The Effect of Review on Student Ability and Test Efficiency for Computerized Adaptive Tests

Mary E. Lunz and Betty A. Bergstrom, American Society of Clinical Pathologists

Benjamin D. Wright, University of Chicago

The effect of reviewing items and altering responses on the efficiency of computerized adaptive tests and the resultant ability estimates of examinees were explored. 220 students were randomly assigned to a review condition; their test instructions indicated that each item must be answered when presented, but that the responses could be reviewed and altered at the end of the test. A sample of 492 students did not have the opportunity to review and alter responses. Within the review condition, examinee ability estimates before and after review were correlated .98. The average efficiency of the test was decreased by 1% after review. Approximately 32% of the examinees improved their ability estimates after review, but did not change their pass/fail status. Disallowing review on adaptive tests administered under these rules is not supported by these data. *Index terms: adaptive testing, computerized adaptive testing, Rasch model, relative efficiency, test information.*

The purpose of educational measurement is to facilitate educational decision making by providing estimates of an individual's knowledge or skill. For certification and licensure, this means making minimum competency pass/fail decisions. In recent years, computers have become more versatile and more accepted for the development and delivery of examinations. One of the most interesting and potentially advantageous methods for test developers and examinees is computerized adaptive testing (CAT). The algorithms for item selection usually depend on item response theory (IRT) (Lord & Novick, 1968; Rasch 1960/1980; Wright, 1977; Wright & Stone, 1979). The items in the test item bank are calibrated to a common scale on which the pass/fail point has been established. The adaptive algorithm selects items that provide the most information about examinee ability given the current ability estimate from responses to the previous items. Because the items administered are tailored to the performance of the examinee, fewer items are needed to reach a decision with the specified level of confidence (Green, Bock, Humphreys, Linn, & Reckase, 1984). CATs frequently vary in length because the stopping rules require a specified level of precision (measurement error) or level of confidence (distance from the pass point) in the accuracy of the decision, rather than a fixed number of items as in paper-and-pencil tests (Weiss, 1985).

In the usual CAT administration procedure, a starting point (typically at the middle of the scale or at the pass point), is selected and an item of that difficulty is presented. If the examinee answers the item correctly, a more difficult item is presented. If the examinee answers that item incorrectly, an easier item is presented. Items with difficulties near the current ability estimate of the examinee continue to be presented, so that maximum information is gained from each item (Weiss & Kingsbury, 1984) until the stopping rule is met.

When the difficulty of an item matches the ability of an examinee, the examinee has a 50% probability of answering correctly. When the item is targeted slightly below the ability of the examinee, the probability of answering the item correctly is greater than 50%. A stopping rule based on confidence in the accuracy of the pass/fail decision requires that an examinee's ability estimate be a specified interval above or below the pass/fail point, usually 1.3 to 1.65

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 33-40

0146-6216/92/010033-08$1.65

standard errors (90% to 95% confidence). A confidence level stopping rule produces more precise ability estimates for examinees with estimated abilities near the pass/fail point because they take longer tests. Examinees with abilities distant from the pass/fail point take shorter tests and consequently have larger errors even though the pass/fail decisions are made with the same level of confidence (e.g., 90% or 95%) (Bergstrom & Lunz, 1991). Each examinee takes a test that measures their ability to the level of precision needed to decide whether they have passed or failed.

The opportunity to review items and alter responses generally has not been allowed when tests are administered adaptively. One concern has been that reviewing and altering item responses may change the estimate of examinee ability such that the sequence of items will become poorly targeted and precision will be lost. This assumes, however, that the response alterations made by the examinee will substantially increase the distance between the current examinee ability estimate and the previously selected item difficulty. This assumption has not been verified with real data.

For example, if an examinee changes several sequential incorrect responses to correct responses, the current ability estimate would increase and all subsequent items could be too easy, thus reducing the information about the examinee gained from the subsequent items. However, systematic increase or decrease in ability estimates as a result of item response review has not been demonstrated.

The opportunity to review items and alter responses is usually important to examinees. Examinees feel at a disadvantage when they cannot review and alter their responses. In addition, there is the fear of entering an incorrect answer accidentally and not being able to change it. The security of a final review can provide comfort and reassurance to the examinee who is taking a test that impacts an important decision about an examinee.

