Reducing the Impact of Inappropriate Items on Reviewable Computerized Adaptive Testing

Yung-Chin Yen, Rong-Guey Ho, Wen-Wei Liao and Li-Ju Chen
Graduate Institute of Information and Computer Education, National Taiwan Normal University, 162, He-ping East Road, Section 1, Taipei 106, Taiwan // Tel: +886 2 77343921 // Fax: +886 2 23512772 // scorpio@ice.ntnu.edu.tw // hrg@ntnu.edu.tw // ljchen@ice.ntnu.edu.tw // abard@ice.ntnu.edu.tw

(Submitted September 20, 2010; Revised May 10, 2011; Accepted August 3, 2011)

ABSTRACT

In a test, the testing score would be closer to examinee's actual ability when careless mistakes were corrected. In CAT, however, changing the answer of one item in CAT might cause the following items no longer appropriate for estimating the examinee's ability. These inappropriate items in a reviewable CAT might in turn introduce bias in ability estimation and decrease precision. An early proposed solution to this problem was rearrangement procedure. The purpose of this study was to implement the 4PL IRT model to reduce the estimation bias introduced by inappropriate items in reviewable CAT. The results of this study indicated that the 4PL IRT model could significantly lower the estimation bias for reviewable CAT, and, while it incorporates with rearrangement procedure, provide more accurate ability estimation. Also, the efficiency of reviewable CAT was promoted by introducing both 4PL IRT model and rearrangement procedure.

Keywords

Item response theory (IRT), Computerized adaptive testing (CAT), Reviewable CAT, Four-parameter logistical (4PL) IRT model, Rearrangement procedure

Introduction

When developing a computerized adaptive testing (CAT), an important controversy was whether examinees should be allowed to review and modify previous responses (Bowles & Pommerich, 2001; Wise & Kingsbury, 2000). In traditional paper-and-pencil (P&P) tests, most examinees took review for granted. They have been trained using the remaining time to review items and catch careless errors since they entered elementary school. In most CATs, however, opportunities for item review and answer change were far less common (Mills & Stocking, 1995; Papanastasiou & Reckase, 2007; Stone & Lunz, 1994; Vicino & Moreno, 1997; Vispoel, 1998; Wise, 1996). Since the testing score would be closer to examinee's actual ability when careless mistakes were corrected, the prohibition of reviewing items in CAT might lead to underestimating examinees' ability (Lunz, Bergstrom, & Wright, 1992; Vispoel, Hendrickson, & Bleiler, 2000; Waddell & Blankenship, 1994). A reviewable CAT was a CAT in which examinees were allowed to review and change the answers of previous items.

The underlying hypothesis of reviewable CAT was that after rereading or rethinking an item, the examinees might correct the careless mistake they made. This hypothesis afterwards led to the fact that even high-ability students might on occasion miss items that they should have answered correctly. However, changing the answer of one item in CAT might cause the following items no longer appropriate for estimating the examinee's ability. These inappropriate items in a reviewable CAT might in turn introduce bias in ability estimation and decrease precision. Early proposed solutions to this problem included limiting review and rearrangement procedure (Chen, 2009; Papanastasiou, 2002).

The same situation was also seen in traditional CAT. In virtue of the underlying characteristics of the traditional item response theory (IRT) model, if a high-ability examinee misses early items carelessly in a test, the following items would be too easy to estimate his/her true ability appropriately. To cope with the underestimation problem, Barton and Lord (1981) proposed the four-parameter logistical (4PL) model allowing a high-ability student to miss an easy item without having his/her ability drastically lowered. In contrast to other well-known IRT models, however, little attention has been given to the 4PL model. The purpose of this study was to implement the 4PL IRT model to reduce the estimation bias introduced by inappropriate items in reviewable CAT. It was also hypothesized that the 4PL model would perform better in alleviating the inappropriate item problems of reviewable CAT than the rearrangement procedure, or at least, improve the performance of rearrangement procedure by incorporating with it.

4PL IRT model

According to the number of parameter describing the item, IRT model could be generally classified into three widely used categories: one-parameter logistic (1PL) model, two-parameter logistic (2PL) model, and 3PL model. In 1PL model, the probability that an examinee with ability θ could answer an item with difficulty b correctly could be mathematically expressed as

$$P_{1PL}(\theta) = \frac{1}{1+\exp[-D(\theta-b)]}. \quad (1)$$

The mathematical form of the 2PL model could be written as

$$P_{2PL}(\theta) = \frac{1}{1+\exp[-Da(\theta-b)]} \quad (2)$$

while new parameter a was called the discrimination parameter which allowed an item to discriminate differently among the examinees (Harvey & Hammer, 1999). In both 1PL and 2PL models, the probability of passing ranges from 0 to 1 as θ goes from -∞ to ∞. On a multiple-choice test, however, the probability of choosing the correct answer did not approach 0 even for low-ability students. Birnbaum (1968) introduced a lower asymptote to handle the situation in which examinees either guessed totally randomly or answered on the basis of their knowledge. The resulting 3PL model was

$$P_{3PL}(\theta) = c + (1-c)P_{2PL}(\theta), \quad (3)$$

where the lower asymptote c represented the probability that an extremely low ability examinee would get the item correct.

In 1PL and 2PL models, the probability that a low-ability student would answer a hard item correctly should approach zero while a high-ability student should answer an easy item with probability approaching one. It was conceivable, however, this assumption might not always hold, since an examinee who knew nothing still had a chance to choose the correct answer in a multiple-choice test. Moreover, the probability might be higher for an examinee who possessed partial knowledge (Bar-Hillel, Budescu, & Attali, 2005; Burton, 2002; Gardner-Medwin & Gahan, 2003; Yen, et al., 2010). On the other hand, high-ability students who are anxious, distracted by poor testing conditions, unfamiliar with computers, careless, or misread the question, might on occasion miss items that they otherwise should have answered correctly (Hockemeyer, 2002; Rulison & Loken, 2009).

