ED 408 331 TM 026 584

AUTHOR Wise, Steven L.; And Others
TITLE The Accuracy of Examinee Judgments of Relative Item Difficulty: Implications for Computerized Adaptive Testing.
PUB DATE Mar 97
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Gains; *Adaptive Testing; *College Students; *Computer Assisted Testing; *Difficulty Level; Higher Education; *Review (Reexamination); Scores; Statistics; *Test Items
IDENTIFIERS *Answer Changing (Tests)

ABSTRACT

The degree to which item review on a computerized adaptive test (CAT) could be used by examinees to inflate their scores artificially was studied. G. G. Kingsbury (1996) described a strategy in which examinees could use the changes in item difficulty during a CAT to determine which items' answers are incorrect and should be changed during item review. In CAT, a correctly answered item will typically be followed by a more difficult item, while an incorrectly answered item will typically be followed by an easier item. The results of two studies involving groups of 77 and 62 undergraduates suggest that examinees are not highly proficient at discriminating item difficulty, a skill needed for successful application of the Kingsbury strategy. In the third study, which used 243 introductory statistics students, the Kingsbury strategy, which examinees would use only for guessed items, was compared to a generalized strategy used for all sequential item pairs. The Kingsbury strategy yielded a small average score gain, while the generalized strategy yielded an average score loss. These results suggest that only the Kingsbury strategy would enable examinees to inflate their scores successfully. (Contains 2 tables, 3 figures, and 14 references.) (Author/SLD)

The Accuracy of Examinee Judgments of Relative Item Difficulty: Implications for Computerized Adaptive Testing

Steven L. Wise, Sharon A. Freeman, Sara J. Finney, Craig K. Enders, and Donald D. Severance
University of Nebraska-Lincoln

Paper presented at the 1997 annual meeting of the National Council on Measurement in Education, Chicago, IL

Abstract

We examined the degree to which providing item review on a CAT could be used by examinees to artificially inflate their scores. Kingsbury (1996) described a strategy in which examinees could use the changes in item difficulty during a CAT to determine which items' answers are incorrect and should be changed during item review. The results of our first two studies suggest that examinees are not highly proficient at discriminating item difficulty—a skill needed for a successful application of the Kingsbury strategy. In the third study we compared the Kingsbury strategy—which examinees would use only for guessed items—to a generalized strategy used for all sequential item pairs. The Kingsbury strategy yielded a small average score gain, whereas the generalized strategy yielded an average score loss. These results suggest that only the Kingsbury strategy would enable examinees to successfully inflate their test scores.

The Accuracy of Examinee Judgments of Relative Item Difficulty: Implications for Computerized Adaptive Testing

There is a debate in the measurement community regarding whether or not examinees should be provided an opportunity to review, and possibly change, their answers on a computerized adaptive test (CAT). Arguments in favor of item review have two basic themes. First, examinees have consistently expressed a strong desire for item review (Baghi, Ferrara, & Gabrys, 1992; Legg & Buhr, 1992; Vispoel, Forte, and Boo, 1996; Vispoel, Rocklin, & Wang, 1994; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). Second, there is substantial evidence that when examinees change answers on multiple-choice tests, they are likely to improve their scores (see Wise, 1996). By implication, denying item review on a CAT would disallow answer changing, which would tend to have a negative effect on test performance.

There are several arguments, however, against provision of item review on a CAT. One is decreased efficiency; providing item review would both markedly lengthen testing time and increase the standard errors of proficiency estimation. Another argument concerns the possibility that some administered items may provide clues to the correct answers of other items. An examinee who recognizes such clues could use item review to change his or her answer to the clued item(s).

A third argument against providing item review during a CAT is that examinees could strategically use item review to artificially inflate their test scores. Wainer (1993) described one such strategy, in which examinees would intentionally fail each item to consequently receive an easier item, and then go back during item review and provide the correct answers to this relatively easy test in order to attain a high final score. The research on Wainer's strategy, however, has indicated that it would not be a highly attractive strategy for examinees to use when taking a CAT scores that provides item review. Wise (1996) provided a summary and discussion of this research.

Kingsbury (1996) described a potentially more serious strategy. His strategy is based on an examinee's monitoring of changes in the relative difficulty of successive test items. In a CAT, a correctly answered item will typically be followed by a more difficult item, while an incorrectly answered item will typically be followed by a less difficult item. This is a source of information that could be exploited by an examinee during item review. After answering a given item, if the examinee could discern that the succeeding item was more difficult, then he or she would know that the answer to the previous item was correct. If the succeeding item was less difficult, then the answer to the previous item was incorrect, and the examinee would know that his or her answer to the previous item should be changed during review. Thus, the feedback provided by changes in item difficulty could inform an examinee which answers to leave alone, and which answers to change.