Allowing examinees to review a CAT also has implications for the perceived fairness of the examination process. An examination may appear to be fairer to an examinee when review and alteration of responses is allowed. Under these circumstances, examinees have a final opportunity to demonstrate their knowledge by checking for entry errors and double checking uncertain responses.

The effect of reviewing and altering responses on the results of CATs has not been studied in detail. Wise, Barnes, Harvey, and Plake (1989) found no significant differences in the mean scores of groups of examinees in review and nonreview conditions. However, the number of responses altered and the direction of the change (correct to incorrect, incorrect to correct, or incorrect to incorrect) has not been studied. If patterns of response during review improve or reduce ability estimates systematically, the information value of subsequent items could be altered and the pass/fail decision could become less precise.

The effect of reviewing items and altering responses on the ability estimates of examinees and the efficiency of CAT was explored. Efficiency refers to the amount of information gained from each item administered. It was hypothesized that the opportunity to review items and to alter responses would not alter significantly either the efficiency of the test or the examinee estimates and resulting pass/fail decisions. It was also expected that the information value of the items presented by the CAT algorithm would remain substantially unaffected by response alterations.

Method

Item Precalibration

This study was implemented in two phases. In the first phase, a database of items was constructed and field tested using a paper-and-pencil examination administered to a national sample of students from medical technology programs. This database was designed to meet the test specifications for the traditional written certification examination. The items were calibrated

using the one-parameter logistic IRT model (Rasch 1960/1980; Wright, 1977; Wright & Stone, 1979).

The fit of the items to the Rasch model was verified by examining the "infit" statistic (the mean of the standardized squared residual weighted by its variance) for the calibrated items (Wright & Masters, 1982, pp. 94–105). For each person/item encounter, the observed response was compared to the modeled expected response. Items that did not fit were removed before the item bank was established. When data fit the model, the infit statistic has a value near 0 and a standard deviation (SD) near 1.0. For the 726 items in the bank, the mean item infit was .045 with a SD of 1.014.

Data Collection

In Phase 2 the calibrated item bank was used to construct CATs. Usable CAT data were obtained from 712 students—83% were white and 81% were female, which is the typical population mix for the certification examination. Students participated in the study because it was part of the review for their certification examination.

CAT Algorithm

The CAT model was an adaptive mastery test (Weiss & Kingsbury, 1984), designed to determine whether an examinee's estimated ability level was above or below a preestablished criterion. Kingsbury and Houser (1990) have shown that an adaptive testing procedure that provides maximum information about an examinee's ability will designate an examinee as above or below a pass/fail point more clearly than a test that peaks the information at the pass/fail point.

The CAT ADMINISTRATOR program (Gershon, 1989) was used to construct CATs according to the test specifications of the paper-and-pencil certification examination (see Table 1). The item with the most appropriate level of difficulty within a given subtest was presented to the examinee. In the first 50 items, blocks of 10 items were administered from subsets 1 through 4, and blocks of 5 items were administered from subsets 5 and 6. After the first 50 items were administered, blocks of 4 items (subsets 1 through 4) and blocks of 2 items (subsets 5 and 6) were administered. Subset order was selected randomly by the computer algorithm, because Maurelli and Weiss (1983) found subtest order to have no effect on the psychometric properties of an achievement test battery. The minimum number of items was set at 50 and covered all 6 content areas according to the established test plan. The number of items administered varied, but the maximum was set at 240.

Items were selected at random from unused items within .10 logits of the targeted item difficulty within the specified content area. While the examinee considered the item presented, the computer selected two items, one that would yield maximum information should the current item be answered incorrectly and another that would yield maximum information should the current

Table 1
Test Plan Distribution (TPD), Number of Items, and Item Difficulty Data for the Medical Technology Item Bank

Subtest	TPD	Number of Items in Bank	Item Difficulty			
			Lowest Difficulty	Mean	Highest Difficulty	SD
Microbiology	20%	147	-2.89	-.06	2.38	.96
Blood Banking	20%	165	-2.21	-.07	2.94	1.00
Chemistry	20%	142	-3.61	-.07	2.97	1.06
Hematology	20%	135	-2.80	-.05	2.97	.97
Body Fluids	10%	72	-2.24	-.09	3.84	.97
Immunology	10%	65	-2.78	.25	2.04	.96
Total Test	100%	726	-3.61	-.02	3.84	1.00

item be answered correctly. This procedure insured that there was no lag time before the next item was presented.