Barton and Lord (1981) introduced an upper-asymptote parameter, expressed by the Greek letter delta (δ), into the 3PL model:

$$P_{4PL}(\theta) = c + (\delta-c)P_{2PL}(\theta). \quad (4)$$

While $P_{2PL}(\theta)$ ranges from zero to one, $P_{4PL}(\theta)$ ranges from the lower asymptote, c, to the upper asymptote parameter, δ, for item-specific "carelessness". Figure 1 illustrates a typical ICC for the 4PL IRT model with b=0, a=1, c = 0.2 and δ = 0.9.

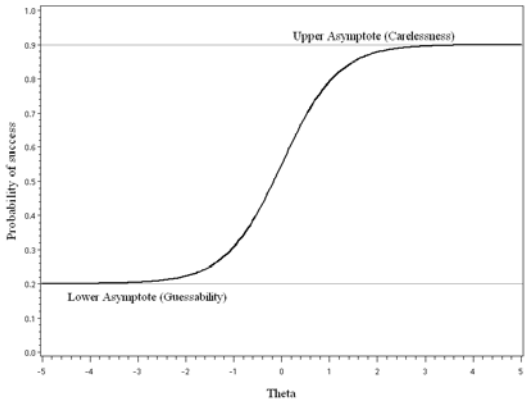
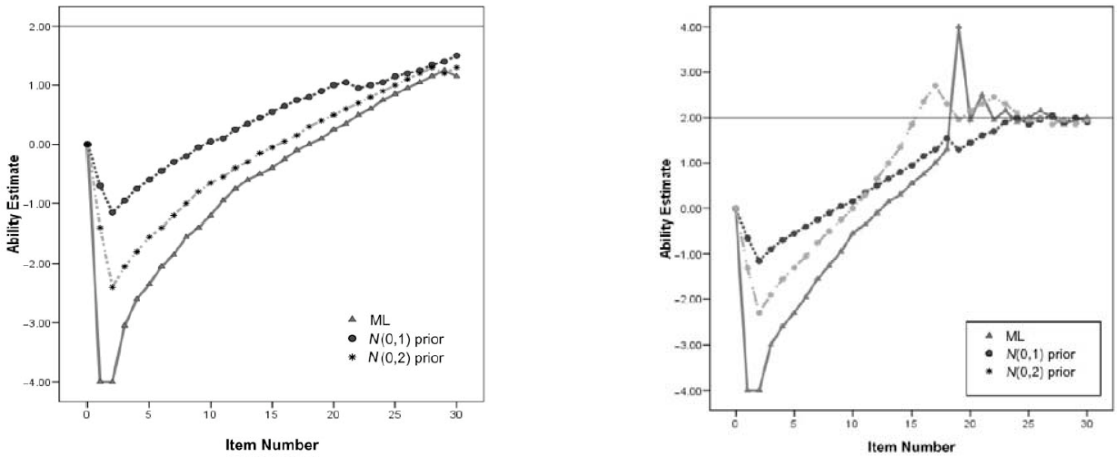


Figure 1. A typical ICC for the 4PL model with b=0, a=1, c = 0.2, and δ = 0.9

To evaluate whether changing the upper asymptote improved scoring on standardized tests, Barton and Lord compared the 3PL model and 4PL model under two upper-asymptote values δ = .99 and δ = .98 by re-estimating test

scores of four data sets: Scholastic Aptitude Test (SAT) Verbal, SAT Math, GRE Verbal, and Advanced Placement (AP) Calculus AB. The results indicated that the changes in ability estimation were too small to be of practical significance (Barton & Lord, 1981). However, it should be emphasized that this study was carried out based on the fixed response data from administered tests in which all examinees received predetermined items from the entire ability range. Hence, the next item was not dynamically selected from item bank according to examinees' accumulated information.

To reevaluate the effect of the upper asymptote on ability estimation in a dynamically CAT environment, Rulison and Loken (2009) conducted two CAT simulation experiments to compare 3PL model with 4PL model in regard to estimation precision for high-ability students with a poor start. In the simulation an examinee (with true θ = 2) missed the first two items, as Figure 2a shows, it was obvious that the initial drop was followed by a very slow ascent in $\hat{\theta}$ under 3PL model.



(a) (b)

Figure 2. Ability estimation in a poor-start 3PL (a) and 4PL (b) model CAT (Rulison & Loken, 2009)

In contrast to Figure 2a, Figure 2b shows the trace plot for a high-ability student (θ=2) who missed the first two items under 4PL model. The errors caused an initial drop almost identical to the drop in the Figure 2a, but $\hat{\theta}$ ascended faster because the upper asymptote of 0.98 discounted the early mistakes. According to Rulison and Loken's study, using 4PL model (δ = 0.98) could lower estimation bias for high-ability students with a poor start. In other words, 4PL IRT model proposed examinees an opportunity to recover from unreasonable responses in CAT.

Reviewable CAT

The term "item review" in testing contexts referred to administrative rules that allowed examinees to change their responses to previously answered items (Vispoel, Rocklin, Wang, & Bleiler, 1999). The opportunity to review items and change answers was usually important to examinees for a variety of reasons (Gershon & Bergstrom, 1995; Harvil & Davis III, 1997; Heidenberg & Layne, 2000; McMorris & Weideman, 1986; Mills & Stocking, 1995; Shatz & Best, 1987):

- Examinees might remember some facts that he/she has not previously recalled, and recognize that they have answered an item incorrectly later in the test.
- Examinees might rethink and conceptualize a better answer for a previous item.
- Review might also be the result of an item occurring at one point in a test cuing the correct answer to a previous item.
- Finally, examinees might want to use the remaining time either to reread the items to avoid careless error, or just to guess another answer to previous questions.

