Adjusting "scores" from a CAT following successful item challenges

Tianyou Wang, Qing Yi, Jae-Chun Ban, Deborah J. Harris, and Bradley A. Hanson
ACT, Inc.

Presented at the Annual Meeting of the American Educational Research Association

Montreal
April 1999

Abstract

This simulation study investigated two strategies for dealing with flawed items (rescoring and deleting), and the positioning of the flawed items, in a CAT environment. In summary, the results show that with the MLE estimates, the deleting strategy seems to produce smoother results and does not systematically inflate or deflate the score estimates. The rescoring strategy seems to produce some unstable results, causing the estimates to change dramatically at a certain ability range. With the EAP estimates, on the other hand, there does not seem to be much difference between the two strategies. However, with the EAP estimates, all these conditions produce some systematic inflation for lower ability examinees and deflation for higher ability examinees. The item position analyses seem to suggest that "flawed" items presented at the end of the CAT appear to have more effect on the differences between initial and 'corrected' scores than do items appearing at the beginning of the CAT.

Adjusting "scores" from a CAT following successful item challenges

The incidence of flawed items actually administered to examinees in many large scale testing programs, such as the ACT Assessment, historically has been very small. For example, in the ACT Assessment forms the present study is based on, there was one item considered problematic out of 8 forms, where each form consists of 75 unique English items, 60 unique mathematics items, 40 unique reading items, and 40 unique science reasoning items. This is true for many large standardized testing programs, who, though they construct multiple forms of an assessment each year and therefore may be developing hundreds of items, also adhere to rigorous test development procedures, such as multiple item reviews and pretesting. Despite the care with which items are developed, however, it is impossible to prevent problematic items from being administered to examinees. For instance, because of the multiple item review processes and the time needed for pretesting, it may easily take 18 months or more from when an item is initially drafted to when it appears as an operational item administered to examinees. In some fields, this may be enough time to change current practice/knowledge such that the item no longer has a single correct answer when it is scored as an operational item. In a liscensure test environment, it may not be possible to pretest items, for security reasons. In addition, in a CAT environment, there may be increased item development to provide items for several item pools, which may result in an increased number of flawed items administered to examinees. The question arises as to what action should be taken in reporting scores for examinees administered one or more flawed items.

Brennan and Kolen (1987) discuss the issues involved when a flawed item is discovered after the operational administration for a paper and pencil test. They view the problem from an equating perspective, but come up with four possible solutions: use the original item key and the original test form equating (hence, the original test scores for the examinees); use the original key with a revised equating (not sensible); use a revised key with the original equating (examinees with a newly-keyed-correct response will receive different scores); or use a revised key and revised equating (examinees receive revised scores). The issue of which action is fairest to examinees is not clear cut.

Within a CAT environment, the options are somewhat analogous. In the case of CATs, all examinees who were administered the now-thought-to-be-flawed item would need to have their 'scores' (whether estimated theta, estimated raw score, or other value) reestimated. Unless the particular item were administered as the last item in a fixed item test, adjusting the score can be very complex, in that subsequent items frequently depend on the examinee responses to earlier items. Other factors, such as how to compensate examinees administered a timed test for the time they spend on a flawed item, and therefore don't have to spend on items that 'count', would be even more complex, as would considering any sort of personality factors, such as frustration, when an examinee receives feedback that is erroneous because the item is flawed (for example, being told his/her response to the item is incorrect during administration).

Potenza and Stocking (1997) undertook a simulation study to look at ways to deal with flawed items in a computerized adaptive test. They considered two types of flawed items: the first type had no correct answer (to make the item very difficult; Potenza and Stocking set the item parameters as a=3, b=10, c=0). The second flaw was more than one correct answer (making the item sill difficult, but with some examinees able to correctly guess a correct response. Potenza and Stocking set the item parameters as a=3, b=10, c=.25). They categorized strategies for dealing with flawed items as either removing the item from the test, or rescoring the item in a "reasonable fashion", and compared these two methods with readministering the item pool to the examinees affected by the flawed item, but with the flawed items removed from the pool. They considered a 30-item fixed length test, incorporating both discreet items and items associated with a common stimulus, and used LOGIST (Wingersky, 1983) to estimate item parameters, the Stocking and Swanson (1993) method for item selection, and incorporated the extended Sympson and Hetter (1985) exposure control methodology.

