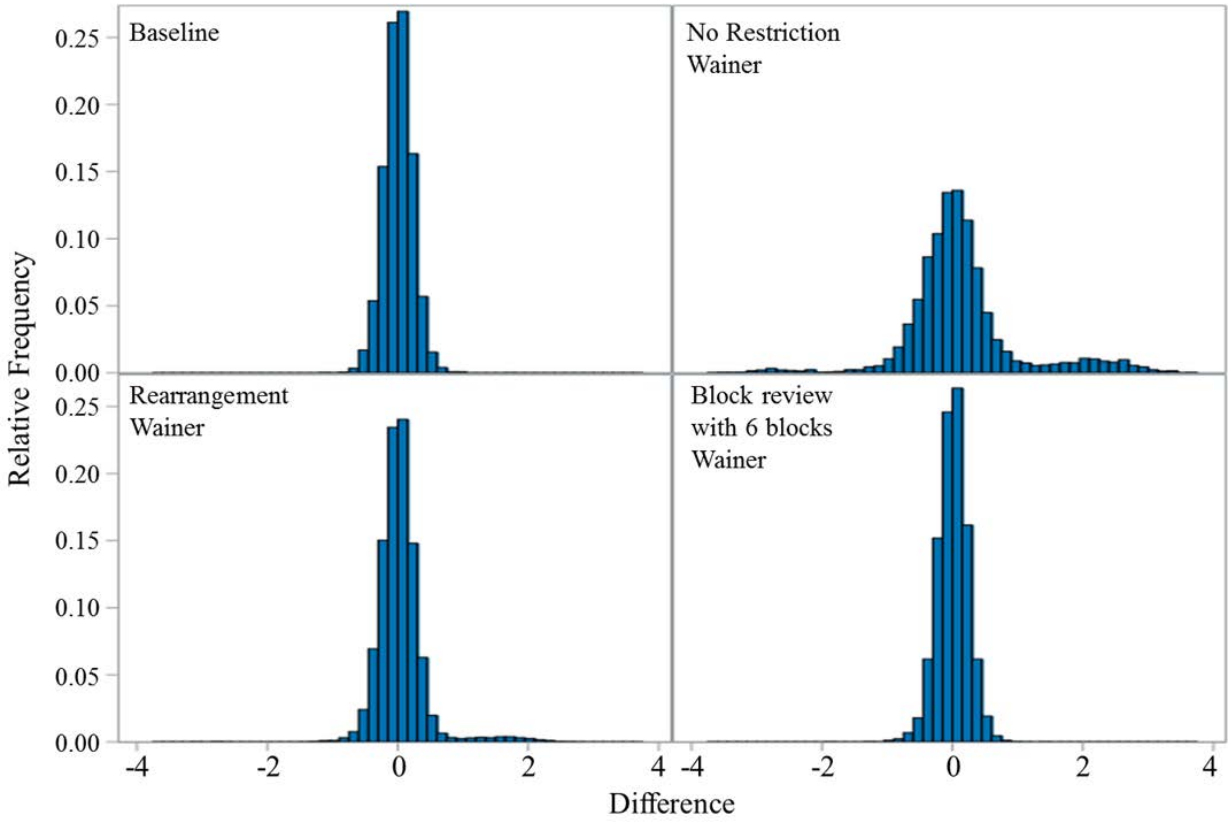
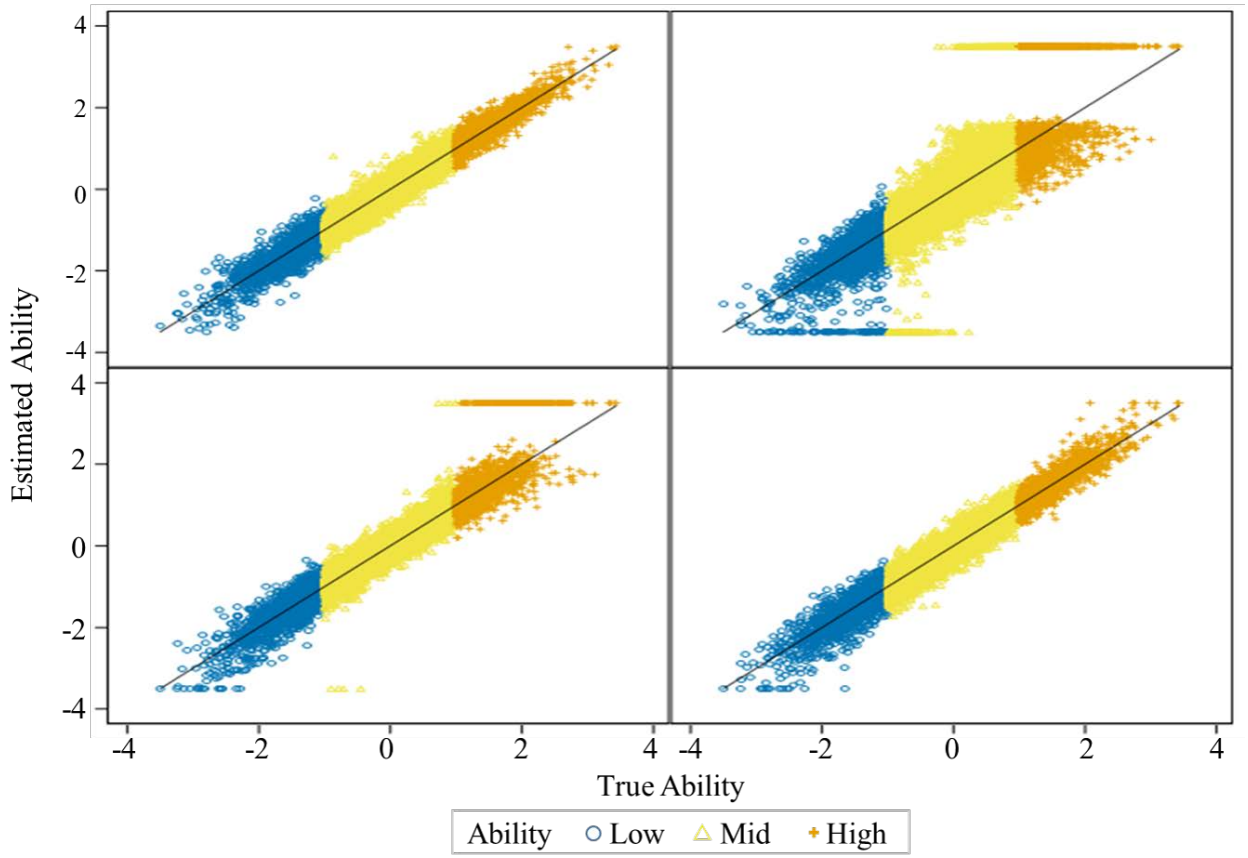