These reasons for choosing to review items could be divided into two major categories: legitimate and illegitimate ones (Wise, 1996). Legitimate reasons were those in which examinees change incorrect to correct answers due to knowledge possessed before the test. This was considered good practice since the final score would reflect the

examinee's ability more accurately. In turn, the validity of the test increased. Illegitimate reasons for those changing answers included the cases in which examinees corrected an incorrect response due to test-wisdom. In this case, the validity of the test decreased (Papanastasiou, 2005). This was the main reason why most CATs did not allow examinees to review items.

Due to the complexity and difficulty of implementation, reviewable CAT was in fact quite rare in practice (Parshall, Kalhn, & Davey, 2002). However, Vispoel et al. (2000) and Papanastasiou (2002) still proposed two algorithms of a reviewable CAT procedure, respectively. These solutions were described in the following two sections.

Limiting answer review and change procedure

Vispoel et al. (2000) proposed the limiting answer review and change procedure that allowed reviewing and changing items within successive m-item blocks. Compared with the traditional CAT, the test items were grouped into n blocks. Figure 3 shows the flow chart of limiting answer review procedure of CAT. In this procedure, examinees were only allowed to review and change answers within the recent block. If an examinee was answering the items in block *j*, he/she was not allowed to review the items in the previous blocks.

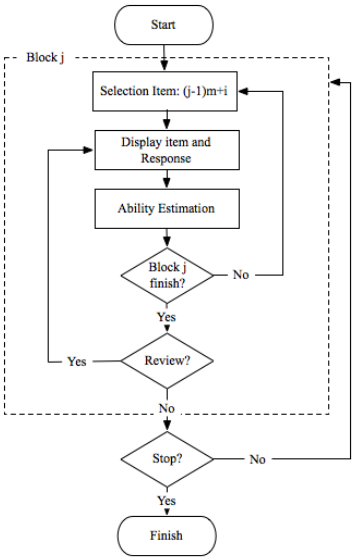


Figure 3. The procedure of limiting reviewable CAT

To verify the performance of the limiting review procedure, an experiment randomly assigning subjects to four CAT conditions was conducted: no review, review of items in 4 blocks of 10 items each, review of items in 8 blocks of 5 items each, or review of all items at the end (Vispoel et al., 2000). In all review conditions, examinees were satisfied with their review options, even when the block size was small. Though the average ability of all review conditions was higher than that of the no review condition, ability estimation did not seem to be related to the various review conditions. As the block size increased, examinees tended to spend more time on the test while there was only 6% difference in testing time between the no review and review in 5-item blocks conditions. The result suggested that block review might function just as well as full review while we reduce item selection problem associated with review.

There were two advantages for using the limiting answer review and change procedure. Firstly, the problem of Wainer strategy could be overcome. That is, the examinees' cheating strategy would not have much effect when they were allowed to review items in a block only (Stocking, 1997; Vispoel et al., 2000). Secondly, there were no significant difference in the accuracy of ability estimation between the limiting review and the no review procedure (Vispoel et al., 2000). This suggested that an examinee could still gain accurate ability estimation by using a reviewable CAT. However, the item administration sequence might become unreasonable after they change answers in this procedure. For example, examinees might correctly change the answers for more difficult items, but such

change would lead to unreasonable item-administration sequence in which some easier items followed a correctly changed, difficult item. These unreasonable response patterns might lead to serious ability estimation bias and decrease testing precision. The rearrangement procedure was proposed to cope with this problem by rearranging the response patterns and re-estimating examinees' ability after answers were changed.

Rearrangement procedure

Papanastasiou (2002) proposed the rearrangement procedure that rearranged and skipped certain items and could better estimate the examinees' abilities. For example, the rearrangement procedure allowed examinees to change up to five of their answers after they finished 30 items in the allotted testing time. After examinees revised their answers, the examinees' final scores were calculated.

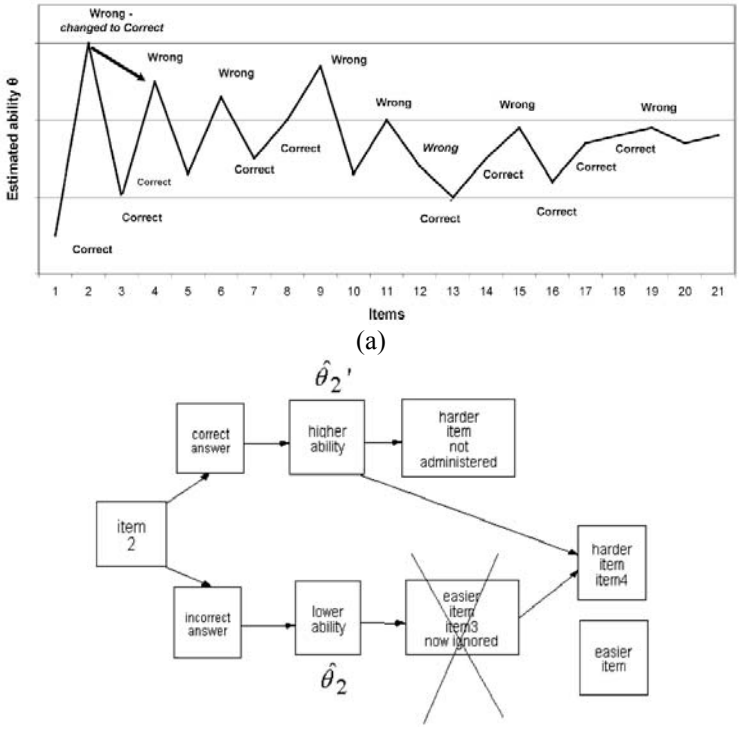


Figure 4. Example of Change C→I (Papanastasiou & Reckase, 2007)

