DOCUMENT RESUME

ED 433 366 TM 030 055

AUTHOR Bishop, N. Scott; Frisbie, David A.
TITLE The Effects of Different Test-Taking Conditions on Reading Comprehension Test Performance.
PUB DATE 1999-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 18-22, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Processes; Elementary Education; *Elementary School Students; *Performance Factors; *Reading Comprehension; Reading Tests; Validity
IDENTIFIERS Iowa Tests of Basic Skills; *Testing Conditions

ABSTRACT

Prior research has shown that test takers use a variety of strategies when taking passage-based reading comprehension tests. The specific effects that these alternative strategies have on actual examinee test performance are largely unknown. Evidence suggesting that performance differences exist across testing conditions would imply that the meanings and interpretations that are associated with the corresponding test scores have limited generalizability. In other words, this is an issue related to the validity of the scores from reading comprehension tests. This study addressed the question of whether different test-taking conditions affect reading comprehension test performance (as indicated by total test scores and work rates) and whether the grade level or item cognitive/process classification interacts with the test-taking approach. More than 300 students in grades 3, 5, and 7 took the Iowa Tests of Basic Skills reading comprehension test in conditions that asked them to read the questions before reading the passage or to read the passage before reading the questions. Significant differences were observed in both test scores and work rates. Students in the passages-first condition (standard test directions) had higher test scores at each grade level. Implications for testing practice are discussed. (Author/SLD)

The Effects of Different Test-Taking Conditions on Reading Comprehension Test Performance

N. Scott Bishop

David A. Frisbie

The University of Iowa

Paper presented at the Annual Meeting of the

National Council on Measurement in Education

Montreal

April, 1999

Abstract

Prior research has shown that test-takers use a variety of strategies when taking passage-based reading comprehension tests. The specific effects that these alternative strategies have on actual examinee test performance are largely unknown. Evidence suggesting that performance differences exist across testing conditions would imply that the meanings and interpretations that are associated with the corresponding test scores have limited generalizability. In other words, this is an issue regarding the validity of the scores from reading comprehension tests. This study addressed the question of whether different test-taking conditions affect reading comprehension test performance (as indicated by total test scores and work rates) and if grade level or item cognitive/process classification interacts with the test-taking approach. Significant differences were observed in both test scores and work rates.

The Effects of

Different Test-Taking Conditions

on Reading Comprehension Test Performance

Despite long-standing interest in and research about the assessment of reading comprehension, there remain several issues that warrant additional investigation. One set of questions relates to how examinees take these particular kinds of tests, which specific skills are (and are not) being measured by them, and how these two factors might interact. For example, if it could be demonstrated that observable performance differences occur when students take reading comprehension tests under different administrative conditions, then there would be clear implications regarding the interpretations that can be associated with the test scores obtained under these different conditions. And because reading, like writing, occurs in specific contexts and with unique purposes, determining the generalizability of the scores from across these different situations represents an important form of validity investigation.

From the onset it is important to highlight the fact that directions for passage-based reading comprehension tests—like those used in many standardized achievement batteries—typically instruct students to read each passage before answering the associated questions. Despite these explicit instructions, previous research has shown that some examinees use alternative strategies when taking such tests. Perhaps the most common of these involves reading the questions before the passages (hereafter referred to as "questions-first"). In an analysis of the test-taking behaviors of 26 college students, Farr, Prichard, & Smitten (1990) found that this was the initial strategy used by 27% of the subjects when taking a passage-based reading comprehension test (*Iowa Silent Reading Test*). And while the majority of subjects (62%) started the test by reading passages first, many abandoned this strategy. In the end, less

than half of the subjects (42%) used the "passages-first" strategy throughout the entire test. The remainder used a questions-first approach or some variation of it (e.g., reading questions then looking at the passage for the answers, one question at a time).

Although the Farr et al. study only involved college students, this issue is very relevant for elementary-school students as well. This is because some teachers encourage their students to use the questions-first strategy, especially when they are participating in standardized achievement testing (Anthony, Johnson, Mickelson, & Preece, 1991; Perlman, Borger, Gonzalez, & Junker, 1988). Given these facts, it is logical to ask why this strategy is being used by so many test-takers and why teachers recommend it to their students. The most obvious answer seems to be that they believe it optimizes reading comprehension test performance. One can certainly argue for the potential advantages of this strategy based on related empirical findings. For example, it has been shown that simply including passage headings or titles can have a facilitative effect on comprehension. Here, Dooling & Lachman (1971) demonstrated that providing a title prior to an ambiguous passage significantly improved passage recall. Similarly, Wilhite (1988) showed that passage headings facilitated performance on a multiple-choice retention test for subjects with high preexisting knowledge on the passage topic. Given that prior exposure to passage headings can improve performance (e.g., by providing contexts or evoking schemas that relate text information to the reader's prior experiences), it seems reasonable to expect that the reading of test questions prior to passages could have a similar effect.