Kingsbury (1996) conducted a real data simulation of the strategy, finding that it yielded (a) substantial gains in estimated proficiency for examinees of low true proficiency, (b) modest gains for examinees of moderate proficiency, and (c) virtually no gains for examinees of high proficiency. This suggests that the strategy would be beneficial to the test performances of many examinees if they were provided item review in a CAT.

Kingsbury's simulation was based, however, on two key assumptions. First, the examinees were assumed to have guessed the answer to any item whose difficulty parameter was more than 1.0 theta units above their final proficiency estimates. Second, if the difficulty of the item succeeding a "guessed" item was at least .50 theta units less difficult than the "guessed" item, then the examinee changed his or her answer during the simulated review process.

Kingsbury acknowledged that these assumptions were open to question and that, under different assumptions, the results may change. In particular, he noted that actual examinees are likely to be inconsistent and make mistakes employing the strategy—an outcome that may lower their proficiency estimates.

In the Kingsbury (1996) strategy, an examinee would consider changing the answers to only those items to which he or she had initially guessed. This strategy could be generalized, however, to consider all sequential item pairs in a test. In this alternative strategy, which we termed the Generalized Kingsbury (GK) strategy, an examinee would make difficulty judgments for all item pairs—whether they involved guessed answers or not—and change the answer to an item whenever he or she judged that the succeeding item was easier.

An important issue regarding the Kingsbury and GK strategies—and a focus of this investigation—concerns the degree to which examinees would be able to accurately discern whether the difficulty levels of successive items increased or decreased. Success in employing the strategy is largely dependent on an examinee's being able to accurately make such difficulty discriminations. Note also that, as a CAT administration proceeds, finer discriminations would be required by the examinee, as the changes in difficulty between successive items decrease in size. Therefore, an examinee who cannot make accurate difficulty discriminations is unlikely to successfully use the Kingsbury or GK strategies to substantially improve his or her score. In fact, for such examinees, many correct answers are likely to be changed, which may result in lowered proficiency estimates.

In reviewing the research literature, we found only one study in which examinees were asked to discriminate between the difficulties of test items. Green (1983) compared subjective judgments of item difficulty with empiricallybased difficulty values, using the method of paired comparisons to establish the subjective difficulty ranking of 10 multiple-choice items. As part of this process, subjects were given each of the 45 item pairs and asked to identify the more difficult item of each pair. For each subject, the resultant subjective difficulty rankings were compared to the empirical rankings using Goodman and Kruskal's (1954) gamma. The mean value of gamma (across subjects) was .20, which suggested that subjects were not highly proficient at judging relative item difficulty.

The purposes of the present investigation were to (a) provide further evidence of the degree to which examinees can discriminate between item difficulty levels of achievement test items and (b) compare the effects of the Kingsbury and GK strategies on test performance. The results of three studies are reported, which should provide a realistic assessment of the likelihood that examinees could use the information from item difficulty changes to artificially inflate their test scores.

Study 1

Study 1 was designed to assess how well examinees can discriminate the relative difficulty of achievement test items when their sole task was difficulty discrimination. The procedure used was similar to that used by Green (1983); students were provided pairs of test items and asked to identify the more difficult item of each pair.

Method

Test Materials. The test items were drawn from a disclosed 60-item form of the ACT Mathematics test. Item difficulty parameters were obtained by fitting the three-parameter logistic IRT model to the item responses from a sample of 1000 high school students who had previously taken the ACT Mathematics test as an operational form. Each of the ACT items had five response options per item.

A subset of 31 items were selected, from which 30 sequential pairs were formed (items 1 & 2, 2 & 3, and so on). We used three criteria in selecting the items and their order. First, the difficulty parameters for each pair of items had to differ by at least the average of their standard errors. Second, the item pairs were selected to exhibit a range of differences between the difficulty parameters; 10 pairs differed by less than .50, 10 pairs differed by between .50 and 1.0, and the remainder differed by more than 1.0. Third, the item pairs were chosen to have a range of average difficulty; for half of the item pairs the average difficulty parameter was less than .40. Each combination of difficulty difference and average difficulty was represented by 5 of the 30 item pairs.

