A 'REARRANGEMENT PROCEDURE' FOR SCORING ADAPTIVE TESTS WITH REVIEW OPTIONS

Running head: Rearrangement procedure

Elena C. Papanastasiou
University of Kansas

Elena C. Papanastasiou
644 Joseph R. Pearson Hall
Psychology and Research in Education
The University of Kansas
Lawrence, KS 66045
USA

Tel. 785-864-9649
Fax: 785-864-3820
e-mail: papanast@ku.edu

Paper presented at the National Council of Measurement in Education,
April 2, 2002,New Orleans, LA

Abstract: Due to the increased popularity of computerized adaptive testing (CAT), many admissions tests, as well as certification and licensure exams have been transformed from their paper-and-pencil versions to computerized adaptive versions. A major difference between paper-and-pencil tests and CAT, from an examinee's point of view, is that in many cases examinees are not allowed to revise their answers on CAT. Examinees prefer item review since they can correct misread or miskeyed items, while some researchers, are afraid that examinees might try to use item review to cheat on the test.

The purpose of this study is to test the efficiency of a '*rearrangement procedure*' that rearranges and skips certain items in order to better estimate the examinees' abilities, without allowing them to cheat on the test. This was examined through a simulation study. The results show that the rearrangement procedure is effective in reducing the bias of the ability estimates. The reliability slightly decreased after the procedure, but this decrease was negligible.

Keywords: Item review, computer adaptive testing (CAT), answer changing, rearrangement procedure, Item Response Theory (IRT), bias, Maximum Likelihood estimation.

Acknowledgements: The author would like to gratefully thank Mark D. Reckase, Edward W. Wolfe, Richard Houang, Maria Teresa Tatto and Frederic Robin for their invaluable contributions to this study.

A 'REARRANGEMENT PROCEDURE' FOR SCORING ADAPTIVE TESTS WITH REVIEW OPTIONS

Computerized adaptive testing (CAT) has gained increased popularity during the last two decades (Reckase, 2000). Consequently, many tests such as admissions tests like the GRE, the SAT, and the TOEFL, have been transformed from their paper-and-pencil versions to computerized adaptive versions. Adaptive tests are now also being used for certification and licensure purposes (Stone & Lunz, 1994).

The popularity of CAT has prompted researchers, measurement specialists and psychometricians to reconceptualize many of the processes that were established for regular paper-and-pencil tests (Mills & Stocking, 1995; Pommerich & Burden, 2000; Reckase, 2000). For example, a major difference between paper-and-pencil tests and adaptive tests, from an examinee's point of view, is that in many cases, examinees are not allowed to revise their answers on CATs (Vispoel, Rocklin & Wang, 1994; Vispoel, 1998b; Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997a; Wise, 1997b). So item review can become a major concern for students who feel anxious while taking tests. This anxiety is especially problematic since anxiety can be an additional source of error in the examinee's ability estimates. So some researchers feel that the validity of a CAT can increase when examinees can revise their answers, since it allows them to rethink their answers and make corrections to items that might have been misread or miskeyed (Vispoel, 1998a). By allowing revisions, the final ability estimate could represent an examinee's ability more accurately because it will be closer to his/her actual ability when small mistakes such as miscodings are corrected (Wise, 1996; Vispoel, Henderickson & Bleiler, 2000).

Other researchers, however, believe that item review can decrease the efficiency and validity of a CAT since item review allows examinees to cheat on the test. An example of a cheating strategy is the Wainer strategy (Wainer, 1993), in which examinees might purposely answer all the items incorrectly when they are first administered so that they can have the easiest items administered to them. The second step of the Wainer strategy involves going back to the test items, and answering all of the items on the test correctly. Answering all the items correctly should not be very difficult for these examinees since the test would consist of very easy items that have low difficulty levels. This would result in an artificial inflation of the examinee's ability estimates.

Other studies on item review have also shown that the efficiency of a test decreases when item review is permitted (Stocking, 1997; Vispoel, Rocklin, Wang & Bleiler, 1999). Consequently, item review is not permitted in most adaptive tests at this time (Vispoel, Henderickson & Bleiler, 2000).

The purpose of this study is to test the effectiveness of a rearrangement procedure that permits examinees to review previously presented items without allowing them to artificially inflate their test scores by using test-wiseness strategies. More specifically, the research questions that will be answered in this study are the following:

1. What are the effects of the rearrangement procedure on the reliability of the estimates?

2. How much statistical bias does the rearrangement procedure create?

Significance Of The Study

The issue of item review is of great importance to examinees who are administered tests, as well as to testing organizations that administer tests. From the perspective of the examinees, tests are stressful situations overall, and even more so when they are high stakes tests. Therefore, examinees would like to have as much control of the testing situation as possible when they are taking tests, so that they can

perform to the maximum extent of their capabilities. Such control is achieved by allowing the examinees to use the test taking strategies that they have been accustomed to. So when examinees are permitted to review answers on a test, the majority of them choose to do so (Bowles & Pommerich, 2001; Wagner, Cook & Friedman, 1998).

When examinees are administered paper-and-pencil tests, each individual uses various strategies while completing the test (Vispoel, Hendrickson & Bleiler, 2000). For example, some examinees choose to go through the test once and answer all the questions immediately no matter how confident they are of their responses. After they answer all the questions, they go over the whole test again, they rethink all of their answers, and they might make any changes that are necessary to their original answers. Other examinees choose to omit questions that they are unsure of and go back to those items after they reached the end of the test (Stocking, 1997).

When taking computer adaptive tests, however, in most cases examinees are not allowed to go back and revise their answers. So the examinees who have been using the strategies mentioned above cannot use those anymore on computer adaptive tests. This may cause stress and anxiety to many examinees, especially if they are taking high stakes tests (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997a; Wise, 1997b). This is an additional source of anxiety since examinees tend to have higher anxiety when they are administered computer adaptive tests than with paper and pencil tests (Powers, 2001). This might cause even bigger problems and stress to international students who have to take high stakes admissions tests in a foreign language to get admitted to universities in the USA. Therefore, a large number of examinees may actually be at a disadvantage when taking computerized adaptive tests, because the no-revision policy might prevent them from performing to the maximum extent of their abilities (Vispoel, 1998a).

The stress that is caused by computer adaptive tests might even cause examinees to make mistakes on questions to which they know the answers. For example, due to stress, examinees might choose an incorrect option accidentally even though they knew the correct answer to a question (Lunz, Bergstrom & Wright, 1992). It is also possible that the stress can cause examinees to make foolish arithmetic errors although they have the ability and skills to answer them correctly. So if examinees lose points on tests due to such reasons, and if they are not allowed to go back and revise their answers, their test scores will not be valid indicators of their true abilities (Vispoel, Henderickson & Bleiler, 2000).