Compared with the traditional CAT, three types of answer changing (Change I→I, Change I→C, and Change C→I) caused the rearrangement procedure in ability estimation. Change I→I involved changing answers from incorrect to incorrect and it would make no difference in ability estimation between the traditional CAT and the rearrangement procedure. The second type (Change I→C) involved changing answers from incorrect to correct and it would result in item skipping in the rearrangement procedure. For instance, if Change I→C was made to item *i,* the ability estimation $\hat{\theta}_i$ would be changed to $\hat{\theta}_i'$. This change in ability estimation would probably make the item *i*+1 not the most informative item for the ability $\hat{\theta}_i'$ since it would be easier and was originally targeted at ability $\hat{\theta}_i$ rather than $\hat{\theta}_i'$. To solve this problem, the ability estimation would skip to item *X* (e.g., item *i*+*k*, with 1<*k*<4) that was answered incorrectly if it was more difficult. It was hypothesized that the new item *X* would be more similar to the item that would have been administered after item *i*. The third type (Change C→I) was to change answers from correct to incorrect and it would also result in item skipping in the rearrangement procedure. For example, if Change C→I was made to item *j*, the item *j*+1 would be ignored in the ability estimation procedure since it would be targeted at a higher ability level than $\hat{\theta}_j$. Therefore, the ability estimation would skip to item *Y* (e.g., item *j*+*k*, where 1<*k*<4) if that was the first easier and correctly answered item after item *j*. So it was hypothesized that the ability estimation

would be more accurate if items *j*+1 through *j*+*k*-1 were ignored from the estimation procedure, and item *j*+*k* was used after the item whose answer was changed.

As Figure 4a shows, a Change I→C (incorrect-to-correct) was made to item 2 of a reviewable CAT. As a result, item 3 became an inappropriate item after the change of this answer since it was selected from the item bank on the premise that the examinee answer item 2 was wrong. In this case, item 3 was skipped because it was targeted at a lower ability level than $\hat{\theta}_2'$. The algorithm therefore jumped to item 4 since that was the first more difficult item answered incorrectly that came after item 3 (see Figure 4b).

The underlying hypothesis of reviewable CAT led to the fact that high-ability students might on occasion miss items that they should have answered correctly. However, almost all previous reviewable studies were carried out based on traditional CAT that assumed that a high-ability student should answer an easy question with probability approaching one. As descried above, the underlying hypothesis of reviewable CAT was consistent with the principle of 4PL model. Besides, the 4PL model might propose examinee an opportunity to recover from the inappropriate responses which introduced by reviewing and changing answer in reviewable CAT. In the present study, therefore, the effect of 4PL model and the rearrangement review solution on reducing estimation bias was investigated.

Methodology

In this study, a simulated experiment was conducted to evaluate the effect of implementing the 4PL model on reducing the impact of inappropriate items on reviewable CAT. The participant demography, item bank characteristics, and procedure are described as follows.

Participants

A group of 13,000 examinees were simulated for this study (1,000 examinees for thirteen equally spaced θ levels). The θ-level groupings ranged from -3.0 to 3.0 at equally spaced intervals of 0.5.

Item bank

The simulated item bank with 250 items was generated according to the specifications proposed by Papanastasiou and Reckase (2007). To determine the item pool characteristic for their simulation study, Papanastasiou and Reckase reviewed psychometric literature to obtain information on the distributions of item parameters of real item pools. Table 1 describes the targeted distributional characteristics of the item pool created for this simulation. As for the upper asymptote, a test-wide δ with value of 1 for 3PL-based CAT and 0.98 for 4PL-based CAT was determined, respectively. By designing item bank of 3PL- and 4PL-based CAT in this way, the only difference between these two models would be the value of upper asymptote, and the results would be correspondent with the intention of this study.

Table 1. Distributional Characteristics of the Item Parameters

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

Simulation procedure

Four versions of 30-item, fixed-length CAT were developed for this simulation experiment (as Figure 5 shows). All these four CATs were reviewable CAT with 6 blocks of 5 items, and the reviewing and changing answer were allowed only within block. The first two CATs (R3CAT and R4CAT) were conducted to compare the performance

of 3PL- and 4PL-based CAT on precision and efficiency. The third one, RR3CAT, was a reviewable CAT with rearrangement procedure while the RR4CAT was a 4PL-based reviewable CAT implementing rearrangement procedure. These two CATs were administered to investigate whether the precision and efficiency of reviewable CAT with rearrangement procedure would be improved by implementing the 4PL IRT model.

1. **R3CAT**: A reviewable CAT based on 3PL IRT model.
2. **R4CAT**: A reviewable CAT based on 4PL IRT model.
3. **RR3CAT**: A 3PL-based reviewable CAT implementing rearrangement procedure.
4. **RR4CAT**: A 4PL-based reviewable CAT implementing rearrangement procedure.

