Impact of Scoring Options for Not Reached Items in CAT

Qing Yi
Heru Widiatmo
Bradley A. Hanson
Jae-Chun Ban
Deborah J. Harris

ACT, Inc.

Paper Presented at the Annual Meeting of National Council on Measurement in Education, April 11-13, Seattle, WA.

Abstract

The purpose of this study was to evaluate the effects of different scoring options for not reached items in computerized adaptive tests (CATs) on examinees' score estimates. Error indices (bias, standard error of estimation, and root mean squared error) resulting from four scoring options for not reached items in CATs were examined. A condition in which an examinee responded to all the items administered was used as the baseline of comparison. The effects of the number of not reached items were also examined. The results showed that as the number of not reached items increased, so did the differences between the error indices obtained from the scoring options and the baseline condition. The error indices resulting from the scoring option in which the final score estimates were taken as the score estimates at the point where not reached items occurred were most similar to the error indices of the baseline condition.

Impact of Scoring Options for Not Reached Items in CAT

Not reached items are defined as the items an examinee does not respond to following the last item the examinee does respond to. In an ideal computerized adaptive testing (CAT) situation, examinees would have ample time to provide a response to every item of a test. In reality, however, the fact that tests are timed can cause some examinees not to answer all the items of a test for various reasons. Some examinees may indeed work slowly and need more time to answer items. Other examinees may simply stop working on the test after answering several items before the allocated testing time expires. The second group of examinees may try to use certain test taking strategies to inflate their scores. There is more bias in ability estimates for short adaptive tests than for long tests (e.g., Anderson & Richardson, 1979; Lord, 1983; Warm, 1989; Weiss & McBride, 1984). For example, with the maximum likelihood estimation (MLE) method, examinees may potentially get a very high score by answering only a few items correct and not answering the other items on a test. Bayesian estimation methods contain a bias that pulls the estimate toward the mean of the prior; examinees may obtain a score at or slightly below the mean by responding a few items.

Policies are required to govern the decisions of how to score not reached items. Generally, raw scores on paper-and-pencil (P&P) tests are based on the number-correct (NC) score or NC score corrected for guessing. Thus, unanswered items are usually scored wrong. In P&P tests, a fixed set of items are administered to examinees on a particular test date, thus, only the number of not reached items is different among examinees who have not finished the tests. In CATs, however, not only the number of not reached items are different for different examinees, items administered may also be different because of the characteristics of CAT (e.g., Wainer, 2000), where each examinee may potentially take a unique test. The method of scoring not reached items for P&P tests may not be applicable to score not reached items in CATs. Therefore, it is necessary to develop scoring procedures to provide a score for examinees who do not complete the whole test when CAT is used as the test delivery mode. Research is needed to evaluate the effects of different scoring options for not reached items on examinees' score estimates in CAT.

A penalty procedure for scoring not reached items developed by Segall (1988) is used on CAT-ASVAB tests. The penalty procedure provides a final score that is equivalent to the score obtained by randomly guessing on not reached items. The size of the penalty for different test length, tests, and ability levels is attained through simulations (Segall, Moreno, Bloxom, & Hetter, 1997). Slater and Schaeffer (1996) evaluated two scoring options for not reached items in the computerized adaptive GRE general test. The not reached items were either scored as wrong or were assigned a random probability of answering an item correctly. They compared the mean reported score differences between the score at the point of completing 80% of the test and the score at the end of the test where all items after the 80% point were either scored as wrong or given random responses. They discovered that there were more mean reported score differences for scoring not reached items all wrong than for providing random guessing to

unanswered items. However, the standard deviations of the reported score differences were similar for the two scoring options. Slater and Schaeffer (1996) indicated that the large standard deviations might suggest that examinees who have similar ability might be penalized differently as a result of using one of these two scoring options.

Although previous research has evaluated certain scoring options for not reached items in CAT, there are additional options to be examined. Thus, the purpose of the current study was to compare and evaluate the effects of several scoring options for not reached items on score estimates in CAT.