The stopping rule required the measured ability of examinees to be 1.3 times the standard error of measurement (Wright & Masters, 1982) above or below the pass point before testing stopped. This stopping rule provided 90% confidence (one-tailed test) in the pass/fail decision. The pass point was set at .15 logits on the bank scale. This was a slightly more rigorous standard than that used for the certification examination.

Test Conditions

Two test conditions were used in this study: a review condition ($N = 220$) and a nonreview condition ($N = 492$); students were randomly assigned to conditions. Examinees in both conditions took a CAT that required them to answer each item when it was presented. The directions for the review condition informed examinees that items must be answered when presented, but that review and alteration of responses would be allowed after the test was complete. Items were presented until the test was completed; instructions for the review procedure were then presented. During review, items were presented in the same order as they appeared in the test and the examinee's initial response was highlighted. Examinees could devote as much time to reviewing and changing answers as desired.

The directions for the nonreview condition specified that examinees must answer each item when presented. Examinees were cautioned to consider their answers carefully because they would have only one opportunity to answer the item.

Two test records were maintained for each examinee in the review condition: a test record before review and one after review. The first record contained the item sequence numbers, item difficulties, examinee responses, current ability estimates, and errors of measurement for the CAT before review. The second record contained response changes made during review, and ability and error of measurement reestimates based on the revised item responses. Comparison of these records indicated the number and accuracy of response changes made during review. From these data, ability and error of measurement estimates and reestimates were compared to ascertain the loss of information and hence loss of test precision caused by response alteration.

Test Efficiency

Test efficiency was determined by the amount of information gained from each item administered. Maximally efficient tests produce the most information about examinee ability with the smallest number of items. The efficiency of CATs depends on targeting items to the examinee's ability so that a pass/fail decision at the specified level of confidence is reached with fewer items. Off-target items provide less information and therefore make longer tests necessary.

Calculation of relative test efficiency depends on the following:

b = examinee ability
d = item difficulty
$p = [\exp(b - d)]/[1 + \exp(b - d)]$ = probability of correctly responding to an item
$q = p(1 - p)$ information value for an item
$\text{SEM} = (1/\sum q)^{1/2}$ standard error of measurement
$I = \sum q$ test information value

To analyze the impact of review, the components of test efficiency were applied as follows:

b_1 = ability estimate before review
b_2 = ability estimate after review
$b_2 - b_1$ = amount of change in measured ability due to review
SEM_1 = standard error of measurement before review
SEM_2 = standard error of measurement after review
$(\text{SEM}_2/\text{SEM}_1)^2$ = relative efficiency of the test after review
I_1 = test information before review
I_2 = test information after review
$I_2 - I_1$ = loss of information due to review
$N_{12} = 4(I_2 - I_1)$ = number of perfectly

targeted items required to recover information loss

The information gained from an item depends on the probability of a correct response. The test information (I) is the sum of the information contributed by the items ($I = \Sigma q$). As each item is administered, more information is acquired about the examinee. If a substantial number of items become "off target" as a result of changes in the examinee's current measured ability due to altering responses during review, the information value may change ($I_2 - I_1$). When an item is perfectly targeted, an examinee has a .5 probability of a correct response. Perfectly targeted items have an information value of .25 logits [$q = p(1 - p)$]. Thus the number of additional items that would have to be presented to an examinee to compensate for information loss can be calculated as $4(I_2 - I_1)$.

The ratio of the squared measurement errors before and after review indicates the relative test efficiency after review. Efficiency depends on the number of items near enough to the examinee's ability to make a tangible contribution to the ability estimate.

Results

Comparison Between Groups

The means and SDs of ability estimates (in logits) were M = .24 and SD = .48 for the review group (after review) and M = .16 and SD = .50 for the nonreview group. These means were significantly different ($t = -2.08$, 710 degrees of freedom, $p < .04$). The examinees who were allowed to review performed significantly, but only slightly (.08 logits), better on average than the examinees who were not allowed to review.