The items were photocopied (with item numbers deleted) from the ACT test booklet and each pair of items was printed, one item above the other, on a separate 8.5" x 11" page. The top and bottom items of each pair were labeled "A" and "B", respectively, and the upper right-hand corner of each page indicated the item pair number. The pages were then collated and spiral bound with an opaque cardstock cover to form the test booklet used in the study.

We developed a separate answer sheet on which students indicated their difficulty judgments. Two judgments were requested for each item pair. First, the student was asked, "Which item is more difficult (A or B)?". Next, the student was asked, "How confident are you with your choice?", and provided four rating options ranging from not at all confident to highly confident.

Participants. The study participants were 77 undergraduates enrolled in an educational psychology course at a large midwestern university. The students provided informed consent and received research credit for their participation.

As an additional incentive, the three students who made the greatest number of correct difficulty judgments were each paid $10.00.

Procedure. Examinees were tested in groups ranging in size from 4 to 15. They were given brief instructions regarding the purpose of the study and informed of the $10.00 payment for the best performers, after which the following testing materials were distributed: test booklets, answer sheets, scratch paper, and pencils. Additional instructions were then given regarding the item difficulty comparison task. The examinees were told that, for each item pair, their task was to identify which of the items most examinees would find harder. They were then directed to judge and record their responses for a given item pair before continuing to the next pair; moreover, they were told not to review and/or change their answers to previously judged pairs. No time limit was imposed, and examinees were encouraged to perform any calculations that they felt would be helpful in identifying which item of a pair was more difficult.

As a measure of mathematics achievement, ACT Mathematics scores were obtained from university records for each examinee. These scores were available for 64 of the 77 examinees.

Results

A total judgment score for each examinee was computed as the number of correct difficulty discrimination judgments made across the 30 item pairs. The distribution of judgment scores, shown in Figure 1, ranged from 9 (30% correct) to 24 (80%), with a mean of 17.42 (57%). The mean judgment score was only slightly higher than 15, which would be the expected score under random guessing, and about a fourth of the examinees scored 15 or lower. In addition, the judgment scores were found to be unrelated to the ACT mathematics scores ($\underline{r}$ (62) = .00, p>.05).

Figure 1: Distribution of Difficulty Discrimination Performance in Study 1

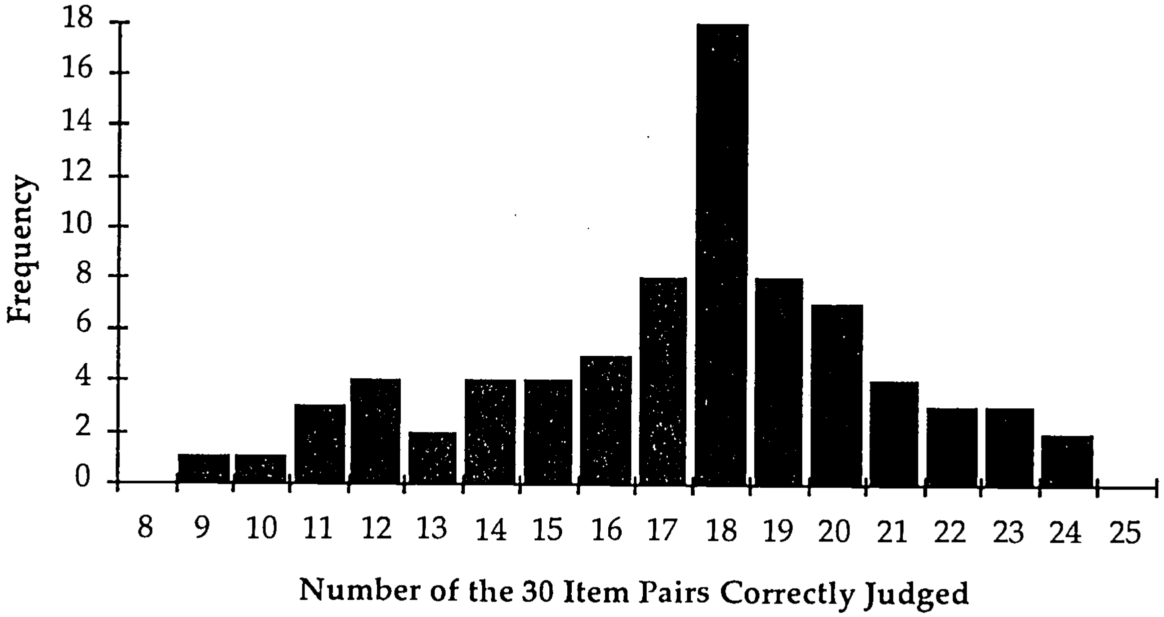


Table 1

Percentages of Correct Difficulty Discrimination Judgments, by Difficulty Difference and Mean Difficulty, in Study 1