Method and Data

This study used simulation methods. Item responses (0/1s) of random equivalent groups of examinees to seven forms of a large-scale achievement test were used to calibrate item parameters. Sixty discrete five-option multiple choice items are included in each test form. Three content areas were considered for content balancing in CAT item selection, with content area one consisting of 40% of the test, and content areas two and three each comprising 30% of the test. Item parameters were calibrated using the three-parameter logistic (3-PL) item response theory (IRT) model by BILOG computer program (Mislevy & Bock, 1990). The *a*-parameters have a mean of 0.965 with a standard deviation of 0.289, and range from 0.296 to 1.933. The mean of the *b*-parameters is 0.183, the standard deviation is 0.966, and the range is from –3.099 to 2.582. The *c*-parameters have a mean of 0.150 with 0.047 as standard deviation, and range from 0.031 to 0.282. The calibrated item parameters were treated as "truth" for item selection, item response generation, and ability estimation in a simulated CAT. Simulees' true abilities were conditioned on 21 equally spaced points on the θ scale from -4 to 4 in increments of 0.4. At each of the θ points, 1,000 simulees were generated.

Simulated CATs started with an initial ability estimate of zero for all the simulees. Items were selected from the designated content areas based on the Fisher information at the current $\hat{\theta}$ level using the Sympson-Hetter (Sympson & Hetter, 1985) method to control item exposure rates. The maximum item exposure rate was controlled at or under 0.15. Item exposure control parameters were obtained through a series of simulated CATs administered to 1000 simulees from a standard normal distribution. Content balancing was achieved through administering a pre-specified percentage of items from each content area (Kingsbury & Zara, 1989). Forty percent of the items were administered from content area one, and 30% from the content areas two and three, respectively. MLE was used as the final ability estimation method while Owen's (1975) Bayes was employed to provide provisional ability estimates.

Simulees' CAT item responses (0/1s) were determined by comparing the probability of a correct response based on the 3-PL IRT model with a uniform random number. If the probability was less than the random number, the examinee received an incorrect response (i.e., 0); otherwise, a correct response (1) was given. The probability of an examinee getting an item correct based on the 3-PL IRT model is

$$P(\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-D a_i (\theta_j - b_i)}} \qquad (1)$$

where

$P(\theta_j)$ represents the probability that an examinee *j* with ability level θ answers item *i* correctly,

a_i is the item discrimination parameter,

b_i is the item difficulty parameter,

c_i is the lower asymptote of item characteristic curve and represents the probability of an examinee with low ability answering an item correctly, and

D is a scaling factor equals to 1.7.

Fixed length CATs of 30 items were simulated for the condition in which examinees provided responses to all the items in a test. This condition served as the baseline to which the different scoring options were compared. For conditions where not reached items occurred, the number of not reached items ranged from one to ten. Not reached items were taken from the baseline condition of the same simulee. That is, out of the 30 items administered to an simulee in the baseline condition, the last one item, two, three, …, or ten items were taken as the not reached items; and scoring options were applied to those not reached items to obtain a final score estimate.

Four different scoring options for not reached items in CAT were examined in this study:

1. The first option was to take the $\hat{\theta}$ at the point where not reached items occurred as the final ability estimate; this option tried to simulate a situation in which examinees are not penalized for not reaching the last item of a test;

2. The second option was to score not reached items as all wrong and obtained a θ estimate; this option was designed to examine if using the same strategy as that of a P&P test to score not reached items in CAT is applicable. However, this option cannot be used in practice only included for comparison purposes;

3. The third option was to select the next administered item based on the last $\hat{\theta}$ obtained from items reached, then assigned a random probability (i.e., 0.2; five-option multiple choice items) of getting the selected item correctly to attain a provisional $\hat{\theta}$, based on this $\hat{\theta}$ another item was selected and provisional $\hat{\theta}$ was obtained. This process continued until the end of the test (30 items) was reached and a final $\hat{\theta}$ was attained; this option was to investigate if it is possible to generate random item responses based on an examinee's $\hat{\theta}$ obtained from items reached, and estimate a final $\hat{\theta}$ according those generated responses; and

4. The fourth option was to transform the $\hat{\theta}$ obtained from the point where not reached items occurred to the NC score metric, and multiply this score by the percentage of the test examinees had answered (e.g., 66.7%, 70%, 76.7%, and so forth, which is 100 times the number of items reached divided by the total number of items); this option is similar to scoring option one, however, examinees are penalized for not completing a test; the fewer items examinees answered, the bigger the penalty.