Potenza and Stocking considered five experimental conditions: the 25 most popular items were considered to be flawed; the two most popular items were considered to be flawed; two typical items were considered to be flawed; the two most popular items as a first item were considered to be flawed, and the two most popular items as the last item were considered to be flawed. They generated data using a second set of item parameters for the flawed items, then used the unflawed item parameters for item selection (simulating a scenario where the item was not flawed during pretesting/introduction into the pool, but subsequently became flawed). They used four rescoring methods, including rescoring with the flawed items removed, and rescoring with the flawed items scored as all-correct. Finally, they deleted the flawed items from the item pool and 'readministered' the CAT to the examinees, using the reduced item pool. Not unexpectedly, the rescaling by scoring all responses as correct resulted in larger increases in test score than in removing the flawed item, or other rescoring methods. Considering the first items as flawed had more of an effect on examinees scores than considering the last item flawed. The authors determined it made no practical difference in terms of fairness to examinees whether the CAT was rescored or if another test was administered using the reduced item pool.

The present study is a simulation study designed to investigate how a rescoring strategy of trying to compensate for administering an examinee one or more flawed items compared with a deleting strategy, in terms of which strategy resulted in scores (raw scores and estimated thetas) closer to the true values, under various simulation conditions.

Methods

Data and CAT Simulation

Computer simulation methods were used in this study. The investigation was conducted under a classical CAT condition where no practical constraints such as item exposure rate control or content balance in a CAT administration were implemented. An actual item pool containing 480 items from eight ACT Mathematics Test forms was used in this study. The ACT Mathematics test is a 60 minute, 60-item 5-option multiple

choice test, focusing on the knowledge and skills attained as the cumulative effects of the school experience. The specifications for the test are given in the Appendix. The three parameter logistic (i.e., 3-PL) IRT model was considered as the model for item parameter calibration. Item parameters were calibrated using the BILOG computer program (Mislevy & Bock, 1990).

The data simulation was conditional on the θ scale. Twenty-one equally spaced points on the θ scale from -4 to 4 in increments of .4 were used. At each of these 21 θ points, responses for 1,000 examinees were simulated.

The CAT administration started with an initial ability estimate of zero for all the examinees. In the CAT administration, maximum item information was used to select the next administered item. A thirty-item fixed length CAT with two ability estimation methods, the maximum likelihood estimation (MLE) and the expected a posterior (EAP) methods were included. Simulees' CAT item responses were determined using the calibrated item parameters and provisional ability estimation. These item responses (0/1s) were generated by comparing the probability of a correct response (i.e., *P*-value) that is based on the 3-PL IRT model with a uniform random number $U(0, 1)$. If the *P*-value was less than the random number, the examinee received an incorrect response (i.e., 0), otherwise, a correct response (1).

Flawed Item Simulation

This study was designed to simulate a situation in which an item was mis-keyed in the test administration. For example, the original answer key had one selection designated as the correct choice. However, after the test administration it was discovered that other choices for this question might also be correct. To realistically simulate the characteristics of the flawed items, the actual item parameters for an actual multiple choice certification test that had flawed items were examined. After correcting the flawed items and re-calibrating all flawed items, the values of the *a* and *b* parameters were lower than the original calibrated values; the *c* parameters were unchanged. Therefore, if an item was mis-keyed, the corrected item then became less discriminating and easier than in the original calibration. In this study, the values of the *a* parameters were changed to 0.2 and the values of the *b* parameters were changed to -3.0 for the flawed items. The reason for choosing these values for the *a* and *b* parameters was to simulate a situation when a flawed item was re-keyed and it then became less discriminating and easier than for the original calibration.

In this study, some items in the item pool were designated as flawed items. Two methods of selecting such items were implemented. One method was to identify a set of most frequently exposed items from a CAT simulation. This method was used to mimic a situation where a set of best items in the item pool were found to be flawed. The second method was to randomly select a number of items. To investigate the effects of the proportion of the flawed items, 1% and 10% of the item pool, that is, 5 or 48 items were selected as flawed items. For those selected items, the values of the *a* and *b* parameters were changed to 0.2 and -3.0, respectively.

The effects of the flawed items were examined when these items were given at the beginning or end of a CAT administration. This study investigated the effects of one or three flawed items at the beginning or at the end of a CAT.

Strategies of Dealing with Flawed Items

Two procedures of handling the flawed items were investigated in this study. One method was to delete the flawed item from the final estimation of examinees' test performance. The other method was to re-score the flawed item with the corrected item parameters (i.e., a =0.2 and b = -3.0). For the investigation of the position effects, the flawed items were deleted from the estimation of examinees' test performance.