Review Group Results

Table 2 summarizes the response alterations during review. The average number of items altered was 2; the minimum was 0 and the maximum was 16 (an average of 96 items were administered). 85 examinees did not alter responses even though the opportunity to do so was available. 135 examinees altered some responses during review. Of these examinees, 30 obtained lower estimates after review, 71 obtained higher estimates after review, and 34 effected no change in their ability estimates as a result of altering

Table 2
Response Alterations During Review

Response Alterations	Number	Percent
Items		
Range of Items Administered	50-240	
Mean No. of Items Administered	96	
Mean No. of Items Altered	2	
Minimum No. of Items Altered	0	
Maximum No. of Items Altered	16	
Responses		
Total Responses	21,082	100%
Responses Altered	456	2.2%
Incorrect to Correct	214	1.0%
Correct to Incorrect	123	.6%
Incorrect to Incorrect	119	.6%
Examinees	220	100%
Altered Responses	135	61%
Higher Estimate	71	32%
Lower Estimate	30	14%
No Change	34	15%
Did Not Alter Responses	85	39%

responses. The examinee who altered 16 responses changed 9 items from incorrect to correct, 6 from correct to incorrect, and 1 from incorrect to incorrect. His test contained 95 items; therefore, 17% of the items were altered. His pass/fail status did not change. Overall, the percentage of item responses altered during review was small (2%). The patterns of alteration did not usually involve changes to a series of sequential items. Rather, one or two item responses were altered at diverse points in the test.

Table 3 presents the summary of ability estimates and test efficiency for the review group. The mean and SD of the ability estimates before review (b_1) were .216 and .47, respectively; after review, mean b_2 = .243, SD = .48. Mean change was .027 logits. Ability estimates before and after review correlated .98. Figure 1 shows the plot of these data before and after review. The most obvious outlier is an examinee whose ability estimate was 1.00 before review and 1.73 after review. This examinee altered 14 responses—10 from incorrect to correct, 2 from correct to incorrect, and 2 from incorrect to incorrect. The pass/fail status of this candidate was not altered as a result of changing responses.

The relative test efficiency ratio was 99% with a SD of 3%. Thus, the average efficiency of the test decreased by 1% after review. Mean information loss ($I_2 - I_1$) due to review was .11. On average, the information loss could be recovered by the addition of one-half item (N_{12}). Of the 135 examinees who altered responses, 108 would require no additional items, 23 would require 2 to 5 additional items, and 4 examinees would require 6 to 14 additional items to recover the information loss.

Table 3
Mean and SD of Ability Estimates, Relative Test Efficiency, and Information Before and After Review (N = 220)

Variable	Mean	SD
Ability (In Logits)		
b_1	.216	.47
b_2	.243	.48
$b_2 - b_1$	.027	.09
Relative Test Efficiency		
SEM_1	.230	.05
SEM_2	.231	.05
$(SEM_2/SEM_1)^2$	.99	.03
Information		
I_1	22.58	12.48
I_2	22.46	12.55
$I_2 - I_1$	-.11	.51
Number of Items to Recover Information		
N_{12}	.45	2.03

The overall effect of altering responses on the pass/fail decisions was also minimal. Table 4 shows the pass/fail tabulation for the review group. Three examinees altered their status from fail to pass after review. These examinees' ability estimates were within 1 error of measurement of the pass/fail point both before and after review, indicating less than 70% confidence in the accuracy of either decision. The average improvement for these three examinees was .05 logits, the result of changing one response from incorrect to correct. The mean improvement in logits for the various decisions was .04 for pass/pass, .01 for fail/fail, .05 for fail/pass, and .03 overall.