Figure 2 shows the scatter plots of the estimated abilities and the true abilities separated by ability groups. This figure enables us to see the impact of different rCAT methods on different ability groups. Because the results for the random method and the Wainer method were similar, we only included results from the Wainer method. The top left panel shows the results for the baseline, where pairs of ability values scattered around the 45 degree line, which is the true relationship between the estimated ability and the true ability if there is no estimation error. When review was allowed without any restriction, as shown in the top right panel, the data points tended to drift away from the true relationship, indicating a loss of accuracy in estimating the abilities. Figure 2 also shows that high ability examinees tended to get inflated scores when the Wainer strategy was used in rCAT without restriction and low ability examinees deflated scores. Compared to the baseline, more high ability examinees received perfect scores and more low ability examinees received the lowest possible scores. This finding suggests that high ability examinees tended to benefit from using the Wainer strategy while low ability examinees tended to be disadvantaged from using it. It should be noted that a few high ability examinees received deflated scores, possibly due to incorrect responses to easier items. This seems to suggest that, besides the likely benefit of inflating scores, there is also a risk for high ability examinees to receive lower scores by using the Wainer strategy.

Figure 2. Scatter plot of the estimated and the true abilities separated by ability groups



When the rCAT was administered with the rearrangement method to control reviewing, as shown in the bottom left panel, the data points tended to move back to the true relationship shown in the top right panel. Although we can still see a few extremely inflated and deflated scores, the rearrangement method yielded considerably less score inflations and deflations than the no restriction method. This finding suggests that the rearrangement method was somewhat effective in reducing the estimation error, caused by the usage of the Wainer strategy in unrestricted rCAT.

The bottom right panel in Figure 2 shows the results for the rCAT with the block review method to control reviewing. As can be seen in this figure, it is hard to tell the difference between the block review method and the baseline regardless of ability group. Being consistent with previous findings, this finding suggests that the block review method performed better than the rearrangement method.

Discussion and Conclusion

The results from this study confirmed the impact of test manipulation strategies on measuring examinees' abilities if rCAT is administered without any restriction, although the impacts of different strategies were not the same. We found that examinees gained very little by using the Kingsbury strategy regardless of how the rCAT was administered. One possible reason may be that examinees could not easily tell the difference in difficulties. This is true especially in the late stage of CAT where the difference in difficulty between two adjacent items is typically small. In their study to examine the effectiveness of the Kingsbury strategy, Vispoel, Clough, Bleiler, Hendrickson, and Ihrig (2002) found that examinees were not very successful in distinguishing the difficulty difference within each item pair. Another possible reason may be that, even after successfully distinguishing the difference in item difficulty, examinees couldn't always change the answer to a correct one. For an item with five options, an examinee can only gain .05 (i.e., the difference between 1/4 and 1/5) in the probability of randomly guessing the right answer. In practice, the Kingsbury strategy may perform better than what we found in the simulation study because examinees with

partial knowledge can upfront eliminate obviously incorrect options. For example, if an examinee could eliminate an obviously incorrect option, the gain in the probability would be .08 (i.e., the difference between 1/3 and 1/4).