The third scoring option actually selected items to use in place of the not reached items and used simulated random responses to those items to determine a score, while the other three options only used the reached items, or in the case of scoring option two, took not reached items from the baseline conditions. The third option also tried to simulate a situation in which examinees run out of time at the end of the test and randomly guessed on the rest of the test. Except the first option, the other three options all administered a penalty to those examinees who did not reach the end of the test; the fewer items the examinees completed, the larger the penalty.

Forty-one conditions were included in this study. One baseline (i.e., examinees answered all 30 items) and forty experimental conditions (i.e., $1, 2, 3, \ldots, 10$ not reached items $\times$ 4 scoring options). Final ability estimates ($\hat{\theta}$) obtained from all the conditions were transformed into the NC score metric. To transform a $\hat{\theta}$ to an NC score estimate, the test characteristic curve (TCC) of the base form (i.e., one of the seven test forms was designated as a base form; note that the base form has 60 items) was used. Under the IRT model, the estimated NC score on the base form associated with a given $\hat{\theta}$ can be found from the TCC as follows:

$$\hat{\tau} = \sum_{i=1}^{60} P_i(\hat{\theta}), \tag{2}$$

where $P_i(\hat{\theta})$ is the probability of correctly answering item *i* given $\hat{\theta}$. An examinee's true NC score (τ) is obtained by substituting the examinee's true θ for $\hat{\theta}$ in Equation 2. Error indices between examinees' true NC score (τ) and estimated NC score ($\hat{\tau}$), that is, bias, standard error of estimation (SE), and root mean squared error (RMSE) were computed at each of the 21 true score levels:

$$Bias(\hat{\tau} \mid \tau) = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{\tau}_j - \tau), \tag{3}$$

$$SE(\hat{\tau} \mid \tau) = \sqrt{\frac{1}{1000} \sum_{j=1}^{1000} \left(\hat{\tau}_j - \frac{\sum_{k=1}^{1000} \hat{\tau}_k}{1000} \right)^2}, \tag{4}$$

$$RMSE(\hat{\tau} \mid \tau) = \sqrt{\frac{1}{1000} \sum_{j=1}^{1000} (\hat{\tau}_j - \tau)^2}. \tag{5}$$

The error indices obtained from different scoring options were compared to the errors resulted from the baseline condition.

Results

Figures 1 to 3 present the error indices obtained from the baseline condition and different scoring options with various numbers of not reached items.

Bias

Figure 1 shows the bias estimates of the baseline condition and of the different scoring options with one to ten not reached items. As the number of not reached items increased, so did the difference between the bias obtained from the baseline condition and from the scoring options. The bias estimates obtained from the scoring option one (i.e., taking the $\hat{\theta}$ from the place where not reached items occurred as the final $\hat{\theta}$) were similar to the bias attained from the baseline condition, especially when the number of not reached items were less than seven (i.e., examinees answered at least 24 items out of the 30). When the number of not reached items increased, bias of the scoring option one began to diverge from that of the baseline condition, especially at the two ends of the score scale (more difference in the bias was observed at the high end; the largest difference was over 3.5 points when examinees only completed 20 items out of the 30). Bias resulting from the scoring option two (i.e., score not reached items as wrong) was most different from the baseline bias at the lower end of the score scale when examinees answered at least 23 items out of the 30 (i.e., not reached items were less than eight), and across the whole score scale when examinees finished less than 23 items out of 30 (i.e., not reached items were larger than seven). The bias obtained from the scoring option four (i.e., taking the $\hat{\theta}$ from scoring option one, transformed it to NC score, and then multiplying it with the percentage of the test an examinee had reached) was the most different from the baseline bias at the higher end of the score scale when examinees completed at least 23 items out of the 30 (i.e., the numbers of not reached items were less than seven). The bias resulted from scoring option three (i.e., generating random response based on the $\hat{\theta}$ from the reached items to obtain a final $\hat{\theta}$) was close to that of the baseline condition when the numbers of not reached items were small (e.g., less than four). For examinees who only reached less than 25 items out of 30 (i.e., not reached items were more than five), those examinees who have higher ability were penalized more by scoring options two, three, and four than examinees who have lower ability.