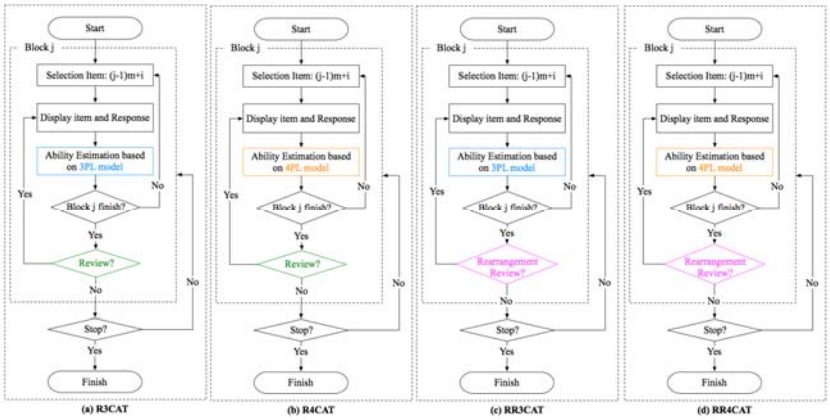


Figure 5. The flowchart of four reviewable CATs

In all four CATs, the first test item started at an item with a middle difficulty. Bayesian expected a posteriori (EAP) estimation was used to estimate examinees abilities in all four CATs while items were selected using a simple maximum-information criterion. Based on the item parameters and simulated θs, the probability (p-value) of an examinee answering an item correctly according to the corresponding IRT model was calculated. This p-value then was compared to a randomly generated number (x) from a uniform distribution $U(0,1)$. If the p-value is larger than or equal to the uniform random number, the simulee got a correct response; otherwise, an incorrect response was obtained for the item. Referring to the study of Rulison and Loken (2009), the Fisher's information for R3CAT and RR3CAT was given by

$$I_j(\theta) = \frac{(1.7a_j)^2(1-c_j)}{(c_j + e^{1.7a_j(\theta-b_j)})(1+e^{-1.7a_j(\theta-b_j)})^2} \tag{5}$$

and Fisher's information function for the R4CAT and RR4CAT was given by

$$I_j(\theta) = \frac{(1.7a_j)^2(\delta-c_j)^2}{(c_j + \delta e^{1.7a_j(\theta-b_j)})(1-c_j+(1-\delta)e^{1.7a_j(\theta-b_j)})(1+e^{-1.7a_j(\theta-b_j)})^2} \tag{6}$$

where

$I_j(\theta)$ was the item information at $\theta = \hat{\theta}$ for item j,

a_j was the item discrimination parameter for item j,

b_j was the difficulty paramenter for item j,

c_j was the lower asymptote parameter for item j, and

δ_j was the upper asymptote parameter for item j.

As five items within each block were answered, the reviewing and changing answer procedure took place. One of three situations might occur in each rearrangement procedure: 1. If the simulee had a 0.80 or higher probability of answering these items correctly (p-value≥0.80), those answers would be changed from incorrect to correct (Change I→C); 2. If the simulee only had a 0.33 or lower probability of answering correctly (p-value≤0.33), the answer would be changed from correct to incorrect (Change C→I) during the revising and changing procedure; 3. For those simulees whose probability to correctly answer an item was from 0.47 to 0.53 (0.47≤p-value≤0.53), the probability of answering item correctly (Change I→C) would be 0.72 and the probability of answering item incorrectly (Change

C→I or Change I→I) would be 0.28. These probabilities described above were determined according to Papanastasiou's (2005) study.

Results

By comparing the estimation bias (Equation 7), median absolute deviation (MAD) (Equation 8), and standard error of estimation (SE) of R3CAT, R4CAT, RR3CAT, and RR4CAT, the precisions of 3PL and 4PL models on reviewable CAT were investigated. On the other hand, by comparing the number of items needed to reach certain SE levels, the efficiency of the four CATs each was evaluated.

$$Bias(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta), \tag{7}$$

$$MAD(\hat{\theta}) = \frac{\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right|}{N}. \tag{8}$$

Table 2 shows that the descriptive statistics of 13,000 simulees' three change types; the conditions of two regular reviewable CATs (R3CAT and R4CAT) were close to the results of meta-analysis by Waddell and Blankenship (1994). One interesting finding revealed that the number of Change I→C (change from incorrect to correct) in reviewable CAT with rearrangement procedure (RR3CAT and RR4CAT) was smaller than that in traditional reviewable CAT (R3CAT and R4CAT). In a rearrangement procedure, if Change I→C was made to item *i*, the following easier items might be skipped since the ability re-estimation probably make these items (item *i* to item *i*+*k*, with $1<k<4$) not appropriate for the newly estimated ability. These skipped items, however, might be the potential targets for Change I→C. This explained the lower rate of Change I→C in RR3CAT and RR4CAT. Though those inappropriate harder items might also be skipped in Change C→I, the higher rate of Change I→C indicated that more inappropriate easier items were skipped than inappropriate harder items in rearrangement procedure. This also explained the higher rate of Change C→I in RR3CAT and RR4CAT compared with that in R3CAT and R4CAT.

Table 2. Percent of three change types (Number of Responses=3,900,000)

Change type \ CAT type	R3CAT (%)	R4CAT (%)	RR3CAT (%)	RR4CAT (%)
Change I→I	21.59	24.75	20.85	22.47
Change I→C	67.55	64.17	52.38	49.12
Change C→I	10.87	11.08	26.76	28.41

To exactly investigate the effect of 4PL model and rearrangement procedure on reviewable CAT, it was essential to distinguish reviewed items and changed answers. The reviewing behaviors included Change I→I, Change I→C, and Change C→I. The changed answers meant those changes of responses that would cause the re-estimation of ability (i.e., Change I→C + Change C→I). Table 3 presents the percent of reviewed items, changed answers, and examinees who changed at least one item during the CAT administration. Since our main concern in this experiment was for those examinees who changed at least one item in the simulation, simulees who never changed answer during the simulation were excluded from the following evaluation.