Study Conditions

Methods of selecting flawed items (i.e., choosing most often exposed items or randomly selecting items as flawed items), proportion of flawed items (i.e., 1% or 10% of the item pool), procedures of dealing with the flawed items (i.e., deleting or re-scoring), and methods of ability estimation (i.e., MLE or EAP) were crossed, which resulted in 16 conditions. In addition, four conditions were included in the investigation of the effects of positioning of the flawed items (i.e., 1 or 3 items at the beginning or end of a CAT administration and two ability estimation methods). Thus, a total of 20 conditions were studied.

Data Analyses

As indicated above, 1,000 examinees were simulated at each of the 21 θ points. In the CAT administration, an examinee's ability estimation was re-estimated when a flawed item was administered to this simulee by one of the two strategies described above (i.e., deleting or re-scoring). The mean differences and mean absolute differences between the original θ estimate and the re-estimated θ were examined. The original θ estimates and the re-estimated θ were also transformed to a raw score scale of the base form (i.e., one of the test forms in the item pool was designated as the base form). The mean differences and mean absolute differences between the original θ estimates and the re-estimated θ were also computed on the raw score scale (still conditional on the θ). The overall mean differences, variance of mean differences, mean absolute differences, and variances of absolute differences were also summarized.

Results

The results on the conditional indices are summarized in figures. The global indices which are the weighted average of the conditional indices over a standard normal distribution are summarized in tables. Figures 1 through 6 contain the plots of the conditional indices for the MLE method, whereas Figures 7 through 12 contain plots of the conditional indices for the EAP method. For the two strategies of dealing with flawed items (deleting and rescoring) and the analyses of positioning of the flawed items, two figures are presented for each of these three sets of analyses, with one for the conditional indices on the θ scale and the other for the conditional indices on the raw score scale. Each of the figures contains two plots, one for the mean differences of the

estimated and re-estimated scores and the other for the mean of the absolute differences. The variances for these differences are not plotted, but their averaged values are contained in the tables.

From Figure 1, we can see that the mean differences for the deleting strategy fluctuate around the zero line, which means there is no apparent systematic increases or decreases in their estimates after re-estimation. There seems to be no clear pattern of effect of the ways of planting the flawed items. The proportion of the flawed items in the pools also does not seem to affect the mean differences. From the plots of the mean absolute differences, there is clear pattern of effects of the item pool conditions. With a higher proportion (10%) of flawed items, the mean absolute differences are larger than with a smaller proportion (1%) of flawed items. The most exposed flawed items seem to produce larger mean absolute differences in the middle ability range than at tails, whereas the randomly selected flawed items seem to produce larger absolute differences at the tails than at the middle ability levels with one exception that the small proportion (1%) of randomly selected flawed items produce almost zero absolute difference at the higher ability level. An examination of the number of flawed items received by the simulees indicate that with such a small number of flawed items, the simulees at high ability levels did not receive any flawed items.

Figure 2 gives the plots for the differences on the raw score scale. The basic pattern of comparisons are similar to that shown in Figure 1 except that the non-linear transformation of the scale causes the effects at the middle ability levels to be inflated.

Figures 3 and 4 give the plots of differences for the rescoring strategy. The most striking observations on these plots are the large abrupt differences at the ability range of 1 to 3 on the θ scale for 3 of the 4 item pool conditions. The one exception is the pool with small proportion of randomly selected flawed items. Again, because there is basically no flawed item received by simulees at this ability level, the differences are mostly zero. A further examination of the plots show that the pool with a small proportion of the most exposed flawed items displays the largest differences at this ability range. It is unclear why this particular ability range has this abrupt large difference and why the pool with a small proportion of flawed items has even larger differences than the pools with large proportion of flawed items at this ability range. Based on previous theoretical and empirical research, it is generally true the MLE is more vulnerable to unusual response patterns such as answering easy items wrong but answering hard items right than the Bayesian methods. It may be that the change of the item parameters and the regeneration of the response patterns may cause some unusual response patterns. At other ability ranges, the patterns of differences seem to be similar with those of the deleting strategy.

Figures 5 and 6 contain the plots for the flawed item positioning analyses. In the positioning analyses, the scores were re-estimated after deleting certain item(s) at a fixed position(s). It is similar to the deleting strategy discussed above except that there are no designated flawed items. The plots show that the mean differences for all but one of the four conditions fluctuate randomly around the zero line. The one exception is the three beginning item condition where there is a decrease in score at the lower end and an increase of score at the higher end. This direction in differences is consistent with the

bias of the MLE estimates under usual CAT conditions. A possible reason for this difference may be that the bias of MLE is increased due to shortening of the test length by 3 items. The mean of absolute differences display a rather clear pattern. The conditions with 3 flawed items appear to have larger absolute differences than the conditions with 1 flawed item. The flawed item(s) at end position seem to cause larger absolute differences than flawed item(s) at the beginning item(s), in contrast to the Potenza and Stocking (1997) findings (however, because the studies differ in methodology, the results are not strictly comparable). This result is probably due to the fact that ending items are more accurately targeted at the simulees' true ability levels. Again, the transformation to the raw score scale causes the shape of these plot to change, with the differences at the middle ability range being inflated whereas the differences at the two tails being deflated.