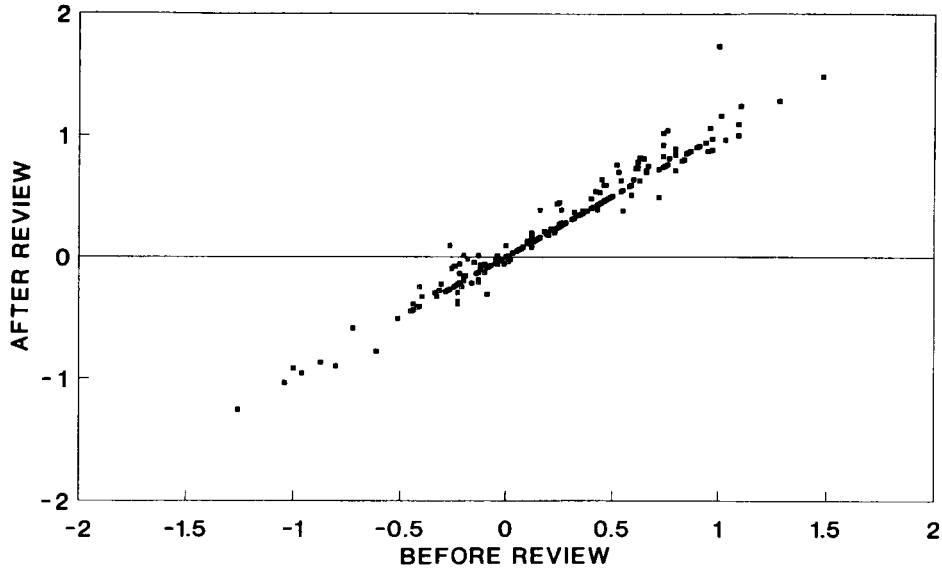
Table 4
Pass/Fail Decisions Before and After Review

	Before Review	
After Review	Pass	Fail
Pass	120	3
Fail	0	97

Discussion

There are two issues involved in allowing examinees taking a CAT to review items and alter responses. The first issue concerns the psychometric implications of including items that provide less than maximum information about the examinee. When the response to an item previously answered correctly or incorrectly is altered, the ability estimate and its standard error are altered, yet the difficulty of the next item administered remains the same. The question is how close to the revised ability estimate the item difficulty must be to maintain the efficiency of the item selection algorithm. In these data, the

Figure 1
Ability Estimates for Examinees in the Review Condition Before and After Review



difficulties of subsequent items were within the related error of measurement of the revised ability estimate, even though some items were altered from incorrect to correct or vice versa. The mean loss of information due to these changes was less than the amount of information provided by one additional perfectly targeted item. The exception was the single examinee who altered 14 items, 10 from incorrect to correct, thus significantly improving his ability estimate (see Figure 1). His passing status, however, did not change.

The second issue is the face validity of CAT. Educators have trained students from elementary school to medical school to review responses after completing a test to catch careless errors and thus render a more accurate demonstration of competence. When adults, who grew up in this system, face a certification examination they usually feel it is their "right" to review their work for careless errors. Indeed, the testing goal is not to penalize an examinee for clerical errors in the name of psychometric integrity, but rather to measure their ability as accurately as possible. These data show that decision accuracy and confidence are comparable before and after review.

A limitation of this study is that the examinees knew that this was not an actual certification examination. They were encouraged to perform their best, but there was no immediate reward attached to acceptable performance. These results, therefore, might not reflect typical examinee review behavior on the certification examination. Examinees may alter more or perhaps even fewer responses. Future research should replicate this study in the context of an actual certification test.

The fact that the group allowed to review performed significantly, but only slightly (.08 logits), better than the randomly equivalent group that was not allowed to review suggests a "greater confidence" hypothesis. That is, the reason for the difference might relate to the familiarity and comfort associated with the knowledge that review and correction of careless errors was allowed. The personal control associated with the opportunity to review may inspire more careful consideration of each item, thus leading to overall improved ability estimates.

References

Bergstrom, B. B., & Lunz, M. E. (1991, April). *Confidence in pass/fail decisions for computer adaptive and paper and pencil examinations.* Paper presented at the annual meeting of the American Educational Research Association, Chicago IL.

Gershon, R. (1989). *CAT administrator* [Computer program]. Chicago: Micro Connections.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Kingsbury, G. G., & Houser, R. L. (1990, April). *Assessing the utility of item response models: Computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Maurelli, V. A., & Weiss, D. J. (1983). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. Expanded edition, University of Chicago Press, 1980.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education, 2,* 235–241.

Wright, B. D (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis.* Chicago: MESA.

Wright, B. D., & Stone, M. (1979). *Best test design.* Chicago: MESA.

Author's Address

Send requests for reprints or further information to Mary E. Lunz, American Society of Clinical Pathologists, 2100 West Harrison Street, Chicago IL 60612, U.S.A.