Item Pair Difficulty Difference	Mean Item Pair Difficulty Less than .40	 Greater than .40
Less than .50	54	51
.50 – 1.00	54	62
Greater than 1.00	55	68

Table 1 shows the percentages of correct judgments for item pairs by level of difficulty difference and mean difficulty. For the easier item pairs (mean difficulty less than .40), the percentage of correct judgments was slightly above the chance level of 50, for all levels of difficulty differences. For the harder item pairs, the percentage of correct judgments increased with degree of difficulty difference. Of greatest relevance to the assumption made in Kingsbury's (1996) simulation was the finding that, for item pairs whose difficulties differed by at least .50, only 60% of the difficulty judgments were correct.

Student judgment confidence was not found to be related to judgment accuracy. Across examinees and item pairs, the mean confidence rating for correct and incorrect judgments was identical (M = 3.03). In addition, the correlation between the mean confidence (across students) and the percentage of correct judgments for each item was nonsignificant (r (28) = .16, p>.05).

Discussion

This study was intended to administer the difficulty discrimination task as simply as possible. Both items of each pair were displayed together, and the examinees did not have to divide their attention between identifying the correct answer for each item and making difficulty judgments. The results indicated that the examinees performed relatively poorly in their difficulty judgments. For items differing by at least .50 theta units in difficulty, the 60% success rate at which students judged the item pairs fell far short of the 100% rate that was assumed in the Kingsbury (1996) simulation.

Two limitations of this study should be noted. First, the examinees made their difficulty judgments under conditions that were nonconsequential, apart from the monetary incentive for the top performers. Therefore, it is difficult to assess the degree of effort expended by the examinees. Second, although examinees were invited to work through the math items in making their difficulty judgments, they were not required to do so. If the Kingsbury or GK strategies were being employed by examinees taking a CAT, they would be required to both answer items and judge item difficulty. It is possible that the act of answering test items provides additional difficulty information that could yield more accurate difficulty judgments. This possibility was investigated in the second study.

Study 2

Although Study 1 was a straightforward study of examinees' proficiency in discriminating item difficulty, it differs in two important ways from the testing experience of an examinee who is attempting to use the Kingsbury or GK strategy during a CAT. First, examinees administered a CAT are shown only one item at a time; they would have to compare the difficulty of the displayed item with their memories of the difficulty of the previous item. Second, in a CAT, an examinee's attention would be divided between identifying the correct answer for each item and comparing its difficulty with that of the previous item. Each of these reasons would probably influence examinees' success using the Kingsbury or GK strategies. Study 2 was designed to more closely approximate the conditions under which an examinee being administered a CAT would have to make item difficulty judgments.

Method

Computer-Based Test. This study used the same 31 ACT mathematics items that were used in Study 1. The items were administered as a fixed-order test via IBM-compatible microcomputers using MicroCAT test administration software (Assessment Systems Corporation, 1988). Each item was administered individually, and examinees had to choose an answer for each item before the next item was displayed. Beginning with the second item, after an answer was chosen for an item, the examinee was asked whether the displayed item was more or less difficult than the previous item. This resulted in difficulty judgments being made for the same 30 item pairs as in Study 1. Examinees were not permitted to review either their answers to the items or their difficulty judgments.

Participants. The study participants were 62 undergraduates enrolled in an educational psychology course at a large midwestern university. The students received research credit for their participation. As an additional incentive, $10.00 was paid to each of the three students who made the greatest number of correct difficulty judgments, as well as to the three students who received the highest scores on the computer-based test.

Procedure. Testing was conducted at a university computer laboratory. The examinees, who were tested in groups ranging in size from 4 to 8, were provided brief instructions by the test administrators concerning the nature of the study. The examinees then completed the computer-based test, which included additional instructions regarding the difficulty judgments that would be made during the test. Each examinee's responses—answers and difficulty judgments—were stored by the testing software. As in Study 1, no time limits was imposed, and examinees were provided scratch paper and pencils to use in their calculations.

Results

The distribution of difficulty judgment scores, shown in Figure 2, ranged from 11 to 25, with a mean of 19.71 (66%). The distribution of total achievement scores on the 31-item test ($\underline{M}$ = 16.73, $\underline{SD}$ = 6.13) provided evidence that the examinees directed considerable effort toward answering the test items. The judgment scores were significantly related to the achievement scores ($\underline{r}(60) = .32$, $\underline{p} < .05$), indicating that the higher achieving examinees made more correct difficulty judgments.