Table 3. Percent of changing behavior (Number of Responses=3,900,000)

Variable \ CAT	R3CAT (%)	R4CAT (%)	RR3CAT (%)	RR4CAT (%)
Reviewed items	14.03	13.26	12.36	12.52
Changed answers	11.00	9.98	9.79	9.71
Simulees changed at least one item	96.38	95.02	97.55	97.35

The precision of four solutions for reviewable CAT

To investigate the performance of 4PL-based IRT model and rearrangement procedure in regard to diminishing estimation bias caused by inappropriate items in reviewable CAT, the estimation bias of four CATs were computed. Figure 6 indicates that the ability underestimation of 3PL-based CAT (R3CAT and RR3CAT) was alleviated in 4PL-based CAT. The bias for the 4PL-based CATs tended to be smaller than that of 3PL-based CAT, except for the top and bottom ability levels.

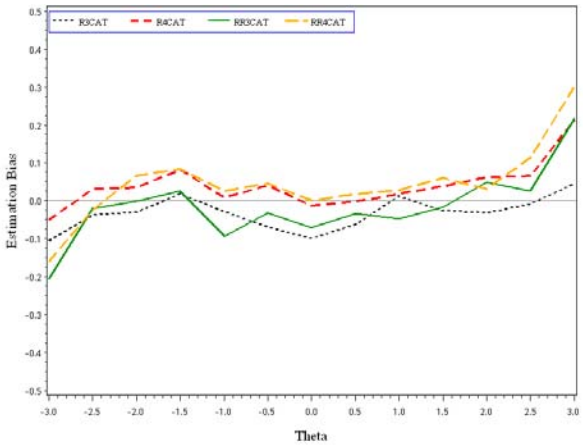


Figure 6. Estimation bias of four CATs

To further evaluate the degree of estimation bias of the four CATs, the estimation MADs of these four CATs were calculated and the ANOVA of MAD was conducted. The value of MAD indicated the distance between the examinees' estimated ability and true ability. The lower the value of MAD was, the more precise the ability estimation was. As Figure 7 shows, the estimation MADs of R4CAT were lower than those of R3CAT across all ability levels. The performance of RR4CAT was also better than that of R3CAT except for some ability levels (θ=-3.0, -0.5, and 3.0). On the other hand, the performance of reviewable CAT with rearrangement procedure (RR3CAT) was similar to that of R3CAT. The lower MAD indicated that the 4PL-based IRT model was a better solution to reviewable CAT.

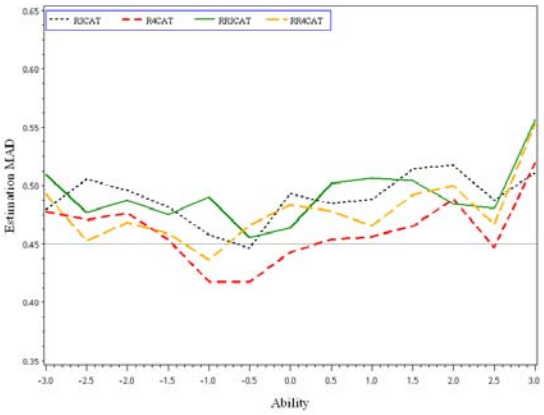


Figure 7. Estimation MAD of four CATs

Since the purpose of this study was to evaluate the performance of 4PL-based IRT and rearrangement procedure by comparing the four CATs, 390 simulees (30 simulees in each ability level) who changed at least one item in each CAT were randomly selected and the repeated measures ANOVA of MAD was conducted. This sampling method was applied to the following repeated measures ANOVA in this experiment. The result of the Mauchly's Test of sphericity indicated that there was no significant difference among the four sets of MAD. A summary of the results of the repeated measures ANOVA is shown in Table 4. As Table 4 shows, though the precision of reviewable was

improved by rearrangement procedure (RR3CAT and RR4CAT), the difference among R3CAT, RR3CAT, and RR4CAT was not significant. According to the results of repeated measures ANOVA, the precision improvement by incorporating 4PL IRT model with rearrangement procedure (RR4CAT) was not statistically significant. Compared with these three CATs, R4CAT significantly improved the precision of ability estimation for reviewable CAT.

Table 4. Repeated Measures ANOVA of MAD (n=390)

Source	*SS*	*df*	*MS*	*F*	Pairwise comparison
CAT type	1.09	3	0.36	2.64*	R4CAT < RR4CAT, RR3CAT, R3CAT
Error	160.47	1167	0.14		

*p<.05.

To further investigate the difference of estimation precision between these four CATs, the estimation SE of these four CATs was evaluated. Figure 8 indicates that SEs of 4PL-based CAT (R4CAT and RR4CAT) tended to be smaller than that of 3PL-based CAT (R3CAT and RR3CAT) across most ability levels, except for the very low ability level (θ=-3.0). The precision improvement introduced by rearrangement procedure was also obvious since the SE of RR3CAT was smaller than that of R3CAT and the SE of RR4CAT was smaller than that of R4CAT across all ability levels. However, according to the results of repeated measures ANOVA, the precision improvement introduced by 4PL IRT model was more significant.

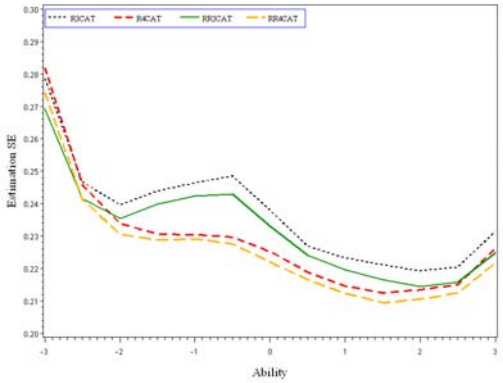


Figure 8. Estimation SE of four CATs

To provide a reliable measurement for repeated measures ANOVA of SE, the Mauchly's test of sphericity was conducted. The results indicated that the data did not meet the assumption of sphericity, thus the Greenhouse-Geisser adjusted *F* test was consulted. It turned out that the values of Greenhouse-Geisser adjusted *F* was the same as those when sphericity was assumed. Table 5 shows the result of repeated measures ANOVA of SE and it revealed that the main effect of CAT was significant. In summary, the precision of 4PL-based CAT was superior to that of 3PL-based CAT, and the precision of 4PL-based CAT (R4CAT) was improved by incorporating with the rearrangement procedure (RR4CAT).