Compared to the Kingsbury strategy, the Wainer strategy was found to be effective in inflating scores for some examinees if the rCAT was administered without any restriction. This finding is consistent with the results from other studies in the literature (Bowles & Pommerich, 2001; Gershon & Bergstrom, 1995; Stocking, 1997; Vispoel et al., 1999; Wang & Wingersky, 1992). In this study, we also found that not all examinees could benefit from using the Wainer strategy equally. In particular, high ability examinees tended to gain scores while low ability examinees tended to lose scores, which seems to suggest that low ability examinees should not use the Wainer strategy to avoid potential score loss.

We included the random strategy because (1) not all examinees know either the Kingsbury strategy or the Wainer strategy, and (2) the random strategy is much simpler for examinees to use than the other two. The results of this study show that the random strategy was more effective than the Kingsbury strategy and as effective as the Wainer strategy. In fact, from the examinees' standpoint, the random strategy may be better than the Wainer strategy because there is no need to spend time trying to find the correct answer and then purposely choosing the wrong one. If an examinee tries to use the Wainer strategy seriously, he/she may run out of time after purposely answering all item incorrectly and leave no time to change answers. By contrast, it does not take much time to randomly choose answers. Thus, the random strategy is more practical than the Wainer strategy for examinees who want to gain scores illegitimately, although it is the test developer's job to minimize this gain, if not eliminate it completely.

In this study, we found both the rearrangement method and the block review method were effective in minimizing the illegitimate score gain. The performance of the block review method was even better. Especially when the rCAT was implemented with six blocks, the results were almost identical to the baseline (i.e., CAT with no review). By contrast, noticeable score distortion can still be seen with the rearrangement method. Although we might be able to get better results for the rearrangement method by tweaking the algorithm, we expect the chance for this method to perform better than the block review method is small. The rearrangement method only allows a small number of answer changes while the block review method allows changes to all items except the last item in each block. Meanwhile, the complexity of the algorithm of the rearrangement method may make it hard for developers to use this method in practice. We believe examinees will also prefer the block review method over the rearrangement method because of more freedom in changing answers.

In applying the block review method, test developers should be cautious in determining the number of blocks. If the number of blocks is too small, for example two, cheating strategies still have the potential to affect the results in a noticeable way as shown in Table 1. On the other hand, if the number of blocks is too large, the block size will be small. An extreme situation in this direction is when the number of blocks is as large as the number of items. In this case, it is equivalent to no review at all. It is obvious that the larger the number of blocks is, the less freedom examinees have in item reviewing. Another consideration is that the size of a block not only depends on the number of blocks but also depends on content. For example, it is logical to put items in the same block if they belong to the same passage.

As pointed out by Wise (1996), examinees want more control over taking a test. In response, the research literature shows efforts in providing review opportunities in CAT while minimizing the effect of cheating strategies. The present study shows that the block review method is promising in a win-win situation - examinees had freedom to review within blocks, and the impacts of cheating strategies were controlled. We have not reached the end of the tunnel, however. Although the score gain using the Kingsbury strategy was small, examinees still could noticeably inflate their scores. And, as long as there are blocks, examinees taking rCAT with the block review method cannot enjoy the same benefit of reviewing items as those taking PnP because, with the block review method, there is no way to turn back to a previous block. A better method in providing review opportunities is still wanted.

References

Bowles, R., & Pommerich, M. (2001). *An examination of item review on a CAT using the specific information item selection algorithm.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Gershon, R., & Bergstrom, B. (1995). *Does cheating on CAT pay: NOT!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC document reproduction service No. ED 392844).

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement,* 37(4), 259-275.

Olea, J., Revuelta, J., Ximénez, M. C., & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicologica,* 21, 157-173.

Papanastasiou, E. C. (2005). Item review and the rearrangement procedure: Its process and its results. *Educational Research and Evaluation,* 11 (4), 303-321.

SBAC (2014). *Computer Adaptive Testing.* http://www.smarterbalanced.org/assessments/testing-technology/ (retrieved on 5/8/2018)

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement,* 21, 129-142.

Vispoel, W. P., Clough, S. J., Bleiler, T. Hendrickson, A. B., & Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement,* 39(4), 311–330.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement,* 37, 21-38.

Vispoel, W. P., Rocklin, T., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computer-adaptive, and self-adaptive testing. *Applied Measurement in Education,* 7, 53-79.

Vispoel, W. P., Rocklin, T. R., Wang, R., & Bleiler, T. (1999). Can examinee use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement,* 36, 141-157.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice,* 12, 15-20.

Wainer, H. (2000). Introduction and History. In Wainer, H. (Ed.) *Computerized Adaptive Testing: A Primer.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, M., & Wingersky, M. (1992). *Incorporating post-administration item response revision into a CAT.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Wise, S. L., (1996). *A critical analysis of the argument for and against item review in computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC document reproduction service No. ED 400267).

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A., & Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education,* 12, 185–198

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.

ACT.org/research