However, item review can also be costly to testing organizations that believe that the efficiency of their tests, as well as the validity of the test scores would be compromised if item review were permitted (Gershon & Bergstrom, 1995). For example, examinees might try to "trick" the computer to artificially inflate their test scores with item review (Wainer, 1993). In addition, when item review is permitted, examinees might have more time to memorize test items, which would jeopardize the security of these tests (Patsula & McLeod, 2000). Therefore, this study will attempt to provide a compromised solution to this problem. This would involve a solution that would allow examinees to revise their answers, without jeopardizing the quality and efficiency of the test.

Methods

The purpose of this study is to assess the effects of a specific ‘rearrangement procedure’ that rearranges and skips certain items in order to obtain a better estimate of the examinee’s ability. It is hypothesized that the rearrangement procedure will improve the ability estimates of the examinee's scores. It is also expected that the rearrangement procedure will have three additional advantages if it were used in real life. First, by using this rearrangement strategy, the estimated ability estimates of the examinees may become more valid since the examinees would have a chance to correct any errors or miscodings that they might have made. In addition, the rearrangement procedure will also help reduce the stress of examinees

because they will have more control over the testing situation when they can revise their answers. Finally, the third advantage of the revision strategy that is essential for testing organizations, is that it will not permit the Wainer strategy from taking place when review is permitted.

To examine the rearrangement procedure, a simulation study was conducted to determine the effect that the rearrangement procedure would have on the accuracy of the examinee's ability estimates. All the simulation procedures for the no-review adaptive testing process, were performed using the Computer-Based Testing Simulation and Analyses Computer Program (CBTS) (Robin, 1999). The simulation of the rearrangement and the item review process was conducted using SAS (SAS Institute Inc, 1999).

Simulation specifications

In order to determine the specifications for this simulation study, the adaptive testing literature was reviewed to make the simulation as realistic as possible (Ban, Wang, Yi & Harris, 2000; Camilli & Penfield, 1997; Camilli, Wang & Fesq, 1995; Eignor, Stocking, Way & Steffen, 1993; Patsula & McLeod, 2000; Stocking, 1997; Vispoel, 1998a; Wang & Vispoel, 1998). So a 30 item, fixed length adaptive test was administered to each examinee in the simulation since a 30 item test would be sufficient to properly estimate the examinee's abilities (McBride, Wetzel & Hetter, 1997). The tests were then created from an item pool that included 250 items.

The psychometric literature was also reviewed to determine the item pool characteristics. The real test items that were used as the basic reference for the creation of this item pool, were obtained from items used in the following references; Wang and Vispoel, (1998); Luecht and Hirsch, (1992); Camilli and Penfield, (1997); Vispoel, (1998a) and Ban, Wang, Yi and Harris, (2000). From the items that were included in these references, the means and standard deviations were obtained for each of the three parameters, to serve as a model for the item parameters of the simulated item pool.

Table 1 describes the targeted distributional characteristics of the item pool created for this simulation. In terms of the distributions of the item parameters, the *a*-parameter, which is the index of discrimination, usually has a log normal distribution. So, the distribution of the a-parameter that was created for this study was a log normal distribution with a mean of 1.10 and a standard deviation of 0.25. The values of the a-parameter were also restricted to range between 0.45 and 2.3. The *b*-parameter, which is the difficulty index, had a uniform distribution. The reason for the use of this distribution was to have an adequate amount of items to assess the ability levels of all the examinees. Although most item pools do not have a uniform distribution of b-parameters, the ideal goal for test developers is to achieve this distribution. The mean for the b-parameter was 0.00 with a standard deviation of 2.0. The values of the b-parameters for the uniform distribution ranged from -3.5 to 3.5. The *c*-parameter, the pseudo-guessing parameter, also had a uniform distribution in this simulation. This is consistent with many studies on adaptive testing (Harwell, Stone, Hsu & Kirisci, 1996; Luecht & Hirsch, 1992). The mean of the c-parameter was 0.17 with a standard deviation of 0.10. Finally, the values of the c-parameter distribution ranged from 0.0 to 0.35. The range of values for all of the distributions of the parameters were chosen to represent the values of the parameters that currently exist in the adaptive testing literature (Eignor, Stocking, Way & Steffen, 1993; Harwell, Stone, Hsu & Kirisci, 1996; Luecht & Hirsch, 1992; Wang & Vispoel, 1998). Items whose upper and lower bounds fell outside of the pre-specified range, were eliminated from the item pools.

The items that were administered in the simulation, were selected from the item pools based on the maximum information procedure (McBride, Wetzel & Hetter, 1997). With the maximum information procedure, the items that are administered to the examinees are the ones that provide the maximum

information at each of the examinees' current ability estimates. To estimate the examinee abilities in the simulation, the Maximum Likelihood (ML) procedure (Lord, 1980) was used

Examinee characteristics and test-taking behaviors

A group of 26000 examinees was simulated for this study. These simulees were selected from 13 equally spaced θ levels so that the distribution of the ability estimates would be approximately normal. The θ level groupings ranged from -3.0, to 3.0, in equally spaced intervals of 0.5.

There were three types of questions that the examinees changed their answers to in the simulation study. The first type of questions were the ones where the examinees made 'stupid mistakes' such as calculation errors, even though they had the ability to answer those items correctly. In the simulation procedure those cases were the questions to which examinees had an 0.80 or higher probability of answering correctly, but were answered incorrectly. So all of those answers would be changed by the examinees from incorrect to correct answers.

The second type of questions that the examinees changed their answers to, were the ones that were very difficult for them. In the simulation procedure, those cases were identified by the questions to which the examinees only had a 0.33 or lower probability of answering correctly, but were answered correctly by chance. Therefore, their answers to those questions were changed from correct to incorrect answers with a probability of 1.0.

The third type of questions that the examinees could change, were the questions that were well matched to their true abilities. In this case, the examinees would be unsure of their answers to such questions, and might wish to reread and rethink their answers. These cases were identified by the questions to which the examinees had approximately a 0.50 probability (0.47- 0.53) of answering correctly. These examinees would have a 0.72 probability of answering the item correctly, and a 0.28 probability of answering them incorrectly. In the cases where examinees had more items that needed to be reviewed than the number of items that were permitted, then the items that would eventually be reviewed were randomly selected by the simulation procedure.

Item Revision Algorithm- The Rearrangement Procedure

The rearrangement procedure will not be visible from the perspective of the examinees. All that they will know is that they will be allowed to change up to 5 of their answers on the test. If this procedure were used in a real testing situation, no additional time would be provided for the examinees to change their answers. Only if the examinees finished answering all 30 items on the test before the end of the allotted time would they be allowed to review their answers.