Table 5. Repeated Measures ANOVA of SE (n=390)

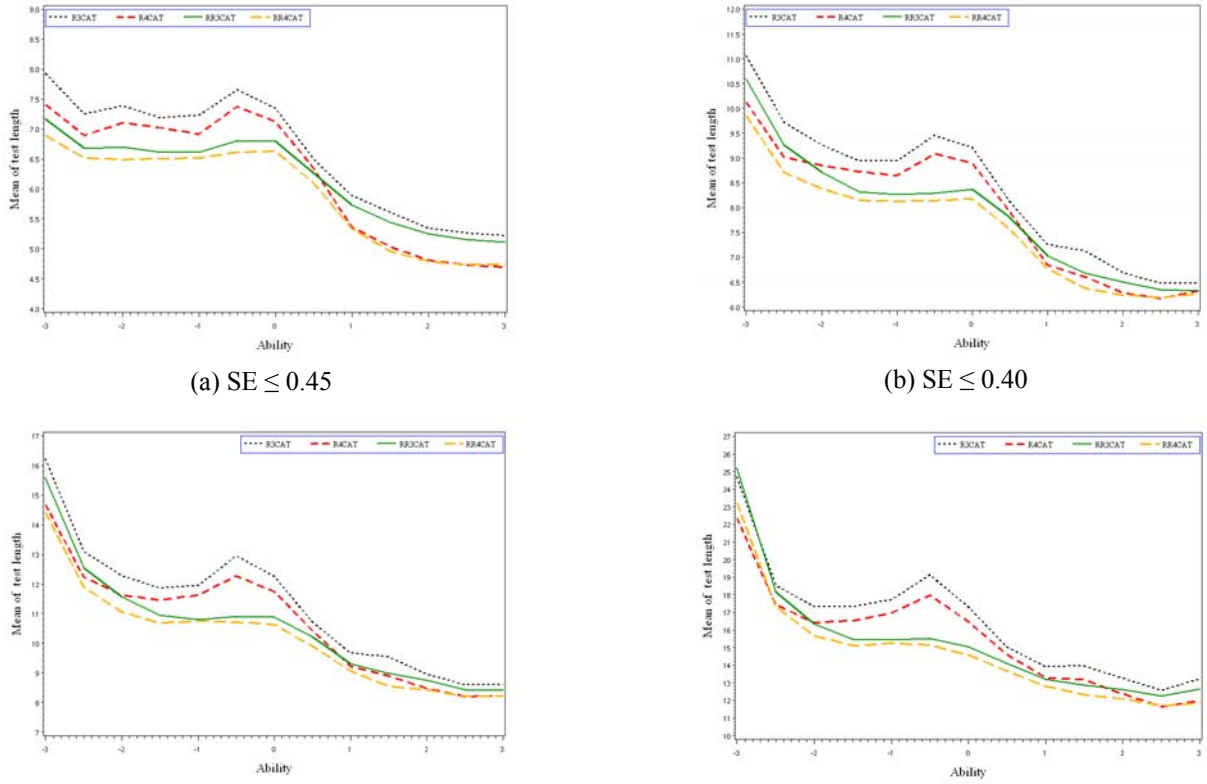
Source	*SS*	*df*	*MS*	*F*	Pairwise comparison
CAT type	0.03	3	0.01	2.64*	RR4CAT<R4CAT<RR3CAT<R3CAT
Error	0.11	1167	0.00		

*p<.05.

The efficiency

The efficiency of CAT was evaluated by comparing all four versions of CAT within the average number of items required for each examinee's ability estimation to reach or surpass a fixed level of precision. Based on a set of stopping criteria including SE reaching to 0.45, 0.40, 0.35, and 0.30, the required numbers of items for R3CAT, R4CAT, RR3CAT, and RR4CAT were compared. Figure 9a depicts the required number in each ability levels for four CATs while SE reaches to 0.45. According to the test length plots, the efficiency of reviewable CAT was improved by introducing the 4PL-based IRT model and rearrangement procedure since the required items of R4CAT,

RR3CAT, and RR4CAT were all smaller than that of R3CAT. The efficiency of R4CAT was superior to RR3CAT at the middle and higher ability levels ($\theta \geq 0$) while the RR3CAT performed better at the rest of ability levels. The efficiency of RR4CAT was superior to that of other three CATs across all ability levels.



(a) SE ≤ 0.45

(b) SE ≤ 0.40

(c) SE ≤ 0.35

(d) SE ≤ 0.30

Figure 9. Mean of test length for four CATs

Figure 9b describes the mean number of items in each ability level for four versions of CAT while SE reaching to 0.40. Like the situation in the previous criterion, the efficiency of reviewable CAT was improved by introducing the 4PL-based IRT model and rearrangement procedure. Compared to RR3CAT, the required number of items for R4CAT was smaller except for the ability levels ranging from -2.0 to 0.5. RR4CAT maintained the best performance of efficiency across all ability levels. As Figure 9c shows, the R4CAT was more efficient than RR3CAT at the higher and lower ability levels. This result was similar to that of previous stopping criterion. While the stopping criterion was set to SE less than or equal to 0.30 (see Figure 9d), four CATs' performance of efficiency was consistent with those of previous two stopping criteria. The efficiency of RR4CAT was identically superior to other three CATs across all ability levels.

