Answer Changing on Multiple-Choice Test Items Among Eighth-Grade Readers

CLIFTON A. CASTEEL
Vernon C. Haynes Middle School
Metairie, Louisiana

ABSTRACT. This study was done to examine the effect of answer changing on multiple-choice test performance among good and poor readers in the eighth grade. Although the gains of poor readers were higher than those of good readers, all subjects profited significantly from changing their answers on items. For all subjects, when a single response was changed, there was a two-to-one chance that the new response would raise rather than lower the final score. Gains from answer changing on test items were slightly higher for poor readers as a group than were those for good readers. However, the result was determined not to be significant. More important, this hypothesis is strengthened by the fact that all subjects profited from answer changing. Therefore, the results were interpreted as lending support to the notion that answer-changing response among young examinees should be encouraged if there is a reasonable doubt about their "first impression."

ACCORDING TO POPULAR BELIEF, an individual's first response when taking a test is the best response, and that answer should not be changed. Thus, students often are told that changing "first impression" (first choice) answers on multiple-choice or true-false tests will probably result in wrong answers and thus produce a lower total test score (Huff, 1961). However, various investigators (Bath, 1967; Berrien, 1939; Costin, 1972; Jacobs, 1972; Jarrett, 1948; McMorris, DeMers, & Schwarz, 1987; McMorris & Weideman, 1986; Mueller & Wasser, 1977; Reile & Briggs, 1952; Reiling & Taylor, 1972; Vidler & Hansen, 1980) have found that most subjects who make one or more changes significantly increase their overall test scores. When taking a true-false test, net gain/loss ratios were reported to range from 1.4:1 to 2.5:1, with a median of 2:1. Net gains for multiple-choice items were slightly higher, with gain/loss ratios ranging from 2.3:1 to 5.3:1 and a median ratio of about 2.7:1.

The present investigation focused on some of the questions raised by earlier research on the answer-changing patterns of good and poor readers

on multiple-choice tests. Some researchers have theorized that poor readers make more changes on objective tests and are more likely to change an answer from correct to incorrect. For example, several researchers found that good readers gain more in answer changing than poor readers (Archer & Pippert, 1962; Crocker & Benson, 1980; Mueller & Schwedel, 1975). However, Berrien (1939) found no consistent correlation between good and poor readers in the amount of revisions made or in the gain made from those revisions. Research published by Crocker and Benson compared younger examinees such as seventh-grade students with college students. They found that although younger examinees made fewer answer changes, the ratio of points gained to points lost because of answer changing was higher for seventh graders than for older examinees.

The purpose of this study was to investigate the answer-changing behavior of young examinees on a multiple-choice test. Specifically, this study investigated the following questions: (a) Which type of reader (good or poor) most often changed his/her answers? and (b) What number of items were changed from right to wrong, wrong to right, and wrong to wrong? Good and poor readers were selected for this study because poor readers seem to lack confidence in their ability to score well on tests, whereas good readers seem to be more positive in their approach to answer changing (Casteel, 1987). If the result of this study proved favorable for both groups, it would be safe to postulate that young examinees would improve their test grades by reviewing answers and making revisions.

An account of the nature of all revisions will be presented first, followed by a test to determine whether the net result of changing responses was a gain or a loss. The results should provide educators with a more prudent and substantial rationale when furnishing subjects with information concerning the skill of test taking.

Method

Subjects

The sample consisted of 53 eighth-grade subjects (34 girls, 19 boys) attending a public middle school. Subjects were judged most likely to be from middle-class families and were predominantly white (87%). The subjects were classified into two groups: good and poor readers. Subjects were primarily classified on the basis of stanine scores obtained on a reading subtest of the Comprehensive Test of Basic Skills (CTBS), which had been administered in the spring of the previous school year. Subjects were classified as poor readers ($n = 27$) if their stanine score on the vocabulary and comprehension reading subtest of the CTBS was 3 or below. Subjects were classified as good readers ($n = 26$) if they had stanine scores of 6 or above. Sub-

jects with stanine scores in the 4 to 5 range were thought of as average readers and unsuited for this investigation. Subjects were also rated by their teachers as either good or poor readers according to such criteria as fluency, oral reading errors, and comprehension. Ages for the subjects ranged from 12 years, 4 months to 14 years, 11 months. No emphasis was placed on sex because similar studies in the past generally reported a nonsignificant difference between the sexes in points gained due to answer changing.

Materials

The Cornell Critical Thinking Test, Level X (grades 5–12) was used in this investigation. The test contained 76 multiple-choice items, all of which were the three-option type.