So, if this were applied to a real testing situation, and if item review were permitted on a 30-item test, and the examinees finished answering all the items before the time limit expired, they would have the option to go back to review and possibly change any of their answers. The test would then officially terminate either at the end of the time limit, or after the examinees finished making up to five changes to their answers on the test, whichever came first. The rearrangement procedure will then take place only after the review has taken place.

A problem might arise (if the procedure was used in a real testing situation), in explaining to the examinees why some of their answers have not be considered for estimation of their final ability estimates because of the rearrangement procedure. So the examinees will have to be warned in advance that certain items will not be used for their ability estimates, just like certain sections of some tests are not used for that purpose either, since they include seed items.

Item Skipping In The Rearrangement Procedure

One of the strengths of adaptive testing, is that the items that are administered are selected to match the examinee's most recent ability estimate. However, with item review, after the answer to an item *i* is changed, the items that follow might no longer be as appropriate for estimating the examinee's ability estimate. Therefore, instead of administering items that are not as appropriate for a new ability level, the rearrangement procedure will skip these items. The rearrangement procedure will then try to find an item *i* +k, that is more appropriate for the posterior ability estimate. It is hypothesized that by administering fewer items that are better targeted to an examinee's ability estimate, the final ability estimate will be less biased and closer to the examinee's true ability level than when less appropriate items are administered. This is consistent with Reckase (1975) who found that the bias of the ability estimates tended to increase by administering extreme items that were not properly targeted to the examinee's ability levels.

Types Of Answer Changing And The Rearrangement Procedure

There are three types of answer changes that could be made by the examinees; changing responses a) from an incorrect to an incorrect response, b) from an incorrect to a correct response, and c) from a correct response to an incorrect response.

Type 1 change. Incorrect to incorrect changes

If an examinee changes an answer from an incorrect option to another incorrect option, then no changes need to be made to the ability estimation of the examinee, and the examinee will obtain the same score as they did before the review. In addition, no change will take place in terms of the accuracy of the standard error of the test.

Type 2 change. Incorrect to correct changes

The second type of answer change that examinees can make, is the change from an incorrect to a correct answer. If this change were made to item *i*, the ability estimation $\hat{\theta}_i$ will be changed to $\hat{\theta}_i'$. However, if this occurs, question *i* +1 would probably not be the most informative item for the ability $\hat{\theta}_i'$ since it would be easier and targeted at lower ability levels than $\hat{\theta}_i'$. This is a problem in adaptive testing, because it will cause the bias of the final ability estimate to increase (Reckase, 1975).

To solve this problem, the computer algorithm will skip question *i* +1 in the ability estimation procedure, since that would no longer be an appropriate item for that ability level. The algorithm of the rearrangement procedure will then jump to the first item X after question *i* +1 (e.g. item *i* +k, with 1<k<4) that was answered incorrectly if it was more difficult. It is hypothesized that this new item *i* +k would be more similar to the item that would have been administered after item *i*, if item *i* were answered correctly in the first place. So after the skipping of items *i* +1 through *i* +k-1, the rest of the test would remain the same, and no changes would be made to the test if no other answers were changed. So the next step would be to recalculate the ability estimate based on the rest of the items in the order that they were presented, until the end of the test. However, in this specific case, a total of 30-(k-1) items would be used to estimate the final ability level.

Figure 1 provides an example in which an incorrect-to correct change was made to item 2 of a test. In this case, question 3 was skipped since it was targeted at a lower ability level than $\hat{\theta}_2'$. So the algorithm

jumped to item 4 since that was the first more difficult item that was answered incorrectly, that came after item 3.

However, it is also possible for the rearrangement procedure to jump 2 or 3 items until it finds the next incorrect item. In case 3 items have been skipped and none of these answers are incorrect, then the 4th item after the answer-changed-item will be the next item that will be used for the estimation of the examinee's ability estimate.

It is also possible for the rearrangement procedure to skip 3 items until it finds the next correct item. In case 3 items have been skipped and none of these answers are correct, then the 4th item after the answer-changed-item will be the next item that will be used for the estimation of the examinee's ability estimate.

Type 3 change. Correct to incorrect changes

If an examinee decides to change another item on the test (e.g. item *I*), and the answer is changed from a correct answer to an incorrect answer, item *I* +1 would be ignored in the ability estimation procedure. The reason for ignoring that item is because item *I* would be targeted at a higher ability level than $\hat{\theta}_I$ so it would be more difficult. This would result in a larger standard error of the final ability estimate. Therefore, the computer would select item Y (e.g. item *I* +K where 1<k<4) if that was the first item after item *I* that was easier since it was answered correctly. So it is hypothesized that the ability estimation would be more accurate if items *I* +1 through *I* +K-1 were ignored from the estimation procedure, and item *I* +K was used after the item whose answer was changed. This is done because it is hypothesized that item *I* +K would be more similar to the item that could have been administered after item *I*, if item *I* was answered incorrectly in the first case. The next step would be to recalculate the ability estimate from the rest of the items in the order they were presented. In this case, a total of 30-(K-1) items would be used to estimate the final ability level.

Making Two Or More Answer Changes: Rearranging Items In The Rearrangement Procedure

Consider the case in which an examinee makes two changes in his/her response patterns. This examinee might change the response to item 2 (from an incorrect to a correct response), and the response to item 13 (from a correct to an incorrect response). When the first change takes place, the algorithm will follow the same procedure as in the type 2 change. So item 3 would be ignored in the estimation procedure, and item 4 would be selected if that were the first item (after item 2) that was answered incorrectly. When the examinee continues through the test and changes the response to item 13 from a correct to an incorrect response, the algorithm would make a comparison to determine which items to use next in the estimation procedure. This determination would be made from the information that is provided by a) item 16, which is the first item after item 13 that was answered correctly, and b) any items that had been skipped in the estimation procedures at previous steps in the algorithm, such as item 3. The item that would provide the most information out of the two at $\hat{\theta}_{13}$ ',would be selected as the item that would replace item 14 that was skipped by the algorithm. The next step would be to recalculate the ability estimate from the rest of the items in the order that they were presented, until the end of the test.

If item 3 was more informative than item 16 at the ability level $\hat{\theta}_{13}$ ', item 3 would be used after item 13 for the estimation of the examinee's ability estimate. So the rearranged order in which the items will be used for the estimation of the final ability estimate is the following: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, **3,** 16, 17, 18, 19, 20 21, 22, 23, 24, 25, 26, 27, 28, 29, 30. Figure 2 describes the above pattern of item responses, with a hypothetical set of data.

Convergence Plots

Figure 3 describes the convergence plot of the ability estimates of a simulee who has a true ability of 0.00. This figure, which is based on the simulated data that are used in this study, reflects the way in which the ML estimation procedure converges to the simulee's final ability estimate. The convergence is examined three times, which is once with each of the three points of the rearrangement process.