One can argue for benefits based on theoretical grounds as well. For example, a questions-first strategy might have the following three benefits: 1) it could reduce working memory load by limiting what the reader must attend to, 2) it could aid examinees in their construction and integration of propositions by alerting them, a priori, to what information is

important; and 3), it could activate relevant schemas and scripts in the test taker's long term memory. As noted above, this approach can provide an important link between the new material and the reader's prior knowledge. (Similar arguments have been associated with "advance organizers.")

Only two prior studies have directly investigated the effectiveness of the questions-first strategy in the context of a standardized achievement test administration (Perlman et al., 1988; Perlman, Borger, Gonzalez-Latin, Hiestand, Junker, & Rosa, 1989). Although there were some circumstances in which this approach was associated with improved performance (e.g., over factual items and for lower achieving students), there were no significant treatment effects overall.

A significant limitation of the Farr et al. and Perlman et al. studies is that the sampled subjects have been restricted in age. Specifically, the Farr et al. study only sampled college students while the Perlman et al. studies only sampled fourth-grade students. As a result, little is known about what test-taking strategies are actually being used by elementary-school students or what affects these alternative strategies have on their test performance. It seems very possible that students at different grade levels might not benefit equally from using the questions-first strategy (e.g., because older students would have more test-taking experience, better short-term memory, etc.).

There are two final issues that need addressing. First, it may be that some strategies are more efficient than others are with respect to the number of items that can be attempted during the allotted testing time. This "work-rate" issue is especially important for the passage-based reading tests under consideration here because there are multiple items that accompany each passage. Thus, if an entire passage were not reached, there would be many items associated with

that passage that could not be attempted. Consequently, no credit would be received for these non-reached items. Second, and more importantly, differential performance across testing conditions would have clear implications for the generalizability of the test scores. Consider the following questions: Which set of test-taking conditions best reflects typical reading situations? What type of reading situation does each condition best reflect? Which conditions reflect reading situations that occur most often in educational settings? What kinds of inferences are reasonable about scores obtained from testing conditions that do not model all kinds of reading that are of interest? Clearly, these are all questions about test score validity.

This study addressed the following questions: 1) Do different test-taking strategies (i.e., questions-first and passages-first) lead to performance differences on reading comprehension tests? 2) Do different test-taking strategies lead to differences in working and completion rates? 3) Is the effect of test-taking strategy mediated by the grade level of students or by items that require different cognitive processing skills (i.e., factual, inferential, and generalization)?

Methods

Subjects

One school system was selected from the state of Iowa on the basis of its size, the grades in which testing was conducted, the school averages from the previous year's *ITBS* reading comprehension results, and the time of year that achievement testing was conducted. (Because the special data collection for this study occurred in the fall, there was no interference with the system's annual achievement testing dates in the spring of the year.) Classrooms from the participating system were selected and assigned to treatment groups so that the expected sample sizes in the two experimental conditions (at grades 3, 5, and 7) would be as similar as possible. (Although there was no data available prior to testing to ensure that the classrooms in each

condition were reasonably similar in their reading achievement, the results from the administration of the *ITBS* conducted later that year suggested that this indeed was the case at grades 3 and 5. Because testing was conducted during the seventh-graders' science classes, it was believed that marked differences in reading achievement between the classroom in each condition at this grade would be unlikely.) Well over 300 students took part in the study, with over 100 students per grade and over 50 students within each condition at each grade. Table 1 shows the exact sample sizes.

Instruments

The *ITBS* Reading Comprehension test (Form H) assesses three levels of understanding (factual, inferential, and evaluative/generalization) over a variety of text types (fables, tales, interviews, fiction, nonfiction, etc.). Reliability information and content/process specifications for this test are documented in the *Manual for School Administrators—Forms G/H* (Hieronymus & Hoover, 1986).