This test is different in format and attempts to measure skills other than those assessed in previous research, which primarily used teacher-constructed tests. This test involves the skills of inference, induction, deduction, and evaluation, with four types of bases for such inferences: the results of other inferences, observations, statements made by others, and assumption. Correlations with tests of subject matter knowledge, critical thinking ability, and scholastic aptitude range broadly around .50. Reliability estimates range from .67 to .90. Because subjects are asked to think critically to determine the accuracy of the statements, the test appears to require strong deliberation. Therefore, this type of test (also used by Vidler and Hansen [1980]) would appear to require a relatively stronger measure of reflection and mental ability than teacher-made tests.

The students used two types of pencils, red and blue, to distinguish between a first and second impression/choice answer. Subjects recorded answers on IBM scantron sheets. One answer sheet was supplied to each student. This process was monitored by two university instructors who assisted with the investigation.

Procedure and Design

Subjects were unaware that their scores would be used in an investigation or that they would be given a second chance to change their answers. This procedure was predicated on the assumption that any answer that was not changed was a first impression answer and that any answer that was changed was a second impression answer. Previous studies have, in most cases, relied upon erasures on answer sheets to analyze the outcomes of changed responses. However, as Jacobs (1972) and Lowe and Crawford (1929) noted, the limitation of this method is that changes can only be detected when they involve erasure. Thus, there is little probability of detecting those decisions that may precede putting pen to paper. Also, there remains a lack

of control in that subjects may work back and forth among items as they look for and utilize clues from one item to respond to another. Thus, the following method was designed to improve the reliability and validity of data and methods of previous studies.

Subjects were given instructions as to the procedures involved in taking this test. A single question was displayed on clear transparency, using an overhead projector for group viewing. Because most previous studies involved analyses of answer sheets to determine the results of erased responses, working only with overt answer changes, there was no possibility of detecting those first impressions that may not have been put to paper. Using a projector and transparency was believed to create an experimental situation where responses and answers to test items could be more easily identified. After a trial run using a preliminary form of the Cornell Critical Thinking Test, a time frame was placed on each item to prevent test takers from being able to make a second impression. The time allocated for each group was as follows:

For good readers: Items with a total word count of fewer than 20 words were exposed for 30 seconds. Items with a word count of 21 to 35 had an exposure time of 45 seconds; items with 36 or more words were exposed for 55 seconds. Then the next question was displayed, and so on.

For poor readers: An additional 10 seconds was added to each item word count level. For example, items with a word count of fewer than 20 words were given an exposure time of 40 seconds. Thus, subjects classified as poor readers were given ample time to read the test. A previous test involving similar subjects indicated that such time and method adjustments were necessary because these subjects were found to read somewhat slower than their counterparts. Thus, the duration was increased to suit this group.

Subjects were initially given blue pens and a scantron answer sheet and told to answer all items rapidly but as carefully as possible; they were also told that they would be given only one chance to answer the items. Subjects were asked to respond to all items with a first impression answer. This procedure was installed to help prevent clerical errors when attempting to analyze responses made by subjects. For example, if a change was made because an item was initially skipped on an answer sheet but answered on the second try, the change would probably be from wrong to right rather than from right to wrong (McMorris & Weideman, 1986).

The next class period subjects were informed that they were being afforded the opportunity to revise their answers. Red pens, answer sheets, and mimeographed copies of the previously administered test were distributed; this stage did not involve transparencies or the overhead projector. This procedure was similar to that employed in a study by Jacobs (1972). In addition, as with the study by Lynch and Smith (1975), the following instruc-

tions were added: "When you reread your examination, you'll probably be tempted to change some of your answers. If you feel strongly that an answer should be changed, change it. But if you aren't certain about either answer, then don't change your initial answer."

In reviewing the collected data, answers recorded in blue (first impression answer) and red (second impression answer) made the scoring much more precise than if relying on erasure marks, as was the case with most previous studies. It should be mentioned that the first testing procedure took place during the last class of the school day. Subjects then made their revisions the first class period of the next day. Because subjects were not told that they would be given the opportunity to make a second impression, the possibility of exchanging information that could affect the outcome of this study was believed to be very minimal. To minimize the possibility of cheating during testing, an oversized room was used, as well as the presence of several teachers serving as monitors at each testing session. As a final precaution, subjects were well spaced and were assigned seats from one test to the next.