This examinee had originally answered item 2 correctly in the simulation. After item review, however, this examinee changed their answer to question 2 to an incorrect answer. Consequently, the rearrangement procedure skipped items 3 and 4, and continued with the use of item 5, which was answered correctly. So after the rearrangement procedure, the posterior θ after item 5 was -0.036. This θ estimate which was closer to the examinee's true ability of 0.00 than the estimate after review that was -0.2094. Eventually, the examinee's final ability estimate after the rearrangement procedure was 0.0098. This was closer to the true score than the estimate before review (-0.1565) as well as the estimate after review (0.0125).

Exceptions to the Rule

A possible problem might exist in the cases where an examinee for example, changes an answer from an incorrect to a correct one, but there are no other appropriate items to replace them. More specifically, there might be no items that were answered incorrectly after the item whose answer was changed that could be used by the estimation procedure. In this case, the procedure would skip three items, and then use the fourth item that comes after the item to which the answer was changed. The same situation could occur when an examinee changes an answer from a correct to an incorrect one, but there are no other items answered correctly that could be used by the rearrangement procedure. Again, like in the previous example, the procedure would skip three items, and then use the fourth item that comes after the item to which the answer was changed.

A second exception to the rule includes the case in which an examinee changes the last item on the test. In this case, no additional changes would have to be made to the estimation procedure, and the final (correct or incorrect) answer would be used to estimate the final ability estimate.

A third exception to the rule would be in the case where more than 3 items have already been skipped. This would cause a problem to the estimation procedure since the examinee's ability estimate would be much worse since there would be too few items that could be used for the estimation. For this reason, no items will be skipped if three items have already been skipped because of the rearrangement procedure.

Stopping rules

In order to avoid possible cheating strategies used by examinees, some restrictions would also have to be made on the revision policy. A meta-analysis conducted by Waddell and Blankenship (1994) found that the mean percentage of items changed in 75 studies was 5.1% when examinees have the option of revision. This means that on a 30 item test, an average of only 1.5 items are changed. So any large deviation beyond 15% might be an indicator that an examinee is trying to cheat. Therefore, a limit would have to be placed on the number of revisions that would be allowable for examinees to make. This would prohibit the Wainer strategy from taking place. So in the case of a 30 item test, a maximum of 5 items would be permitted to be changed for the rearrangement procedure. This should not appear as a major restriction to the examinees since the typical examinee would only change about 2 items out of 30. It should also be noted that if an examinee changes their answer to the same question two times, that would count as one revision, not two.

Dependent Variables

The effects of the rearrangement procedure can be judged in many ways. Two dependent variables were used to help determine the effects that the rearrangement procedure had on the examinees' ability estimates were the bias and the reliability estimate (Kim & Nicewander, 1993). The bias of the final ability estimate, as described in Equation 1, was calculated to determine how much the examinees' estimated scores deviated from their true scores.

$$\text{Bias}_i = \hat{\theta}_i - \theta_i \quad (1)$$

Where $\hat{\theta}_i$ is an examinee's estimated ability and

θ_i is an examinee's true ability

Another way of judging the quality of the results was by estimating the reliability of the ability estimates before review, after review, and after the rearrangement procedure. Equation 2 was used to estimate the reliability of the examinees' ability estimates.

$$\rho_{\hat{\theta}\theta} = \frac{\sigma_{\theta}^2}{\sigma_{\theta}^2 + \sigma_{\hat{\theta}/\theta}^2} \quad (2)$$

where σ_{θ}^2 is the variance of the examinee's true ability and

$\sigma_{\hat{\theta}/\theta}^2$ is the conditional variance of the ability estimates

Results

To determine the accuracy and effectiveness of the rearrangement procedure, the ability estimates were obtained three times, at the three points of the rearrangement process; before review, after review, and after the rearrangement procedure. The before review time point is the one before the examinees in the simulation had the opportunity to change their answers on the test. The point after review describes the ability estimate after the examinees in the simulation had the opportunity to revise and change their answers. Finally, the after the rearrangement procedure (ARP) time point describes the ability estimates in the simulation after the rearrangement procedure was used. At each of the three points, the bias and the reliability were estimated.

Overall, 41.66% of the simulated examinees made correct-to-incorrect, or incorrect-to-correct changes to their answers. These types of changes are the only ones that will be discussed since the rearrangement procedure takes place only when such changes have occurred. Table 2 describes the percentage of actual answer changes, that are divided in the four categories that were discussed in the previous condition in the simulation.

As can be seen from Table 2, the majority of the changes that were made (41.66%) were from incorrect to correct ones. In addition, the majority of those changes were from examinees that made one or two such changes throughout their test. There were also 15.51% of examinees that made correct-to-incorrect changes to questions to which they had approximately a 0.50 probability of answering correctly.

Only 6.23% of the simulated examinees had made 'stupid mistakes' that were then changed to correct answers. Finally, there were also 3.58% of the same examinees that changed their answers to incorrect answers to an item that was originally answered correctly just by chance.

After the examinees reviewed their items on the test in the simulation, the rearrangement procedure was used. Because of the rearrangement procedure, there were 894 examinees (5.89%) that used item review, to which 1 of their items on the test were ignored. There were also 766 simulated examinees (5.05%) that used item review, to which 2 of their items on the test were ignored. Finally, there were also 7853 examinees (51.77%) of the examinees that used item review, to which 3 of their items on the test were ignored.

When items were rearranged because of the rearrangement procedure, the amount of information that was provided at each ability level was used as an indicator for which items should be selected to be used next. The average amount of information that was gained by rearranging the items was 0.0514 with a standard deviation of 0.0421. The minimum amount of information that was gained was 0.0001, while the maximum information that was gained was 0.3396.

Results Based On Bias

Table 3 presents the overall pattern of bias that exists at the three time points of before review, after review, and after the rearrangement procedure.

As described in Table 3, the ML bias estimate before review was -0.1374. After review, the ML bias dropped in magnitude to 0.0718. After the rearrangement procedure, the bias decreased in magnitude further to 0.0610. This was a 15.6% decrease in the bias when compared to the after review bias.

The results of the bias are described more analytically in Table 4, which presents the bias at each of the 13 ability levels from which the examinees were sampled. In most cases the after review bias was smaller than the before review bias. However, at some ability levels such as at the θ level of 1.5, 2.0 and 3.0, the after review bias was larger than the before review bias. The reason for this increase is because in certain cases, the review process eliminated the randomness from the examinee's responses. This resulted in a mismatch between the examinee responses and the IRT model. For example, examinees with an ability of θ=2.0 might have had a 90% probability of answering item i correctly. Consequently, it is expected that 90% of the examinees with a θ of 2.0 would answer item i correctly, and 10% would answer the item incorrectly. However, if 100% of these examinees answer the item correctly, then their response patterns do not match the IRT model, which consequently will increase the after review bias of the ability estimates.