Because it was vital that students actually took the test according to the instructions for the respective conditions, two significant modifications were made to the standard administration procedures for this test. First, the written and oral directions were adapted to be congruent with the protocols for each condition. To further ensure that students understood the testing procedures, the modified directions included an illustrative example (also consistent with each testing condition) that was demonstrated to the students prior to testing. Second, the question/passage sets were reformatted, as described further below, to be compatible with the respective protocols for each condition. This latter change also helped teachers monitor whether students followed the directions for each condition.

Procedures

Students took an on-level version of the *ITBS* Reading Comprehension test (Form H) that was modified to be consistent with the requirements for each experimental condition. Students in the questions-first condition were specifically instructed to read the questions for each passage before reading the passage itself. In order to ensure that these instruction were followed, the test booklets for this condition had the questions for each passage printed on a separate page that preceded the associated passages. Thus, the questions appeared alone, without their respective passages, on the page prior to each passage. (Each question was printed in its entirety, with the stem and all alternatives.) After the questions for a particular passage had been read, the students had to turn to the next page in the test booklet in order to read the associated passage. The questions were reprinted for the students on the page facing the passage; thus, the students had passages available to them when they responded to the test items.

In the passages-first condition, students were instructed to read each passage before reading and answering the associated questions. In order to insure that these instructions were followed, the test booklets for this condition had each passage printed alone, without the questions, on a separate page that preceded the associated questions. After each passage was read, the students had to turn to the next page in the text booklet in order to see the questions for that passage. As before, the passage was reprinted for the students on the page facing the questions. Consequently, these students also had passages available to them when they answered the questions.

Three points about the study's protocols should be highlighted. First, students in both conditions had passages available to them when they answered the test questions, as is the case in the standard test administration. Second, the passages-first condition essentially models what the

standard test instructions intend for students to do. That is, students were specifically asked to read the passages before looking at the corresponding questions. On the other hand, the questions-first condition is unique from many actual reading situations in that advanced knowledge of the specific questions to be asked later is often not available.

Work-rate data in both conditions were collected in the following manner: At both 10 and 20 minutes into testing, students were asked to stop working. They were then told to circle, in their test booklets, the number of the last item they had answered. Students were told to resume working after this was done. This procedure was not a significant intrusion in the testing process because it was done so quickly. Additionally, it is not unusual for teachers to allow their students a short rest period during this test, as test directions suggest they might.

Data Analysis

Performances on the *ITBS* Reading Comprehension test by students (grades 3, 5, and 7) in the two experimental conditions were compared. The outcome measures of interest included: 1) Overall test performance, within each grade, expressed on the grade equivalent scale; 2) Performance on items targeting three different content/process skills (i.e., factual, inferential, and evaluative/generalization), within each grade, expressed as the number of items correct; and 3) Working and completion rates, within each grade, expressed as the number of items attempted by the 10-minute and 20-minute criteria, and the end of testing, respectively.

Differences in overall test performance—expressed on the grade equivalent scale—were analyzed using a two-way ANOVA (Grade x Test-Taking Condition). The interest here was only in the interaction effect and the main effect for test-taking condition. The main effect for grade was not of experimental interest because it would be expected to be significant. Independent-samples t-tests, between the two test-taking conditions at each grade, were used to analyze

performance on items targeting different content/process skills as well as the working and completion rates. Note that it would be inappropriate to analyze the latter measures with a two-way ANOVA because, unlike total test performance, they cannot be placed on a common developmental scale regardless of grade. (Strictly speaking, the tests used at each grade are not equivalent across the different test levels. One might argue, therefore, that this is a confounding factor in this study. However, there is ample evidence that suggests that each test level provides a measure of the same construct. And consequently, there is good reason to believe that the grade equivalents that are associated with total test scores at each level are reasonable estimates of growth along this developmental continuum.)

Results

Descriptive information regarding the distributional characteristics of each outcome measure, as well as the results from the other statistical analyses, are presented in Tables 1 through 4. At each grade, clear differences were observed between the two test-taking conditions on nearly every outcome of interest—total test scores; factual, inferential, and generalization scores; and working and completion rates. As indicated in Table 1, the students in the passages-first (or standard test directions condition) had higher total scores than the students in the questions-first condition. The differences were substantial and reflected not only by the mean differences, but by differences across the entire distribution of test scores (i.e., the 10th, 25th, 50th, 75th, and 90th percentiles). Not surprisingly, the ANOVA results (Table 2) showed a statistically significant main effect for test-taking condition for the total test performance expressed on the GE scale. The interaction effect was not significant at the .05 level. Figure 1 shows the conditional group means associated with this procedure. While these lines are generally parallel, the gap between the means for the two test-taking conditions is slightly greater

at grade 7 than at the other two grades.