Figure 2: Distribution of Difficulty Discrimination Performance in Study 2

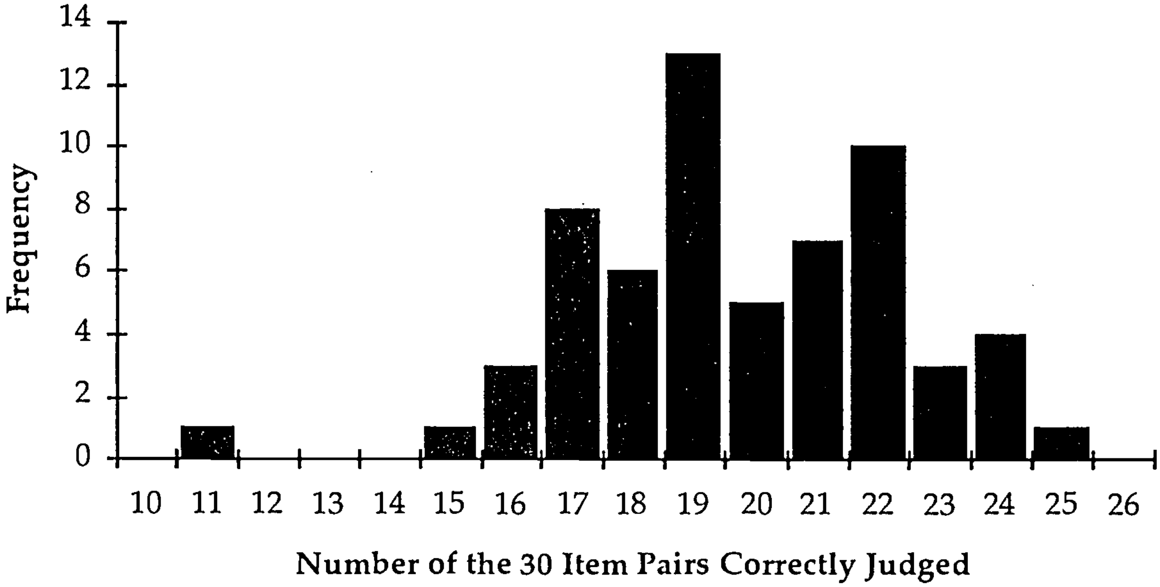


Table 2

Percentages of Correct Difficulty Discrimination Judgments, by Difficulty Difference and Mean Difficulty, in Study 2

Item Pair Difficulty Difference	Mean Item Pair Difficulty Less than .40	 Greater than .40
Less than .50	53	50
.50 – 1.00	61	68
Greater than 1.00	77	85

The percentages of correct difficulty judgments, broken down by item difficulty difference and mean pair difficulty, are shown in Table 2. As was found in Study 1, examinees generally made better difficulty judgments for the more difficult item pairs. Not surprising, moreover, was the finding that difficulty judgment accuracy increased with item difficulty difference. For item pairs whose difficulty parameters differed by at least .50, examinees made correct difficulty judgments 73% of the time.

Discussion

The examinees in Study 2 exhibited markedly better performance than those in Study 1 on the difficulty discrimination task. This performance difference is likely due to the Study 2 examinees being required to try to solve each item; this activity apparently led to more accurate difficulty judgments. Nevertheless, average difficulty discrimination performance again was well below the 100% level assumed in Kingsbury's (1996) simulation.

Study 3

The purpose of the third study was to study the effects of the Kingsbury and GK strategies on test performance when examinees make imperfect difficulty judgments. A real data simulation of both strategies was conducted in a manner similar to that used by Kingsbury (1996).

Method

The data used in the simulations was based on a sample of 243 response records that were selected from introductory statistics students who were administered a 20-item CAT at a large midwestern university during 1995. This CAT, developed to measure student proficiency in the algebra skills needed in an introductory statistics course, provided proficiency estimates that were calculated using a maximum-likelihood procedure. Each multiple-choice item contained four response options. Details regarding the item pool and the CAT procedures can be found in Wise, Plake, Johnson, and Roos (1992).