Results

The number of revised answers made by the 53 subjects were analyzed in a 2 (Reader Type: good/poor) × 3 (Response Type: wrong to right, right to wrong, wrong to wrong) ANOVA with the latter as a within-subjects factor. The within-subjects factor indicated the type of revised answer made by each subject: wrong to right, right to wrong, and wrong to wrong (see Table 1). As can be seen, the main effect for reader type, $F(1, 153) = .14$, $p = .71$, as well as the two-way interaction between reader type and response type, $F(2, 153) = 0.22$, $p = .80$, was nonsignificant. These latter findings indicate that level of reading ability did not influence the number of answers changed by any given reader. That is, both good and poor readers were inclined to make the same number of changes to initial answers. In addition, the nonsignificant two-way interaction indicates that the type of revision

TABLE 1
Analysis of Variance for Reader Type and Response Type

Source	*df*	*SS*	*MS*	*F*	*p*
Reader type	1	2.02	2.0	0.14	.71
Response type	2	832.96	416.4	28.43	.05
Interaction (Reader Type × Response Type)	2	6.45	3.2	0.22	.80
Within-subject (error)	153	2241.0	14.6		
Total score	158	3082.5			

made was not affected by reading ability. For example, poor readers were no more inclined to make right to wrong or wrong to wrong responses than were good readers. Conversely, this latter nonsignificant interaction indicates that good readers did not make more wrong to right responses than did poor readers. The most revealing data of the ANOVA was the significant main effect for response type, $F(2, 153) = 28.43$, $p < .05$. This latter finding indicates that readers made some kind of revisions significantly more than others.

As can be seen from the data in Table 2, post hoc comparisons revealed that readers significantly made more wrong to right ($M = 7.8$) revisions. In both groups, making wrong to wrong ($M = 1.8$) revisions was very minimal. Such findings indicate that both good and poor readers tended to change incorrect answers to correct responses when told that answer changing was acceptable.

Relative to these latter findings, further analysis revealed that poor readers improved their scores more ($M = 13.3$) than did good readers ($M = 11.3$), $t(51) = -2.2$, $p < .05$, as a result of revising their first impressions. An independent t test was performed on the mean difference between the total number of revisions for good and poor readers, respectively. The standard error of the differences between means was obtained, and the significance of the difference is shown in Table 3. As can be seen, the mean dif-

TABLE 2
Tukey (HSD) Pair-Wise Comparisons of Mean Number of Revisions by Response Type of Good and Poor Readers

Response type	*M*	Good	Poor
Wrong to right	7.8[a]	7.4	8.2
Right to wrong	2.6[b]	2.2	3.0
Wrong to wrong	1.8[b]	1.7	2.0

Notes: Group means sharing common notation within are not significantly different from one another. Critical Q value = 3.314, rejection level = 0.05; critical value for comparison = 1.7261; standard error for comparison = 0.736; error term used: Reader × Response × Subject, 153 *df*.

TABLE 3
Two Sample *t* Tests for Mean Number of Revisions for Good and Poor Readers

Variable	*M*	*n*	*SD*	*SE*
Revisions made by good readers	11.3	26	3.498	0.415
				$t(51) = -2.2, p = .05$
Revisions made by poor readers	13.3	27	3.012	0.358

ference between good and poor readers for the total number of revisions was significant, $t(51) = 2.20, p < .05$. These data were obtained by finding for each individual the differences between wrong to right changes and right to wrong changes and finding the means and standard deviations of the resulting data.

The number, percentage, and type of revised answers made by the two levels of readers on the multiple-choice test are presented in Table 4. Among the 53 students in the investigation, 652 revisions were made in answers on the 76-item test, of which 415 or 64% represented changes from wrong to right answers, resulting in a net gain in the scores. On the other hand, 139 or 19% of those revisions made were from right to wrong answers, which lowered scores accordingly. There were 4,028 total responses. Group differences with respect to the three types of response changes were small. Poor readers were more likely than good readers to make a revision of wrong to right.

Ninety-eight percent of all subjects (96% good and 100% poor readers) changed at least 1 answer; almost two-thirds of the subjects changed responses to at least 11% of the 76 items. The ratio of subjects gaining to subjects losing points was 10:3 for poor readers and 5:1 for good readers. Moreover, the ratio of changes for gains to changes for losses was about 2:1 for both good and poor readers. Simply stated, when subjects of both groups made revisions, their changes resulted in a net gain in points twice that of points lost through revision.

Discussion

The major findings of this investigation are consistent with results from previous studies discussed in the literature, where it was found that answer changing significantly increased final scores on test items of examinees.

TABLE 4
Number, Percentage, and Type of Revised Answers of Good and Poor Readers on the Cornell Critical Thinking Test

Ability level	Revisions								Average number per student
	Wrong to right		Right to wrong		Wrong to wrong		Total		
	n	%	*n*	%	*n*	%	*n*	%	
Good ($n = 26$)	192	65	58	18	44	15	294	15	11
Poor ($n = 27$)	223	62	81	21	54	14	358	18	13
Total ($N = 53$)	415	64	139	19	98	14	652	16	12

Notes: Total number of responses = 4,028. Total number of changed responses = 656. Four changed responses not included in the three categories above were right to right.