In every grade the students in the passages-first group had higher scores over the factual, inferential, and generalization items (see Table 3). As in the total test performance, the differences favoring the passages-first group were generally reflected across the entire distribution of item subset scores. The only case where the performance for questions-first group exceeded that of the passages-first group was for the 90th percentiles of the grade 3 generalization items. (The total number of items in each subset are also given in Table 3 and may be used to convert the item raw scores to proportion correct scores.)

Nearly without exception, the students in the passages-first condition had significantly higher working and completion rates (see Table 4). Differences favoring the passages-first group in Grades 5 and 7 were large, while the differences at grade 3 were much less extreme. In fact, at the 10-minute interval, the working rates for the two groups were very similar at this grade. But by the end of testing, there was a two-item difference between the median completion rates favoring the passages-first group.

Discussion

By any reasonable standard, the results of this study would probably be considered significant. At every grade, the students in the passages-first group obtained higher test scores than their counterparts. This was true for both the total test performance and for performance over the factual, inferential, and generalization items. And although some of the differences would not be considered statistically significant (especially if a Bonferroni-type correction were used to control for the overall Type I error rate), the fact that students in passages-first condition received higher scores in every possible comparison is certainly noteworthy. There was no convincing evidence suggesting that item cognitive/process classification or grade level interacts

with test-taking condition (although the grade × test-taking condition interaction did approach significance).

To help place the observed differences in total test scores in a more practical context, Table 1 also reports the raw scores and percentile ranks (PRs) that correspond to the grade equivalents for each group. For example, the differences in the median total raw scores at grades 3, 5, and 7 were 6.5, 6, and 8 respectively. The corresponding differences in PR units were approximately 20, 16, and 23. Similar comparisons can be made at other percentile points; however, it is clear that indices of both status (PRs) and growth (GEs) are markedly influenced by the between group differences.

Clearly, differences of such magnitude could have consequences in practice. Consider a teacher who forms reading groups based on her students' reading comprehension PRs from the previous year's test results. A student's group placement could be influenced by the conditions under which the student took the test during the previous year. Similarly, a teacher's impression of a student's growth could be substantially different depending on the conditions under which the test was taken from year to year. Based on these results (and all other things being equal) GEs from the questions-first then the passage-first administrations would show relatively more growth than GEs from the passage-first then the questions-first administrations. As discussed further below, these simple examples illustrate that it may not be appropriate to give the same interpretations to scores obtained under different test-taking conditions.

Just as important as the differences in observed test scores is the fact that the differences in working and completion rates also favored the passage-first group. Given these two facts, a very logical inference would be that students in the questions-first condition received lower scores, at least in part, because they were not able to attempt as many items. Perhaps this was

because this condition required a greater amount of reading since the test items had to be read twice. Or, it may be that it simply takes more time to read passages when one is looking for specific information. Whatever the reasons for the difference, it is natural to wonder what the performances of the two groups would have been like independent of the completion-rate differences. Two possible methods of investigating this issue are outlined below. However, it is not possible to answer this question in a completely satisfactory manner with only the data available from this study.

One approach to this problem would be to analyze, at each grade, the total test performance for the two groups using the number of items completed as a covariate (i.e., to statistically control for the differences in completion rates). The results of this analysis of covariance for this study showed that completion rate accounted for a significant proportion of the variability in total-test scores at all three grade levels. Moreover, significant differences remained between the two testing conditions at grade 3 ($p = .009$) and grade 7 ($p < .001$), with the predicted group means at both grades again favoring the passages-first condition. At grade 5 the predicted mean was greater for the questions-first group; however, the difference was not statistically significant at this grade ($p = .696$).

Another approach to determining whether performance differences would have manifested themselves independently of the differential completion rates would involve comparing, at each grade, the performance of the two groups over a limited range of test items (i.e., only over the items that the majority of test-takers in both conditions attempted). In this study, the 10th percentiles of the completion-rate distributions for the questions-first groups at each grade were designated as the stopping points for this analysis. The results were identical to the covariance analyses reviewed above. Specifically, significant differences were observed at

grade 3

(p = .0083) and grade 7 (p < .0001) but not at grade 5 (p = .5372), with all conditional partial test-score means favoring the passages-first group. In nearly every case, the conditional partial test-score means over the corresponding subsets of factual, inferential, and generalization items favored the passages-first group (except for facts at grade 5). Differences at grade 7 were significant for each of these item sets (p < .0002 in every case); differences at grade 3 were generally marginal (p < .10 in every case); and no differences at grade 5 were statistically significant.