SE

Figure 2 displays the standard error of estimation (SE) obtained from the baseline condition and from different scoring options with various numbers of not reached items. The difference between SE obtained from the baseline condition and from different scoring options tended to increase as the number of not reached items increased. However, the SE of scoring option one was similar to the SE of the baseline condition, especially when the number of not reached items was less than nine (finishing at least 22 items out of the 30). SE from scoring option four also was similar to the SE of the baseline condition when the number of not reached items was less than six (completing at least 25 items out of the 30). The difference between the SE of scoring option three and the baseline condition was the largest at the middle of the score

scale across all numbers of not reached items. The SE of scoring option two also differed from the SE of the baseline condition, especially at the middle of the score scale.

RMSE

Figure 3 presents the root mean squared error (RMSE) from the baseline condition and different scoring options. Similar to what was observed in Figures 1 and 2, the number of not reached items affected the discrepancy between the RMSE resulting from scoring options two, three, and four and the RMSE from the baseline condition. When the number of not reached items was greater than five, this discrepancy became more noticeable, especially when NC score was larger than 20. On the other hand, the RMSE of scoring option one was closest to the RMSE of the baseline condition.

Discussion

More and more testing programs are offering CATs. In an ideal testing situation, examinees should be given ample time to answer every item in a test and should be motivated to do so. In practice, however, tests are timed so there are always some examinees who do not reach the end of a test for various reasons. The rules for scoring not reached items for P&P tests may not be applicable to CATs due to different characteristics of these two test administration modes. In P&P tests, examinees are administered a fixed set of items, and not reached items are typically scored as wrong. In CATs, there is not a fixed set of items administered to examinees and we do not know which items are not reached items if an examinee does not finish the test. However, a scoring algorithm is still needed to provide a score to those examinees who do not finish the test. Different procedures may be used in CAT to score not reached items, however, the effectiveness of those procedures have not been fully investigated. Thus, it is important to evaluate the effects of scoring options for not reached items in CAT on score estimates. This paper compared four scoring options for not reached items in CATs.

The study results indicated that errors obtained from the scoring option one were the closest to those of the baseline condition. The number of not reached items affected the differences between the errors of the various scoring options and the baseline condition. The higher the number of not reached items, the larger the differences. It seems that scoring option one probably was the best approach to handle not reached items in the simulated CATs of the current study.

This study did not consider the effects of time on score estimates. For a fixed time test, examinees who do not finish the test may spend more time on items they provide responses to, however, examinees who do finish the test have to spread the allotted time over all the items. Wang and Hanson (2001) have been developing an item response model that incorporates response time into ability estimation. For future studies, it may be used to incorporate a model that considers response time into the investigation of scoring options for not reached items in CAT. As indicated above, scoring option one appeared to have the errors closest to those of the baseline condition. However, this option is still affected by the number of not reached items. Thus, it has the potential to be used by some examinees to their advantage, especially if they only respond a few items. Scoring option four does take the percentage of test examinees have completed into account, the fewer items examinees reached, the more penalty this scoring option

provides. However, the results from this study indicated scoring option four had error indices that were quite different from those of the baseline condition. Approaches of employing scoring option one to score not reached item but preventing some examinees of using certain strategies to inflate their scores need to be investigated in future research.

References

Anderson, J. A., & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics, 21,* 71-78.

Kingsbury, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2(4),* 359-375.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-form reliability. *Psychometrika, 48,* 233-245.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* [Computer program]. Chicago, IL: Scientific Software.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70(350),* 351-356.

Segall, D. O. (1988). *A procedure for scoring incomplete adaptive tests in high stakes testing.* Unpublished manuscript. San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). *Psychometric procedures for administering CAT-ASVAB.* In W. A. Sands, B. K. Waters, and J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 131-140). Washington, DC: American Psychological Association.