Table 4 shows that there were 9 out of the 13 ability levels where the ML bias decreased in magnitude after the rearrangement procedure. These improvements existed at the θ levels from -1.5 to 2.5. So the effects of the rearrangement procedure were generally more evident at the positive rather than the negative end of the θ scale when the ML estimation procedure is used.

Figure 4 describes the percentage of bias reduction that has occurred from the after review estimates to the ARP estimates. The ML estimation procedure appears to work well at most ability levels since it has a small positive percentage of improvement at 9 of the 13 true ability levels. This means that the ML bias

decreased in magnitude, and the accuracy of the ability estimates improved at most ability levels due to the rearrangement procedure. However, the ARP estimates had much larger bias at the extremes of the distribution, such as at the levels of -3.0 and 3.0. The reason why the ARP estimates were less accurate at the extremes of the distribution is a function of the failure of the ML procedure to converge for examinees whose abilities are at the extremes of the distribution. This occurs when examinees get their answers on the test either all correct, or all wrong. So after review, and after the skipping of items in the rearrangement procedure, it is more likely that more examinees at the extremes of the distribution will get their answers either all wrong, or all correct. Consequently, the ML estimation after the rearrangement procedure will have problems converging for these examinees, which in turn decreases the accuracy of the ability estimates.

Reliability Of Test Scores

The reliability of the ability estimates was also compared when a maximum of 5 reviews were permitted by the examinees. The results of this study show that the reliability of the ML estimates before review was 0.817. After the examinees changed their answers on the test, the ML reliability estimate dropped to 0.811. After the rearrangement procedure, the reliability dropped further to 0.806. However, this drop in reliability was too small to have a significantly negative effect on the quality of the examinee's final ability estimates.

Conclusions

Due to the increased popularity of computerized adaptive testing, many high stakes tests, certification, or even achievement tests are now being converted to a computer adaptive format. However, as researchers are trying to improve many of the components of CAT, students are trying to familiarize themselves with the new testing format and processes of CAT (Pommerich & Burden, 2000; Reckase, 2000). One of the components of adaptive tests that creates some tension between students and CAT researchers, and which has not been conclusively resolved yet, is that of item review. On the one hand, examinees prefer to have the option of changing their answers on adaptive tests (Bowles & Pommerich, 2001). They argue that item review allows them to perform to the maximum extent of their abilities since they are able to rethink over their answers, as well as to correct questions that might have been misread, miskeyed, or miscalculated. This is especially important for examinees that have test taking anxiety (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997) and who are very likely to make careless errors on such tests.

However, some researchers believe that item review should not be permitted on CAT since it does not follow the logic on which adaptive tests are based on (Wise, 1996), and since item review might actually hurt the accuracy of the examinee's ability estimates. For this reason, a rearrangement procedure was proposed in this study. This rearrangement procedure that is used after item review takes place, was hypothesized to improve the accuracy of the examinee ability estimates, without allowing the examinee's to artificially inflate their ability estimates.

How does the rearrangement procedure affect the bias and reliability of the ability estimates?

When the ML estimation was used for obtaining the examinee ability estimates, the bias became smaller after the rearrangement procedure. When looking at the conditional bias of the ML estimates, the overall pattern of bias showed that the ML bias tends to increase from a negative to a positive bias as the ability of the examinees increases. Therefore, examinees with lower ability estimates tend to have a negative ML bias, while examinees with higher abilities tend to have a positive ML bias. This means that examinees at the lower end of the distribution have lower estimated scores than true scores when the ML is used with the rearrangement procedure. The situation is the opposite for examinees at the higher end of the ability distribution where a positive bias exists. This is expected since the ML estimator is more biased towards

the extremes (Kim & Nicewander, 1993; Lord, 1986; Wang & Vispoel, 1998). However, this is a reflection of the properties of the estimator rather than a reflection of the rearrangement procedure.

Only at the extremes of the distribution does the rearrangement procedure fail to produce more accurate ability estimates. This is also a function of the ML estimation procedure that fails to converge the examinee's ability estimates when their answers on the test either all correct, or all wrong. However, since these are such a small percentage of cases, it does not hurt the overall accuracy of the ability estimates since the overall results have shown that the bias improves by 15.6% after the rearrangement procedure.

The changes in reliability because of the rearrangement procedure were very small and they probably do not reflect significant effects on the ability estimates. More specifically, the reliability dropped by 0.05 after the rearrangement procedure. However, this decrease is too small to have a significant effect on the overall results of the study since the overall reliability after the rearrangement procedure was approximately 0.81.

Implications for practice

Overall, the rearrangement procedure has shown some positive and promising results. In addition, the rearrangement procedure is associated with item review that permits examinees to change any mistakes that they might have made, such as miskeyed, miscalculated, or misread, items. These corrections will make the examinees' final ability estimates more valid since the careless errors will be removed from the final test scores (Vispoel, 1998a). In addition, many examinees will have less anxiety when they realize that they can go back and change some of their answers on the test (Wise, Roos, Plake & Nebelsick-Gullett, 1994; Wise, 1997). This will also allow them to pace themselves better throughout the test when they know that they can come back to an item and spend more time on it after they have reached the end of the test.

However, item review is also associated with two main cheating strategies, the Wainer (Wainer, 1993) and the Kingsbury strategy (Kingsbury, 1996). The Kingsbury strategy should not be a major issue for item review since the current research has shown that examinees are not able to use this strategies effectively to artificially inflate their test scores, even when they are taught to do so (Vispoel, Clough, Bleiler, Henderickson & Ihrig, 2001). The examinees that want to cheat will not be able to perform the Wainer strategy either, since that requires changing the answers to all of the items on a test. Since the rearrangement procedure only permits up to five item reviews, then this cheating strategy cannot be effectively used.

In addition, the rearrangement procedure itself does not reduce the efficiency of CAT. If it were used in a real testing situation, it would not require extra testing time for the examinees, and it would not require the administration of additional items either. The rearrangement procedure is just an algorithm that can be used with the ability estimation procedures after the test has ended.

In terms of the rearrangement procedure itself, it is effective in the sense that it can reduce the overall ML bias of the ability estimates. For this reason, if the rearrangement procedure were adopted, the ability estimation procedure that would be used, would have to depend on the dependant variable that would be used as an index of the effectiveness of this procedure.

Limitations

A large component that is missing from this study is the use of real data. Since very few adaptive tests allow review, obtaining real data to base this model on was very difficult. For this reason, the specifications of this study were based on prior studies that have dealt with CAT and with item review.