As noted earlier, both approaches are of limited usefulness in evaluating performance differences independently of the differential completion rates. For example, partial-test scores calculated over a restricted range of items would never be of interest in practice because of the truncated coverage of the content domain they would represent. Similarly, the statistical adjustments made to the total test scores from the analysis of covariance provides a far-from-perfect indication of how subjects with lower completion rates would have performed if they had answered all the remaining questions. In fact, neither approach provides a satisfactory indication of how subjects would have performed on the non-reached items. This is because neither procedure can account for the fact that individual students can vary substantially in their performance across specific types of passages.

On the other hand, the fact that both procedures led to similar results is hard to ignore. But it is difficult to explain why the questions-first approach, after controlling for completion-rate differences, was associated with poorer performances at grades 3 and 7, but not at grade 5. A more detailed analysis of these results (at the item level) indicated no discernable pattern that favored the questions-first group at this grade (i.e., there was no particular passage, item

classification, or item position that consistently favored the questions-first group at grade 5). It was the case, however, that some students in grade 5 had previous experience with the questions-first strategy. One grade 5 teacher reported, during an informal discussion about the study, that her class had used this strategy during classroom reading activities. (This teacher's class had been assigned to the passage-first condition, and she voiced concern about her students having to take the test without being able to use the questions-first strategy.)

Another possibility is that the results from the other grades might be unique in some way. In particular, the grade 3 results should probably be interpreted cautiously because of the limited test-taking experience of these subjects. It was originally hoped that the grade 3 results might serve as an informal baseline to which the other results might be compared. This was because it seemed unlikely that these students would have received any prior instruction about test-taking strategies, given their grade, and the fact the school system's annual testing occurred in the spring. However, several third-grade teachers in both conditions reported that their students had difficulty with the "standardized" testing procedures. And observations of the students in the questions-first condition indicated that this group might have experienced more difficulty taking the test than their counterparts. Thus, a potential confounding variable in interpreting the results at this grade would be the extent to which the questions-first condition placed differentially more test-taking demands on the students.

Because this study's findings were so different from those of the Perlman et al. studies—where no significant differences were observed between the test-taking conditions—it is important to consider the potential reasons why such a discrepancy might have occurred. First, the Perlman studies made no mention of reformatting the testing materials so that they were consistent with the protocols for each condition (as was done in the present study). This is

important for monitoring whether students actually followed the test directions because the passages and the questions appear together, on the same page or facing pages, in the unmodified test booklets. If a substantial proportion of subjects in either condition did not take the test as instructed, then the data from the Perlman studies would have been tainted.

Second, the Perlman studies provided subjects with more extensive training and practice with the questions-first strategy than the present study (which only used one extended example prior to testing). It may be that extensive training and practice with this strategy can negate the completion-rate differences that were observed in this study. Unfortunately, the Perlman studies did not investigate completion and/or working rates; therefore, it is impossible to make an informed comparison of the studies based on this dimension. However, two possible scenarios seem likely.

First, unlike the current study, the students in the Perlman studies might have had very similar completion-rates between the testing conditions. If this indeed were the case it would have suggested that the performance differences in the current study were solely attributable to the completion-rate differences. However, this argument is weakened by the fact that significant differences were observed in the present study (at grades 3 and 7) even after controlling for completion rates. Second, the questions-first students in the Perlman study could have had lower completion rates, as in the current study, but performed so well on the questions they did answer that this compensated for the difference in the number of items attempted. In short, although higher completion rates tend be associated with higher scores, the relationship is not perfect. If one attempted fewer items but answered the majority of them correctly, it would be possible to obtain a higher total score than someone who attempted more items. (This can occur, especially when the items are of at least moderate difficulty.)

The results from the current study, as well as those of the Perlman studies, call into question the beliefs that some hold regarding the superiority of the questions-first strategy. Ironically, some teachers likely recommend this strategy to their students because they believe it will improve their students' test scores. (Whether or not this is a desirable practice involves the larger issue of how achievement test scores should be used.) However, it appears that this strategy can actually hinder student performance, especially if students do not receive extensive training and practice using the strategy. And as discussed further below, such teachers may not be getting the kind of measure of their students' reading comprehension that is of the most interest to them.