Slater, S. C., & Schaeffer, G. A. (1996, April). *Computing scores for incomplete GRE general computer adaptive tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceeding of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H. (2000). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T., & Hanson, B. (2001, April). *Development and calibration of an item response model that incorporate response time.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

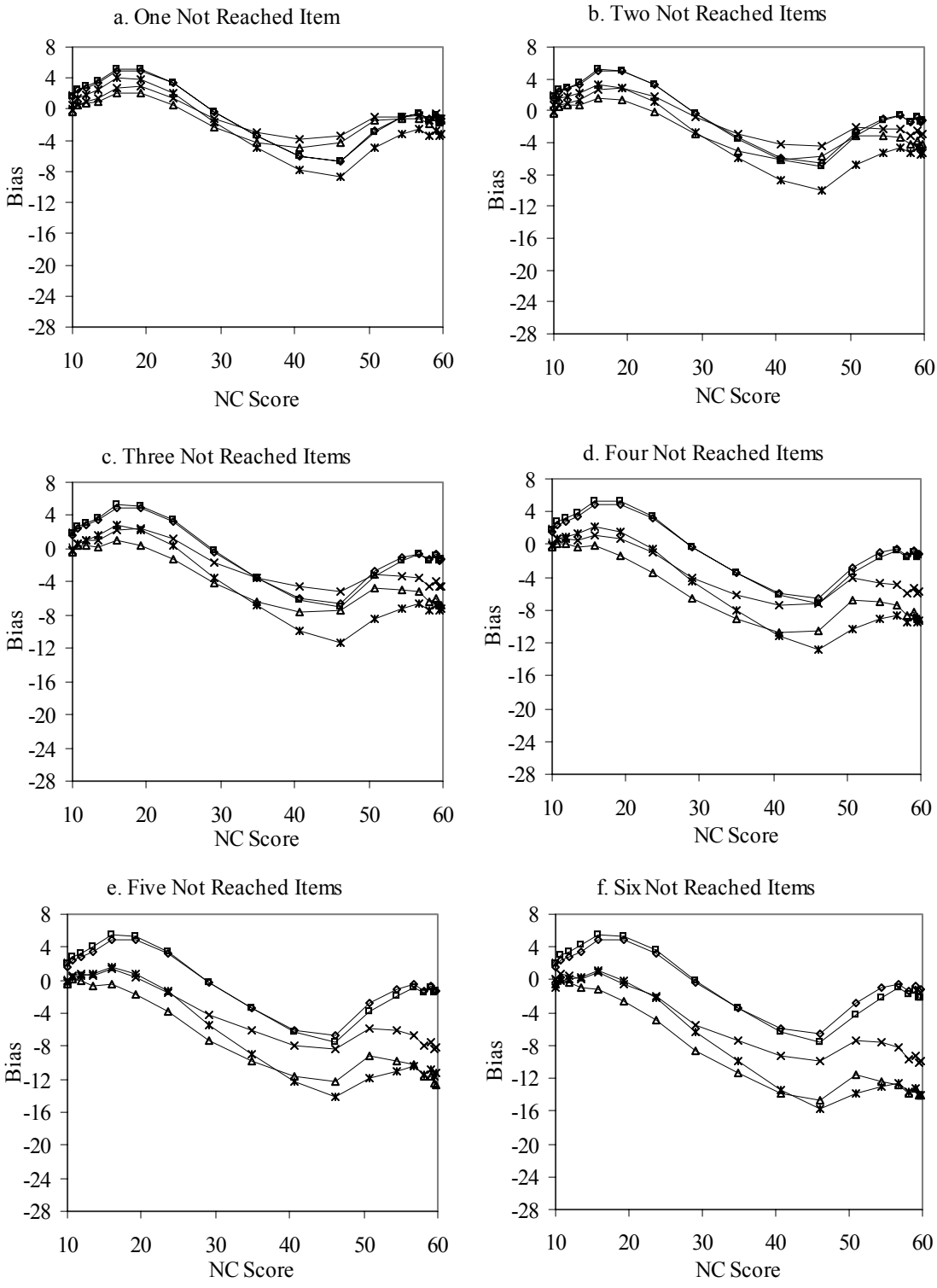

Figure 1. Bias obtained from baseline condition and different scoring options for not reached items.