The only aspect of this simulated CAT, on which no prior research has been done, was on the characteristics of the items that are reviewed by the examinees. For this reason, more research needs to be done with real data from adaptive tests that permitted review, to ensure that the positive effects of the rearrangement procedure can be replicated.

It might also be argued that it is difficult for testing organizations to explain why certain items have been omitted from the examinee's final ability estimates. However, many CATs administer seed items in their tests to pilot them and judge their quality. Such items are not used for the estimation of the examinee's abilities either. So the examinees can just be informed at the beginning of the test that some additional items might be omitted from their test scores in order to improve their ability estimates. In case that it is too risky to omit items on high stakes tests, this procedure could still be used for general achievement and aptitude tests.

Another possible limitation of the rearrangement procedure, is that omitting items from the test will reduce the total amount of test information (Lunz, Bergstrom and Wright, 1992; Wainer, 1993). This is correct. A possible solution to overcome this problem would be to add as many additional items at the end of the test, as the number of items that have been omitted. This would create a large increase in the test information since these additional items would be perfectly targeted to the examinee's 'corrected' ability estimate after review. However, testing organizations would have to judge the feasibility of this solution since administering more items would be more costly to them.

Finally, it is essential for test developers to pilot the use of the rearrangement procedure before applying it to CAT. It is possible that this procedure might have different results when real item pools are used. Consequently, future research needs to examine the effects of the rearrangement procedure with a) real data, b) variable length adaptive tests, c) varying item selection procedures, d) with item selection constraints such as item exposure controls, e) with polytomous items, as well as f) with other estimation methods.

Bibliography

Ban, J. C., Wang, T., Yi, Q. & Harris, D. J. (2000). Effects of nonequivalence of item pools on ability estimates in CAT. Paper presented at the annual meeting of the National Council of Educational Measurement, April, New Orleans, LA.

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.

Camilli, G. & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenzel statistic. Journal of educational measurement, 34 (2), 123-139.

Camilli, G., Wang, M. & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admissions Test. Journal of educational research, 32 (1), 79-96.

Eignor, D. R., Stocking, M. L., Way, W. D. & Steffen, M. (1993). Case studies in computer adaptive test design through simulation. (Research report RR-93-56). Princeton, NJ: Educational Testing Service.

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 392 844).

Harwell, M., Stone, C. A., Hsu, T. & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied psychological measurement, 20(2), 101-125.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. Psychometrika, 58 (4), 587-599.

Kingsbury, G. G. (1996) Item review and adaptive testing. Paper presented at the annual meeting of the National Council of Measurement in Education, NY.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum associates.

Lord, F. M. (1986). Maximum Likelihood and Bayesian parameter estimation in item response theory. Journal of educational measurement. 23 (2), 157-162.

Luecht, R. M. & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. Applied psychological measurement, 16(1), 41-51.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied psychological measurement, 16 (1), 33-40.

McBride, J. R., Wetzel, C. D & Hetter, R. D (1997). Preliminary psychometric research for CAT-ASVAB: Selecting an adaptive testing strategy. . In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing. From inquiry to operation (pp. 83-95). Washington, DC: American psychological association.

Mills, C. N., & Stocking, M. L. (1995). Practical issues in large-scale high-stakes computerized adaptive testing. (Research Report 95-23). Princeton, NJ: Educational Testing Service.

Patsula, L. N. & McLeod. L. D. (2000). Detecting test-takers who have memorized items in computerized-adaptive testing and multi-stage testing: A comparison. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA. .

Pommerich, M & Burden, T. (2000). From simulation to application: Examinees react to computerized testing. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April 2000.

Powers, D. E. (2001). Test Anxiety and Test Performance: Comparing Paper-based and Computer-Adaptive Versions of the Graduate Record Examinations (GRE) General Test. Journal of educational computing research, 24 (3) 249-273.

Reckase, M. D. (1975). The effect of item choice on ability estimation when using a simple logistic tailored testing model. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Reckase, M. D. (2000). Computerized testing- The adolescent years: Juvenile Delinquent or positive role model? Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April 2000.

Robin, F. (1999). CBTS: Computer-based testing simulation and analyses [computer program]. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.

SAS Institute Inc. (1999). Language Reference: Concepts, Version 8. Cary, NC.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. Applied psychological measurement, 21 (2), 129-142.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. Applied measurement in education, 7(3), 211-222.

van der Linden, W. & van Krimpen-Stoop, E. M.L.A. (2001). Using response times to detect aberrant behavior in computerized adaptive testing. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Vispoel, W. P., Clough, S. J., Bleiler, T., Henderickson, A. B. & Ihrig, D. (2001). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Vispoel, W. P. (1998a).Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and text anxiety. Journal of educational measurement, 35 (2), 155-167.

Vispoel, W. P. (1998b). Review and changing answers on computerized adaptive and self-adaptive vocabulary tests. Journal of educational measurement, 35 (4), 328-347.

Vispoel, W. P., Henderickson, A. B. & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. Journal of educational measurement, 37(1), 21-38.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: a comparison of fixed-item, computerized-adaptive, and self-adapted testing. Applied measurement in education, 7(1), 53-79.

Vispoel, W. P., Rocklin, T. R., Wang, T. & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased estimates on a computerized adaptive test? Journal of educational measurement, 36 (2), 141-157.

Waddell, D. L. & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. The journal of continuing education in nursing, 25, 155-158.

Wagner, D., Cook, G., & Friedman, S. (1998). Staying with their first impulse? The relationship between impulsivity/reflectivity, field dependence/field independence and answer changes on a multiple-choice exam in a fifth-grade sample. Journal of research and development in education, 31 (3), 166-175.

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

Wang, T. & Vispoel, W. P. (1998). Properties of estimation methods in computerized adaptive testing. Journal of educational measurement, 35 (2), 109-135.

Wise, S. L. (1996). A critical analysis of the arguments 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 400 267).

Wise, S. L. (1997a). Examinee issues in CAT. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC document reproduction service No. ED 408 329).

Wise, S. L. (1997b). Overview of practical issues in a CAT program. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC document reproduction service No. ED 408 330).

Wise, S. L., Roos, L. R., Plake, B. S., & Nebelsick-Gullett, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. Applied measurement in education, 7(1), 81-91.