Figures 7 through 12 show that with EAP as the ability estimation method, the mean differences for all these three sets of analyses all display a consistent pattern. At the lower part of the ability scale, there are increases in score re-estimates, whereas at the upper end of the scale, there are decreases in the re-estimates. A plausible explanation for this pattern of results is that with the re-estimation procedures, the prior of the EAP tend to play a larger role in the posterior distribution while the likelihood play a lesser role. This is clearly understandable in the deleting strategy and the positioning analyses because some response data were simply deleted in the re-estimation. With the rescoring strategy, even though the flawed items are not simply deleting, the change of the item parameters to be very easy and less discriminating make the likelihood corresponding to these data smaller and thus play a lesser role in the posterior distribution.

With the EAP estimates, the mean of the absolute differences do not seem to change much from those with the MLE estimates. There are only some small decreases at the two tails of the ability scale.

One interesting result for the rescoring strategy is that the abruptly large differences at the mid-high ability range disappeared with the EAP estimates. It is suspected that because the Bayesian methods are more robust to unusual response patterns, that abruptly large differences were alleviated.

Tables 1 and 2 give the overall values for the difference indices and their variances. For the MLE estimates, the rescoring strategy has larger overall mean differences and absolute differences than the deleting strategy, probably due to the large difference at the mid-high ability range. With the EAP estimates, the two strategies have nearly the same overall mean differences and absolute mean differences. Higher proportion of flawed items produced larger overall mean absolute differences, but not necessarily larger overall mean differences.

Table 3 contains the overall mean and the variance of the numbers of flawed items received by a population with standard normal distribution for each of the conditions. It also contains the proportions of simulees who received at least one flawed items. It is clear that with high proportion of flawed items in the pool, simulees received many more flawed items and a larger proportion of simulees received at least one flawed item than with the low proportion of the flawed items. Also, with the most exposed items

being flawed, simulees received many more flawed items and a larger proportion of simulees received at least one flawed items than with randomly selected items being flawed.

Conclusions and Discussions

This study investigated with simulations two strategies for dealing with flawed items and the positioning of the flawed items. In summary, the results show that with the MLE estimates, the deleting strategy seems to produce more smooth results and does not systematically inflate or deflate the score estimates. The rescoring strategy seems to produce some unstable results, causing the estimates to change dramatically at a certain ability range. This phenomenon should be further investigated with more replications and with some change in the conditions. Before more light is shed on this issue, the rescoring strategy should be avoided when used with the MLE estimates. The deleting strategy seems to be a desirable option when used with MLE estimates. With the EAP estimates, on the other hand, there does not seem to be much differences between these two strategies. However, with the EAP estimates, all these conditions produce some systematic inflation for lower ability examines and deflation for the higher ability examines. The positioning analyses seem to suggest that ending items seem to have more effects on the differences than the beginning items.

References

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement*, *11*, 279-290.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software, Inc.

Potenza, Maria T., & Stocking, Martha L. (1977). Flawed items in computerized adaptive testing. *Journal of Educational Measurement.,34*, 79-96.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hamilton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Appendix

Specifications for the ACT Assessment Mathematics Test

Items are classified according to five content categories. These categories and the approximate proportion of the test devoted to each are given below.

Content Area	Proportion of Test	Number of Items
Pre-Algebra and Elementary Algebra	.40	24
Intermediate Algebra and Coordinate Geometry	.30	18
Plane Geometry	.23	14
Trigonometry	.07	4
Total	**1.00**	**60**

a. **Pre-Algebra.** Items in this category are based on operations with whole numbers, decimals, fractions, and integers. They also may require the solution of linear equations in one variable.

b. **Elementary Algebra.** Items in this category are based on operations with algebraic expressions. The most advanced topic in this category is the solution of quadratic equations by factoring.

c. **Intermediate Algebra and Coordinate Geometry.** Items in this category are based on graphing in the standard coordinate plane or on other topics from intermediate algebra such as operations with integer exponents, radical expressions and rational expressions, the quadratic formula, linear inequalities in one variable, and systems of two linear equations in two variables.

d. **Plane Geometry.** Items in this category are based on the properties and relations of plane figures.

e. **Trigonometry.** Items in this category are based on right triangle trigonometry, graphs of the trigonometric functions, and basic trigonometric identities.