There are many important issues relating to the procedures used by standardized achievement tests to measure reading comprehension that deserve further deliberation. Consider the following questions that relate to the typical formatting and administration of these tests: Should test directions explicitly tell students to read the passage first? Should testing materials be reformatted in order to improve the monitoring of examinees? What, if any, time limits should be used on these tests? Should separate test norms be developed for students who take such tests under different conditions? What should test manuals tell teachers about giving test-taking advice to their students?

Somewhat enmeshed with these issues is the fact that there are overarching validity concerns. This is because the differential effects observed in this study suggest that reading comprehension tests, under standard test-taking conditions, might not measure the broader construct of reading comprehension so purely for every examinee. More specifically, these results call into question the extent to which the scores obtained under standard testing procedures generalize to reading situations other than the ones specifically like those modeled by

the standard conditions. For example, if an educator is most interested in how students comprehend when given a prior set of questions, perhaps the test conditions should model this situation. Such an adaptation would reflect the belief that the processes involved in reading are just as important as the products (i.e., the test scores themselves).

It is clear from the shear breadth of the issues noted above that there is much work left to be done in this area. Indeed, any of the issues raised above represents a legitimate area for future research (e.g., disentangling completion rate from performance differences between test-taking conditions). Additional research suggestions include identifying the exact strategies that elementary-school students use when taking these tests. Here, protocol analyses, videotaping, post-test interviews, or computer simulations might help capture the unique aspects of what these students actually do when they are taking these tests. Related to this issue is the need to learn what specific advice teachers are giving their students regarding how they should take these tests and how pervasive each type of advice is (e.g., what percentage of teachers tell their students to use the questions-first strategy?). Another issue relates to which teacher-suggested strategies are unique to testing and which are promoted more as general reading strategies? And as alluded to above, it is important to determine whether teachers provide their students with practice on the testing strategies they recommend. Finally, what conditions characterize the informal reading techniques with which teachers assess their students' reading comprehension and how are these consistent or inconsistent with the test-taking strategies that they recommend to their students?

But perhaps the most important need is to identify the nature of the reading situations that are considered important in educational settings. For example, educators may or may not want a measure of reading comprehension in which passages are available. The historical practice has been that passages are read first and are accessible to students when they answer questions about

the passage. (The theoretical and/or empirical foundations underlying this practice are unclear because the historical roots of this convention are difficult to trace.) While it may be that such a testing procedure models an important kind of reading of interest to teachers, there may be many others that are of interest as well. Indeed, there are many educational practices where the interest is in how well comprehension occurs without reference to the original reading material (e.g., research using magazines and encyclopedias). There are also some situations where questions are given prior to reading assignments (e.g., many textbooks now give prior questions or advance organizers at the start of each chapter).

It would be a rather simple matter to modify current testing procedures so that they are consistent with any one of these situations. But given the marked performance differences observed in this study, it may be that separate norms are needed for each alternative testing condition (much like the separate norms that are provided for mathematics tests when calculators have been used). Perhaps an even more appropriate alternative would be for future reading comprehension tests to contain a mixture of reading conditions, just as they currently contain a mixture of passage types. Reading comprehension tests include a variety of passage types because test users wish to make generalizations about student performance over a variety of passage types (even though students can vary substantially in their performance over individual passages). The results of this study suggest that there are also performance differences across different reading conditions. Therefore, if test users are similarly interested in how well students comprehend in a variety of different reading conditions, then it would seem reasonable to incorporate these conditions into the tests. Despite the practical limitations associated with varied conditions being used in a single test administration, the gains in validity may just justify such procedures.

References

Anthony, R. J., Johnson, T. D., Mickelson, N.I., & Preece, A. (1991). *Evaluating Literacy: A Perspective for Change*. Portsmouth, NH: Heinemann.

Dooling, D. J. & Lachman, R. (1971). Effects of comprehension on the retention of prose. *Journal of Experimental Psychology, 97*, 404-406.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209-226.

Hieronymus, A. N. & Hoover, H. D. (1986). *Iowa Tests of Basic Skills, Forms G and H, Manual for School Administrators*. Chicago: Riverside Publishing Co.