Figure 1. Example of an incorrect-to-correct answer change on a CAT (type 2 change)

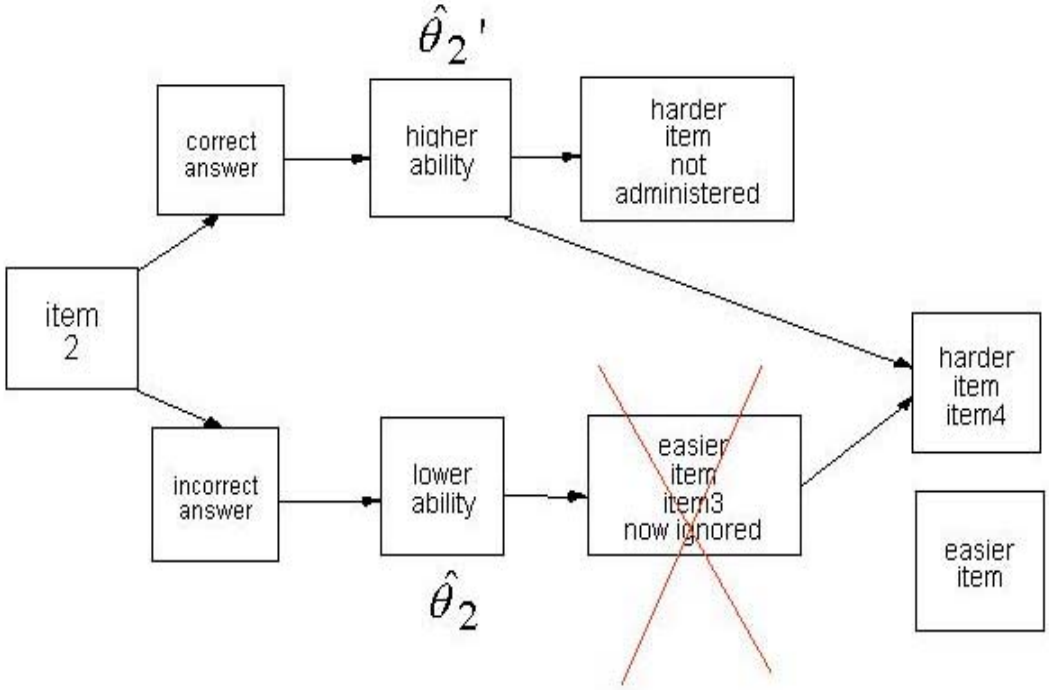


Figure 2. Rearrangement procedure with a rearrangement of the item order

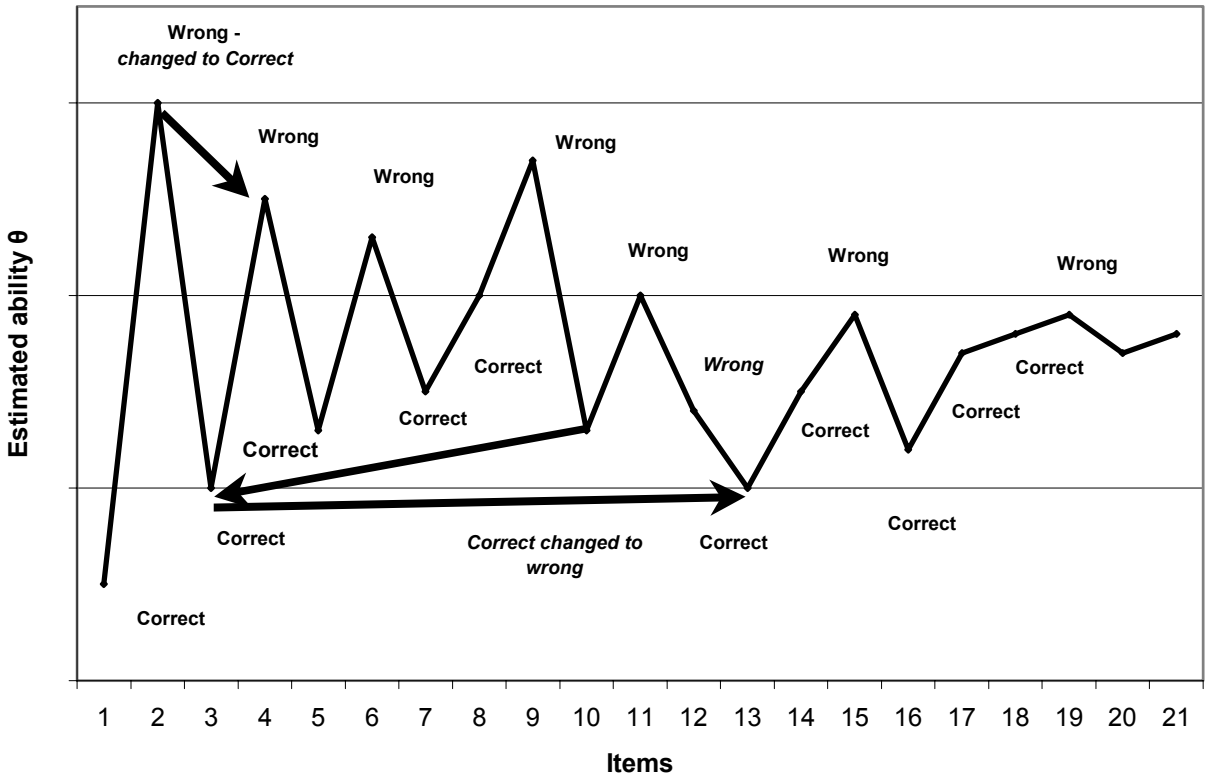


Figure 3. Convergence plot with correct-to-incorrect change

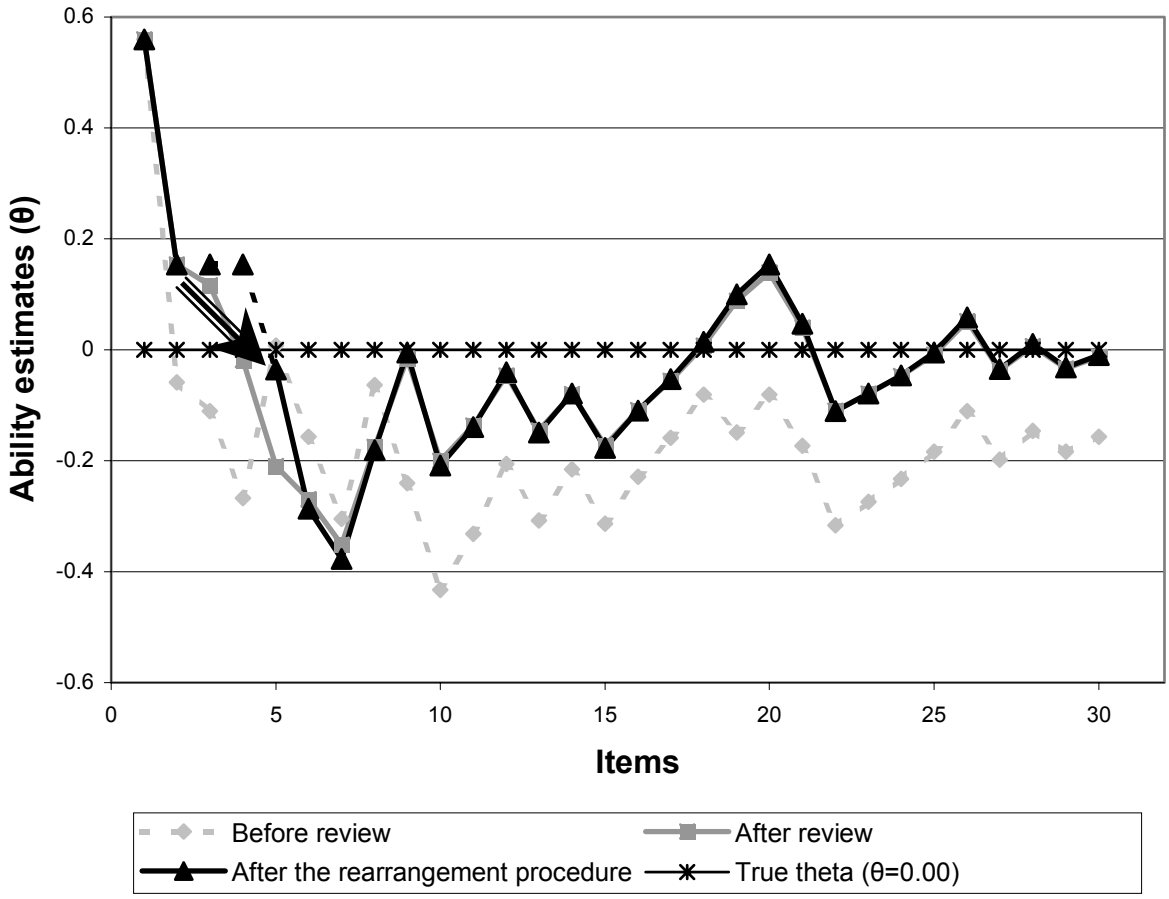
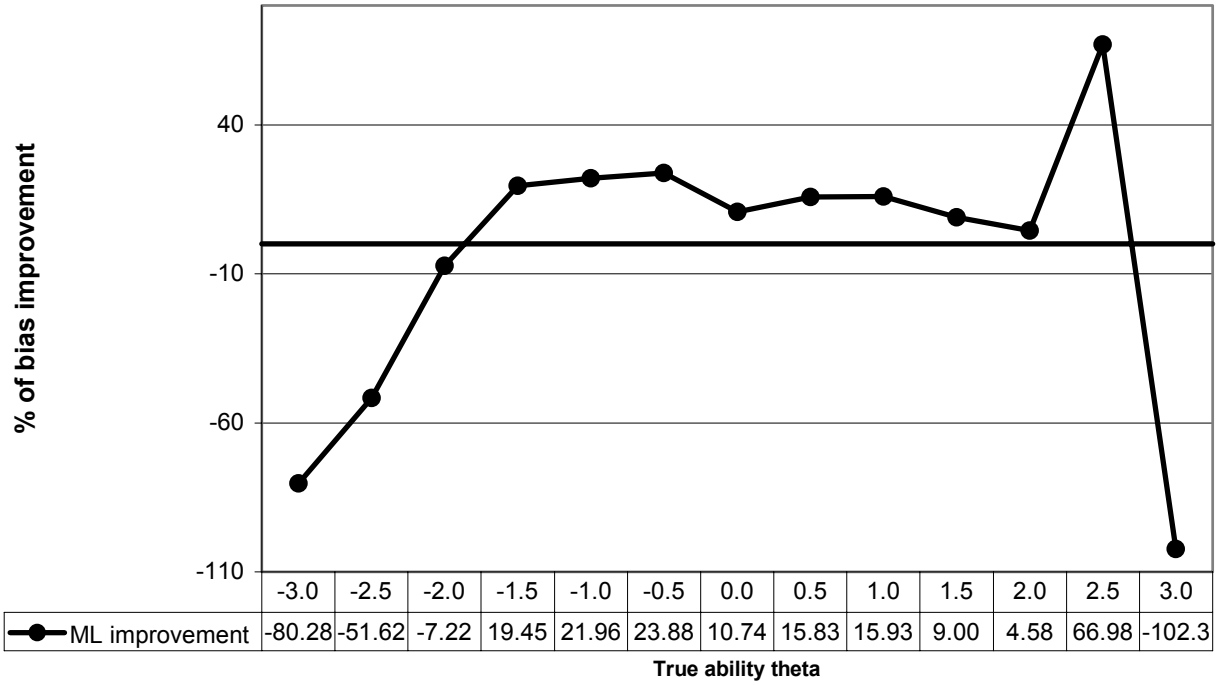


Figure 4. Percentage of bias improvement after the rearrangement procedure



	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
ML improvement	-80.28	-51.62	-7.22	19.45	21.96	23.88	10.74	15.83	15.93	9.00	4.58	66.98	-102.3

Table 1. Target distributional characteristics of the item parameters

Parameters	Mean	SD	Type of distribution	Minimum	Maximum
a parameter	1.10	0.25	Log normal	0.45	2.30
b parameter	0.00	2.00	Uniform	-3.50	3.50
c parameter	0.17	0.10	Normal	0.00	0.35

Table 2. Percentage of actual answer changing patterns

Type of change	**Number of examinees**	**Percentage of examinees (out of 26000)**