Simulation Assumptions. The assumptions made were extensions of those made by Kingsbury (1996) that reflected examinees making imperfect difficulty judgments. For the Kingsbury strategy, the following assumptions were based on the difficulty judgment accuracy estimate from Study 2:

1. An examinee was assumed to have guessed the answer to any item whose difficulty parameter was more than 1.0 theta units above his or her final proficiency estimate.
2. If the difficulty parameter of the item succeeding a guessed item was at least .50 theta units lower (easier), then an examinee had a .73 probability of correctly judging this difference and consequently changing his or her answer to the guessed item. This probability is based on our findings in Study 2 that examinees were able to correctly judge the relative difficulties of 73% of the items whose difficulty parameters differed by at least .50. When an answer to a guessed item was changed, an examinee had a .33 chance of passing the item, as he or she was assumed to randomly guess among the remaining three options.
3. If the difficulty parameter of the item succeeding a guessed item was at least .50 theta units higher (harder), then an examinee had a .27 (i.e., 1 – .73) probability of incorrectly judging that the succeeding item was easier and consequently changing his or her answer to a guessed item that had actually been passed. In this case, any answer change would result in the guessed item being failed.

The assumptions for the GK strategy were essentially the same as those made in simulating the Kingsbury strategy, with the exception that an examinee judged and potentially changed answers for all sequential item pairs, not just those in which the first item was guessed.

Results and Discussion

The changes in proficiency estimates resulting from the Kingsbury and GK strategies are shown in Figures 3 and 4, respectively. For the Kingsbury strategy, the changes were relatively small ($\underline{M}$ = .01, $\underline{SD}$ = .02), ranging from -.08 to .16. For the GK strategy, the changes were more variable ($\underline{M}$ = -.03, $\underline{SD}$ = .18), and ranged from -.53 to .78. Thus, the Kingsbury strategy yielded a small mean gain in estimated proficiency, while the GK strategy yielded a small mean loss.

The changes in estimated proficiency for the Kingsbury strategy in our study were generally smaller in magnitude than those found by Kingsbury (1996). This is probably due both to differences in test length and item pool characteristics across the two studies, and to differences in assumptions made regarding examinee difficulty judgment accuracy. Note also that, when judges make imperfect difficulty judgments, the Kingsbury strategy can result in lowered proficiency estimates.

As expected, most of the answer changes tended to be made during the initial items of the CAT. For the Kingsbury strategy, 22 (96%) of the 23 changed answers occurred for the first five items. For the GK strategy, answer changes made during the first five items accounted for 173 (82%) of the 211 answer changes.

General Discussion

This investigation was directed toward assessing the degree to which providing item review on a CAT could be used by examinees to artificially inflate their scores. Although the generalizability of our findings are possibly limited by our use of only mathematics items in the three studies, some tentative conclusions can be presented.

Figure 3: Changes in Proficiency Estimates Resulting From Use of the Kingsbury Strategy