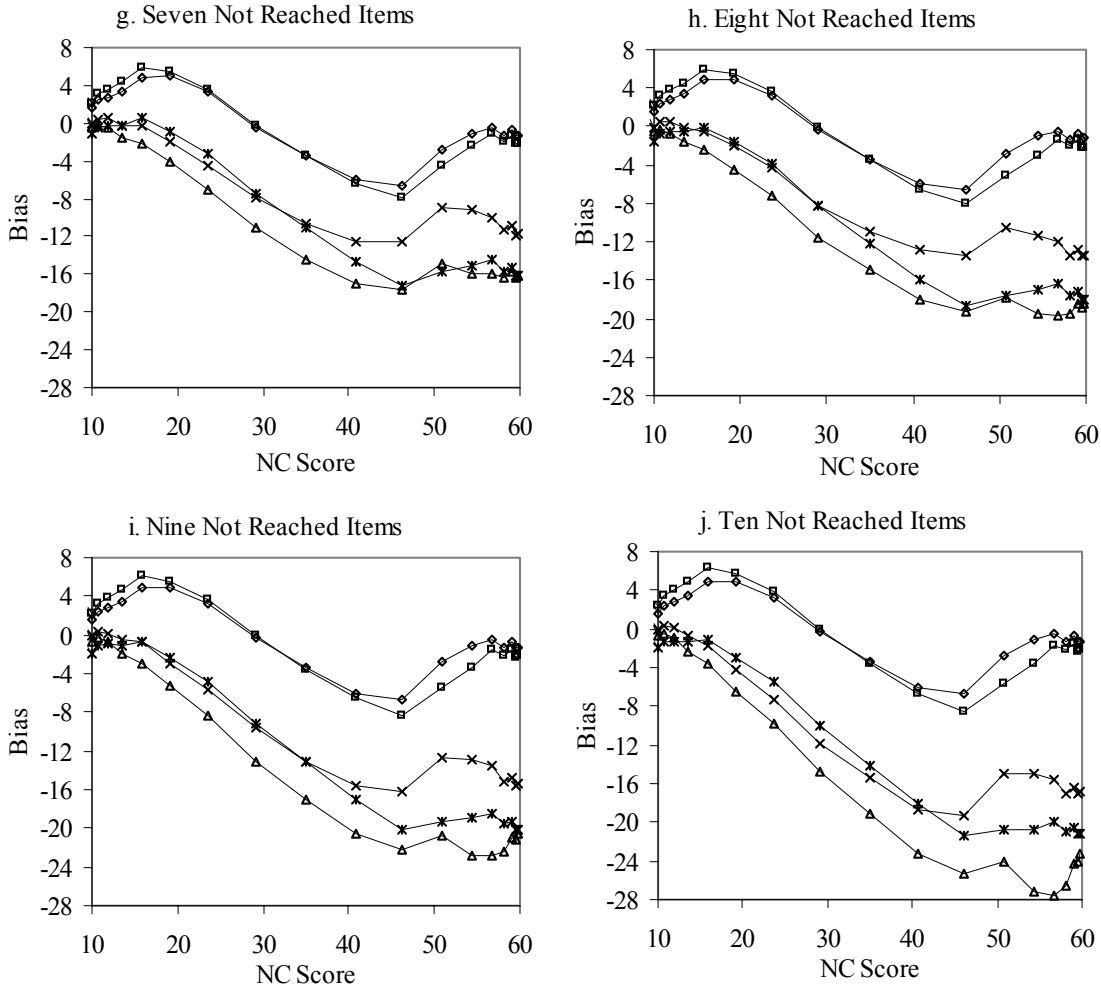

Figure 1 Con't.