Perlman, C.L., Borger, J., Gonzalez, C. & Junker, L. (1988). *Should they read the questions-first? A comparison of two test-taking strategies for elementary students*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Perlman, C. L., Borger, J., Gonzalez-Latin, C., Hiestand, N., Junker, L. & Rosa, M. (1989). *How distracting are the distractors? A comparison of three test-taking strategies.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Pressley, M. & Ghatala, E. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23*(4) 455-464.

Wilhite, S. C. (1988). Reading for a multiple-choice test: Headings as schema activators. *Journal of Reading Behavior, 20*(3) 215-228.

Table 1. Distributional Statistics for Reading Comprehension Test Performance (Expressed as Grade Equivalents, Raw Scores, and Percentile Ranks) by Grade and Testing Condition

		Grade Three		Grade Five		Grade Seven	
		Passages	Questions	Passages	Questions	Passages	Questions
Scale	Statistic	First	First	First	First	First	First
Grade	N	63	54	61	59	100	84
Equivalent	M	3.9[a]	3.3[a]	6.1[b]	5.5[b]	7.5[c]	6.3[c]
	SD	1.2	1.2	1.4	1.4	1.5	1.3
	P_{90}	5.5	5.0	8.0	7.1	9.4	7.9
	P_{75}	4.8	3.9	7.0	6.4	8.5	7.3
	P_{50}	3.8	3.2	6.2	5.5	7.6	6.4
	P_{25}	3.3	2.6	5.2	5.0	6.6	5.6
	P_{10}	2.2	1.8	4.4	3.6	5.7	4.6
Raw	k	44	44	54	54	54	54
Score	M	27.6[d]	22.4[d]	33.5[e]	29.3[e]	33.9[e]	25.8[e]
	SD	9.0	9.4	10.3	10.2	9.6	8.4
	P_{90}	37	36	46	42	47	37
	P_{75}	35	30	41	37	41.5	32
	P_{50}	29	22.5	35	29	34	26
	P_{25}	23	15	27	24	27	20.5
	P_{10}	12	10	19	14	21.5	16
Percentile	P_{90}	96.2	92.4	95.8	89.0	88.0	64.1
Rank	P_{75}	90.4	73.2	87.6	77.6	55.2	52.2
	P_{50}	70.4	50.7	73.6	57.4	58.4	35.2
	P_{25}	54.2	29.7	50.2	45.3	38.9	21.0
	P_{10}	18.4	9.4	30.4	14.4	22.6	8.4

Note. [a]$t_{115} = 2.869$, $p = .0049$; [b]$t_{118} = 2.294$, $p = .0236$; [c]$t_{182} = 5.848$, $p < .0001$; [d]$t_{115} = 3.0727$, $p = .0026$; [e]$t_{118} = 2.2580$, $p = .0258$; [f]$t_{182} = 5.9688$, $p < .0001$.

Table 2. Grade by Condition ANOVA Summary Table for Total Test Scores (GEs)

Source	Type III SS[a]	df	MS	F	p
Grade	793.68	2	396.84	222.83	.000
Condition	65.09	1	65.09	36.55	.000
Grade x Condition	10.05	2	5.02	2.82	.061
Error	739.09	415	1.78		

Note. [a]Although there was some disproportionality in the cell sample sizes, it is not believed that this is a significant obstacle in interpreting these results.

Figure 1. Plot of the Conditional Means Corresponding to the Grade by Condition ANOVA