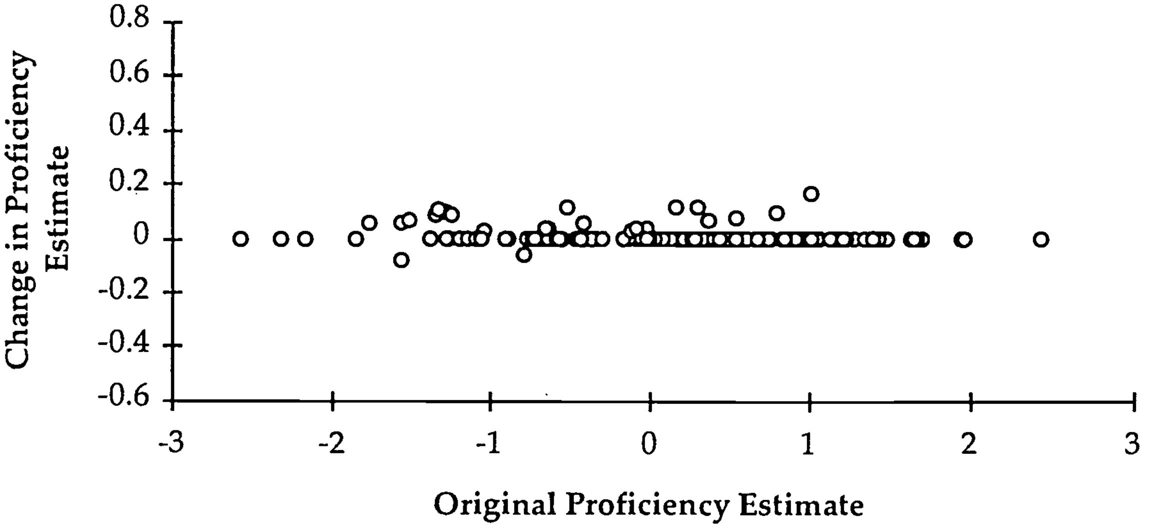
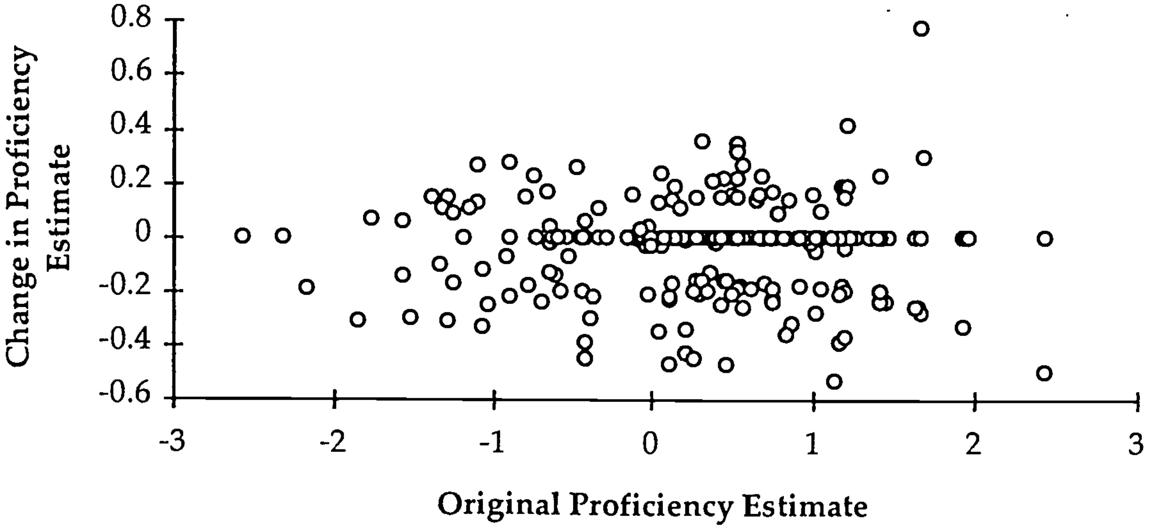


Figure 4: Changes in Proficiency Estimates Resulting From Use of the GK Strategy



First, the results of the first two studies suggest that examinees are not highly proficient at discriminating item difficulty—a skill that would appear necessary for a successful application of either the Kingsbury or GK strategies. Imperfect difficulty judgment accuracy restricts the effectiveness of strategy-directed answer changing in two ways. First, if an examinee cannot correctly recognize that item _x_ is succeeded by an easier item, he or she will miss an opportunity to potentially change the answer to item _x_ from incorrect to correct. Second, and more importantly, if an examinee erroneously judges that item _x_ was succeeded by an easier item, then his or her correct score to item _x_ will be changed to incorrect. Thus, whenever an examinee judges that a succeeding item is easier and an answer change is therefore warranted, the potential score gain resulting from a correct difficulty judgment must be considered relative to the certain score loss resulting from an erroneous judgment.

In Study 3, the Kingsbury and GK strategies were found to have markedly different effects on test performance. The Kingsbury strategy tended to yield small (if any) increases in estimated proficiency. These changes were attributable primarily to answer changes made during the first several items of the test, when a CAT is most likely to administer items that are not closely matched to an examinee's true proficiency level. In contrast, changes in estimated proficiency yielded by the GK strategy tended to be much more variable and, on the average, negative.

The differential effects of the two strategies on test performance can be explained by considering the differences in the percentage of passed items among those for which answer changes are considered during item review. In the Kingsbury strategy, answer changes are considered only for items for which the initial answer was a guess. This implies that an examinee is likely to have passed a proportion of items that does not differ greatly from that expected by chance (e.g., about 20% of a set of five-option multiple-choice items). In the GK strategy, however, an examinee is likely to have passed a much higher percentage of his or her items prior to review. A CAT adjusts item difficulty to match examinee proficiency, which will typically result in an examinee passing about 50-60% of his or her items. Thus, in the GK strategy there proportionately more passed items for which erroneous difficulty judgments could be made—which would result more answers to previously passed items being changed to incorrect.

An example demonstrates this differential effect of the Kingsbury and GK strategies on the numbers of items passed on a hypothetical test. Suppose that an examinee has taken a CAT consisting of 51 multiple-choice items, each with 5 response options. Prior to beginning item review, the examinee has (a) chosen the correct answer for 3 of the 15 items for which guesses were made and (b) chosen the correct answer for 30 of the first 50 items. Moreover, the examinee is capable of making correct difficulty judgments 67% of the time.