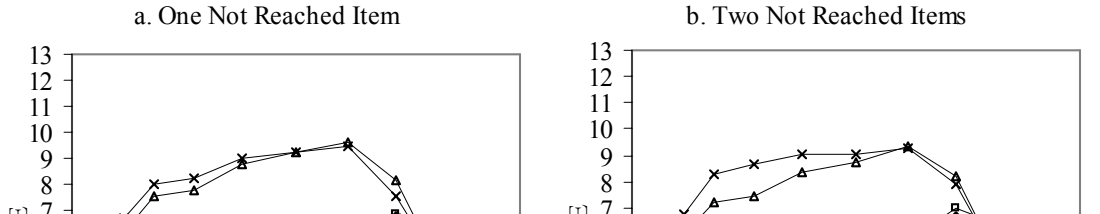

Figure 2. SE obtained from baseline condition and different scoring options for not reached items.

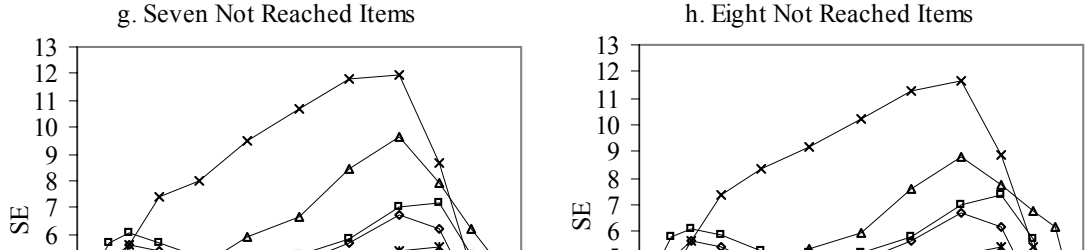

Figure 2 Con't.

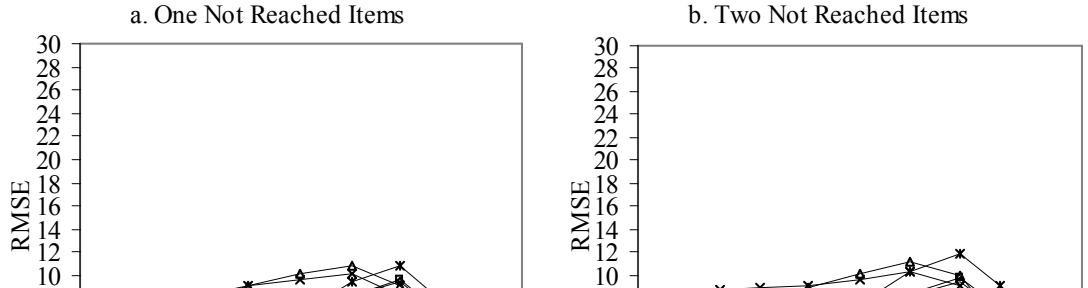

Figure 3. RMSE obtained from baseline condition and different scoring options for not reached items.

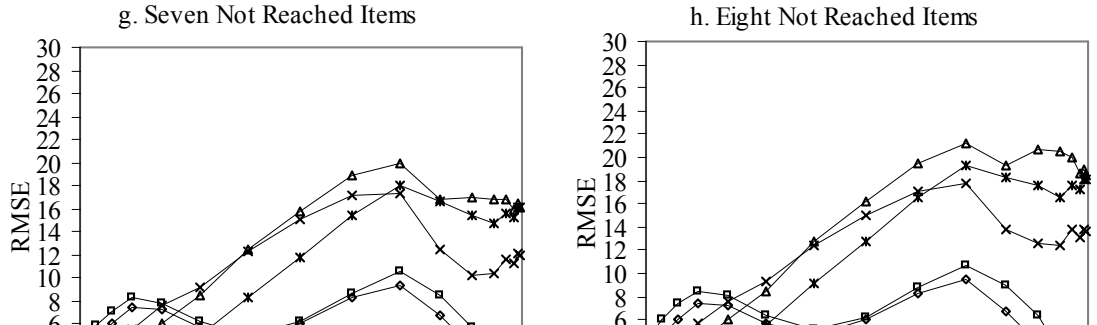

Figure 3 Con't.