On the other hand, the results here seem to be in conflict with several previous studies that reported that gains made by good readers as a result of revisions made on test items were significantly higher than those of poor readers. However, for the most part, these investigators' studies involved college subjects whose attitudes and confidence could have had a strong effect upon the results. Also of interest is the fact that most instruments used in previous studies were teacher-made tests that could have been less rigorous than the standardized tests used in this investigation. For example, Berrien (1939) and Mueller and Wasser (1977) speculated that, on easy tests, the better students may be more confident in their first answer and reap little benefit from changing answers. Jarrett (1948) found the following:

> Teacher-made tests could contain "catch questions" or ambiguous questions which might increase the total number of changes made especially from wrong to right changes. Furthermore, for the better reader, the chance of increased wrong to right response factor was even greater than for the less able reader. Also, this factor might operate as a chance-factor. (p. 248)

Also of importance are the questionable procedures and conditions that characterized previous investigations. For example, contrary to the procedure used in this study, most investigators informed their subjects before the test that they would be given the opportunity to record a second response if they felt uncertain about their first impressions.

As for the amount of revisions and gains made by poor readers in this study, these findings could be somewhat attributed to subjects being able to reread the test item with greater comprehension. McMorris and Weideman (1986) found that rereading the item and understanding the question better were popular reasons for revisions among all subjects.

Conclusions

After carefully analyzing the results of this investigation, it seems safe to suggest that young examinees' test scores will significantly improve if revisions are made on multiple-choice items. Therefore, students should be encouraged by educators to carefully deliberate on items when they are not sure of the correct answer and to change that first impression response if they feel it is incorrect.

The results should not be overgeneralized, however. For example, no attempt was made in this study to examine systematically the relationships of test-specific variables to answer-changing behavior, and the prevalence of cheating on a test of this magnitude was not addressed. Jarrett (1948) believed that the possibility of cheating should operate to increase the number of wrong to right responses because the student is not likely to copy from one who is known to be less informed than him/herself. Cheating on

multiple-choice examinations is a serious problem among young test takers, but Belleza and Belleza (1989) believe that such an obstacle can be overcome if proper precautions are used (such as those used in this study). Also, there is the possibility that the attitudes of young examinees toward testing could have significantly affected the outcome of this type of study. Nevertheless, the present investigator concluded that these influences are relatively small and should not outweigh the positive factors associated with answer changing. Because relatively little is known about the decision-making procedure used by young examinees involving standardized test items, further investigation may lead to a better understanding of the validity of answer-changing procedures.

ACKNOWLEDGMENTS

The author extends special thanks to Gwendolyn Maxmillion of Fireman's Fund Insurance Company of Metairie, Louisiana, and to Tim Vance, Psychology Department, Xavier University of Louisiana.

REFERENCES

Archer, S., & Pippert, R. (1962). Don't change the answer. *The Clearing House, 37,* 39–41.

Bath, J. A. (1967). Answer-changing behavior on objective examinations. *Journal of Educational Research, 61,* 105–107.

Belleza, F., & Belleza, S. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16,* 151–155.

Berrien, F. K. (1939). Are first impressions best on objective tests? *School and Society, 50,* 319–320.

Casteel, C. A. (1987). *Answer changing on objective tests: Poor readers can profit.* Unpublished manuscript.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement, 32,* 1035–1038.

Crocker, L., & Benson, J. (1980). Does answer changing affect test quality? *Measurement and Evaluation in Guidance, 12,* 233–239.

Huff, D. (1961). *Score: The strategy of taking tests.* New York: Ballantine.

Jacobs, S. S. (1972). Answer changing on objective tests: Some implications for test validity. *Educational and Psychological Measurement, 32,* 1039–1044.

Jarrett, R. (1948). The extra-chance nature of changes in subjects' responses to objective test items. *Journal of General Psychology, 38,* 243–250.

Lowe, M. L., & Crawford, C. C. (1929). First impression vs. second thought in true-false tests. *Journal of Educational Psychology, 20,* 192–195.

Lynch, D., & Smith, B. (1975). Item response changes: Effects on test scores. *Measurement and Evaluation in Guidance, 7,* 220–225.

McMorris, R., DeMers, L., & Schwarz, S. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement, 24,* 131–143.

McMorris, R., & Weideman, A. (1986). Answer changing after instruction on answer changing. *Measurement and Evaluation in Guidance, 18,* 93–101.

Mueller, D. J., & Shwedel, A. (1975). Some correlates of net gain resultant from answer changing on objective achievement test items. *Journal of Educational Measurement, 12,* 251–254.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14,* 9–13.

Reile, P. J., & Briggs, L. J. (1952). Should subjects change their initial answers on objective-type tests? More evidence regarding an old problem. *Journal of Educational Psychology, 43,* 110–115.

Reiling, E., & Taylor, R. (1972). A new approach to the problem of changing initial responses to multiple-choice questions. *Journal of Educational Measurement, 9,* 67–70.

Vidler, D., & Hansen, R. (1980). Answer changing on multiple-choice tests. *Journal of Experimental Education, 49,* 18–20.