Incorrect-to-correct changes (0.5 probability)	10831	41.66%
Correct-to-incorrect changes (0.5 probability)	4032	15.51%
Stupid' mistake corrections (incorrect-to-correct)	1621	6.23%
Unlucky guess' changes (correct-to-incorrect)	930	3.58%

Table 3. Overall bias of the ability estimates

ML Bias	**Mean**	**Standard Deviation**
Before review	-0.1374	0.3229
After review	0.0718	0.3346
After the rearrangement procedure	0.0610	0.3394

Table 4. Conditional Maximum Likelihood bias

Ability θ	Before Review	After Review	After Rearrangement procedure (ARP)	Improvement from rearrangement procedure
-3.0	-0.1261	0.0613	-0.1105	
-2.5	-0.2234	-0.0476	-0.0722	
-2.0	-0.1765	-0.1073	-0.1151	
-1.5	-0.1284	0.0694	0.0559	Yes
-1.0	-0.1469	0.0437	0.0341	Yes
-0.5	-0.1595	0.0438	0.0333	Yes
0.0	-0.1398	0.1167	0.1041	Yes
0.5	-0.1422	0.0770	0.0648	Yes
1.0	-0.1034	0.0242	0.0204	Yes
1.5	-0.0883	0.1940	0.1765	Yes
2.0	-0.1442	0.2080	0.1985	Yes
2.5	-0.1087	-0.0198	-0.0065	Yes
3.0	-0.0501	-0.0520	0.1053	