If the examinee employs the Kingsbury strategy, what would be the expected result? Of the 9 guessed items that were initially failed, the examinee would be expected to correctly identify 6 (67%) whose answers should be changed. Assuming that the examinee guessed randomly among the 4 remaining options, we would expect that the correct answer would be chosen for 2 of these items. Of the 3 items that were initially passed, we would expect the examinee to erroneously identify 1 (33%) whose answers should be changed. For this item, the answer will be changed from correct to incorrect. Thus, we would expect the incorrect judgments to result in 1 additional item being passed as the result of using the Kingsbury strategy.

In contrast, if the GK strategy is employed by the examinee, we would expect him/her to correctly judge about 14 of the 20 item that were originally failed, and subsequently guess the correct answer to about 3 items. Of the 30 items that were initially passed, the examinee should erroneously change answers to—and consequently fail—10 items. We would therefore expect the GK strategy to result in about 7 fewer items being passed.

Overall, the results of the three studies indicate that the Kingsbury strategy would be an attractive strategy for examinees to use if item review were provided on a CAT. It provides the possibility of a modest increase in an examinee's proficiency estimate, with little risk of a decrease. The GK strategy, however, would be far less attractive. Although it potentially could yield larger increases in estimated proficiency than the Kingsbury strategy, it is more likely to yield lower proficiency estimates.

We should note that although few examinees would likely think of the Kingsbury strategy on their own, they could be readily coached by others to use it. The strategy is relatively simple and straightforward, and examinees could be quickly taught to routinely employ it during the first several items of a CAT—when the changes in item difficulty tend to be the greatest, and therefore more reliably discerned. If used in this fashion, the Kingsbury strategy offers a chance for modest score gains, without requiring an examinee to spend a great deal of time and attention keeping records regarding which answers were guessed and the judged difficulty of the succeeding items.

The results of this investigation have important implications regarding whether or not item review should be provided on a CAT. The potential for examinees to use item review to artificially inflate their performance on a CAT represents a serious threat to the validity of its proficiency estimates. Not providing item review on a CAT, however, also poses a potential threat to score validity. By denying item review—and thereby disallowing answer changing—examinees are denied an opportunity to engage in a test-taking behavior that has consistently been shown likely to improve test performance (Benjamin, Cavell, & Shallenberger, 1984; Waddell & Blankenship, 1995). Thus, there are validity concerns associated with each side of the item review issue. And without clear validity-based reasons for settling this issue, it appears likely that CAT developers' decisions regarding whether or not to provide item review will continue to made on the basis of testing efficiency.

References

Assessment Systems Corporation. (1988). User's manual for the MicroCAT testing system, version 3. St. Paul, MN: Author.

Baghi, H., Ferrara, S. F., & Gabrys, R. (1992, April). Student attitudes toward computer-adaptive test administrations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Benjamin, L. T., Cavell, T. A., & Shallenberger III, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? Teaching of Psychology, 11, 133-141.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732-764.

Green, K. E. (1983). Subjective judgment of multiple-choice item characteristics. Educational and Psychological Measurement, 43, 563-570.

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

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. Educational Measurement: Issues and Practice, 11, 23-27.

Vispoel, W., Forte, E., & Boo, J. (1996, April). Effects of answer review and test anxiety on the psychometric and motivational characteristics of computer-adaptive and self-adaptive vocabulary tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

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

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). How review options and administration modes influence scores on computerized vocabulary tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Waddell, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. The Journal of Continuing Education in Nursing, 25, 155-158.

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

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. Journal of Educational Measurement, 29, 329-339.

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

AERA CONFERENCE 1997

TM026584

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: THE ACCURACY OF EXAMINEE JUDGMENTS OF RELATIVE ITEM DIFFICULTY: IMPLICATIONS FOR COMPUTERIZED ADAPTIVE TESTING

Author(s): STEVEN L. WISE, SHARON A FREEMAN, SARA J. FINNEY, CRAIG K. ENDERS, DONALD D. SEVERANCE

Corporate Source: UNIVERSITY OF NEBRASKA - LINCOLN

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	The sample sticker shown below will be affixed to all **Level 1** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 1**	The sample sticker shown below will be affixed to all **Level 2** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 2**	☐ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title: STEVEN L. WISE

Organization/Address:
DEPT. OF EDUCATIONAL PSYCHOLOGY
UNIV. OF NEBRASKA
LINCOLN, NE 68588-0345

Telephone: 402/472/2736

FAX:

E-Mail Address: SWISE@UNLINFO.UNL.EDU

Date: 3/25/97

(over)