To investigate the differences in efficiency among these four CATs, repeated measures ANOVA of test length based on four preset stopping criteria was conducted. Though the variances in the differences between the four sets of test length did not meet the assumption of sphericity according to the result of Mauchly's test of sphericity, the adjusted *F* was the same after the Greenhouse-Geisser adjustment was consulted. Results of repeated measures ANOVA is provided in Table 6.

In summary, there was a significant difference in efficiency among these four CATs. Both 4PL-based IRT model and rearrangement procedure improved the efficiency of reviewable CAT significantly. Though the average required number of item for R4CAT was smaller than that for RR3CAT across all stopping criteria, the differences were not statistically different. The efficiency of reviewable CAT was significantly improved while the 4PL IRT model incorporated with rearrangement procedure (RR4CAT).

Table 6. Repeated Measures ANOVA of Test Length (n=390 in each CAT type)

Criterion	*Source*	*SS*	*df*	*MS*	*F*	Pairwise Comparisons
SE ≤ 0.45	CAT type	98.13	3	32.71	40.37***	RR4CAT<R4CAT,
	Error	945.63	1167	0.81		RR3CAT<R3CAT
SE ≤ 0.40	CAT type	104.86	3	34.95	31.93***	RR4CAT<R4CAT,
	Error	1277.39	1167	1.09		RR3CAT<R3CAT
SE ≤ 0.35	CAT type	236.18	3	78.73	33.68***	RR4CAT<R4CAT,
	Error	2727.82	1167	2.34		RR3CAT<R3CAT
SE ≤ 0.30	CAT type	694.11	3	231.70	53.79***	RR4CAT<R4CAT,
	Error	5026.64	1167	4.31		RR3CAT<R3CAT

***p<.001.

Conclusion

In a test, the testing score can be closer to examinee's actual ability when careless mistakes are corrected. The rearrangement procedure proposed by Papanastasiou (2002) offers examinees opportunity to review and change administered items in a CAT. By limiting the number of reviewing items, the rearrangement can not only avoid cheating strategy caused by reviewing, but also require no extra testing time. Moreover, the rereading and rethink process is important since taking a test is not just a passive mechanism for assessing student. Taking a test actually helps student learn, and it may work better than a number of other techniques.

In CAT, however, changing the answer of one item in CAT might cause the following items no longer appropriate for estimating the examinee's ability. These inappropriate items in a reviewable CAT might in turn introduce bias in ability estimation and decrease precision. This study implemented the 4PL IRT model as a solution to the problem of estimation bias introduced by inappropriate items in reviewable CAT. The simulation result indicated that the 4PL IRT model could significantly lower the estimation bias for reviewable CAT, and provide more accurate ability estimation while incorporating with rearrangement procedure. Also, the efficiency of reviewable CAT was promoted by introducing both 4PL IRT model and rearrangement procedure.

Acknowledgements

This work was supported in part by the Science Education Division of Taiwan National Science Council under the Grant No. NSC 99-2511-S-003-010-MY2. The project principal investigator, Rong-Guey Ho, is also the corresponding author of this article. The authors would like to thank the NSC of Taiwan for financial support.

References

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society, 4*(1), 3–12.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Services.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). London: Addison Wesley.

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

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education, 36*(9), 805–811.

Chen, L. J. (2009). *Effects of block-review and rearrangement computerized adaptive test on ability estimation and test anxiety.* Unpublished doctoral dissertation, National Taiwan Normal University, Taiwan.

Gardner-Medwin, A. R., & Gahan, M. (2003). *Formative and summative confidence-based assessment.* Paper presented at the 7th International Computer-Aided Assessment Conference, Loughborough, UK.

Gershon, R. C., & Bergstrom, B. (1995, April). *Does cheating on CAT pay: NOT!* Paper presented at the the annual meeting of the American Educational Research Association, San Francisco, CA.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353–384.

Harvil, L. M., & Davis III, G. (1997). Medical students' reasons for changing answers on multiple-choice tests. *Academic Medicine, 72*(10), S97–S99.

Heidenberg, A. J., & Layne, B. H. (2000). Answer changing: A conditional argument. *College Student Journal, 34*(3), 440–450.

Hockemeyer, C. (2002). A comparison of non-deterministic procedures for the adaptive assessment of knowledge. *Psychologische Beiträge, 44*(4)*,* 495–503.

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.

McMorris, R. F., & Weideman, A. H. (1986). Answer changing after instruction on answer changing. *Measurement and Evaluation in Counseling and Development, 19*(2), 93–101.

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

Papanastasiou, E. C. (2002, April). *A 'rearrangement procedure' for scoring adaptive tests with review options*. Paper presented at the the National Council of Measurement in Education, New Orleans, LA.

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

Papanastasiou, E. C., & Reckase, M. D. (2007). A "rearrangement procedure" for scoring adaptive tests with review options. *International Journal of Testing, 7*(4), 387–407.

Parshall, C. G., Spary, J. A., Kalhn, J. C., & Davey, T. (2002). *Practical considerations in computer based testing*. New York: Springer-Verlag.

Rulison, K., & Loken, E. (2009). I've fallen and I can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83.

Shatz, M. A., & Best, J. B. (1987). Students' reasons for changing answers on objective tests. *Teaching of Psychology, 14*(4), 241–242.

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.

Vicino, F. L., & Moreno, K. E. (1997). Human factors in the CAT system: A pilot study. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 157–160). Washington, DC: APA.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35*(4), 328–345.

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*(1), 21–38.

Vispoel, W. P., Rocklin, T. R., Wang, T., & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability 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. *Journal of Continuing Education in Nursing, 25*(4), 155–158.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the the annual meeting of the National Conference on Measurement in Education, New York, NY.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica, 21*(1–2), 135–155.

Yen, Y. C., Ho, R. G., Chen, L. J., Chou, K. Y., & Chen, Y. L. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Educational Technology & Society, 13*(3), 163–176.