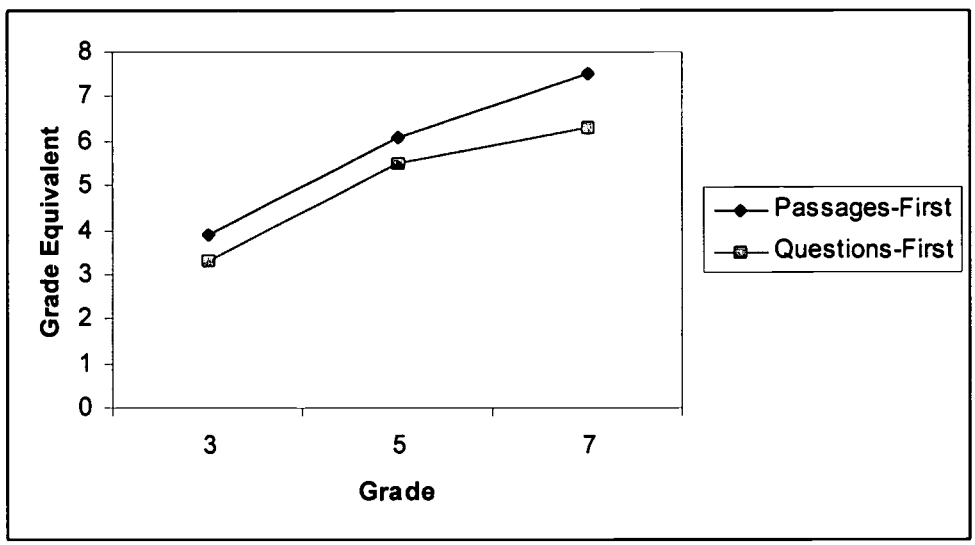

Table 3. Distributional Statistics for the Facts, Inferences, and Generalization Subtest Raw Scores by Grade and Testing Condition

Item Classification	Statistic	Grade Three Passages First	Grade Three Questions First	Grade Five Passages First	Grade Five Questions First	Grade Seven Passages First	Grade Seven Questions First
Facts	k	18	18	17	17	18	18
	N	63	54	61	59	100	84
	M	13.0[a]	10.4[a]	10.1[b]	8.6[b]	12.6[c]	10.1[c]
	SD	4.2	4.7	4.0	3.9	2.8	3.0
	P_{90}	17	16	15	14	16	14
	P_{75}	16	15	13	11	15	12
	P_{50}	14	11	11	9	13	10
	P_{25}	11	6	8	6	10.5	8
	P_{10}	6	4	5	3	9	6
Inferences	k	11	11	16	16	15	15
	N	63	54	61	59	100	84
	M	7.2[d]	5.8[d]	10.0[e]	8.5[e]	8.7[f, h]	6.4[f, h]
	SD	2.7	2.5	3.2	3.2	3.3[g]	2.7[g]
	P_{90}	10	9	14	13	13	10
	P_{75}	9	8	13	11	11	8
	P_{50}	8	6	10	9	8.5	6
	P_{25}	5	4	8	6	6	4.5
	P_{10}	3	3	6	4	4	3
Generalization	k	15	15	21	21	21	21
	N	63	54	61	59	100	84
	M	7.4[i]	6.2[i]	13.3[j]	12.1[j]	12.6[k]	9.4[k]
	SD	3.0	3.2	4.1	3.9	4.5	3.9
	P_{90}	11	12	18	17	19	14
	P_{75}	10	7	16	15	16	12
	P_{50}	8	5	14	13	12	9
	P_{25}	5	4	11	9	9	7
	P_{10}	3	3	7	6	7	5

Note. [a] $t_{115} = 3.2168$, $p = .0017$; [b] $t_{118} = 2.0164$, $p = .0460$; [c] $t_{182} = 5.8985$, $p < .0001$; [d] $t_{115} = 2.8162$, $p = .0057$; [e] $t_{118} = 2.6086$, $p = .0103$; [f] $t_{182} = 5.2395$, $p < .0001$; [g] $F_{99, 83} = 1.49$, $p = .0608$; [h] $t_{181.9} = 5.1592$, $p = .0001$; [i] $t_{115} = 2.0896$, $p = .0389$; [j] $t_{118} = 1.7270$, $p = .0868$; [k] $t_{182} = 5.0968$, $p < .0001$.

Table 4. Distributional Statistics for Number of Items Attempted (at the 10 and 20 Minute Criteria and then End of Testing) by Grade and Testing Condition

		Grade Three		Grade Five		Grade Seven	
		Passages	Questions	Passages	Questions	Passages	Questions
Criteria	Statistic	First	First	First	First	First	First
10- Minute	N	60	46	60	58	97	80
	M	8.8[a]	9.0[a]	12.1[b]	9.0[b]	11.6[c, e]	9.3[c, e]
	SD	3.4	4.0	3.2	2.9	3.9[d]	3.2[d]
	P_{90}	14	14	15	13	16	15
	P_{75}	10.5	10	15	11	16	11
	P_{50}	9	8.5	12.5	8	11	8
	P_{25}	6	6	9.5	7	8	8
	P_{10}	5	6	8	7	8	6
20-Minute	N	59	46	60	58	97	80
	M	20.5[e]	19.2[e]	26.9[f]	20.6[f]	24.1[g,,i]	19.9[g, i]
	SD	6.4	7.8	7.3	6.1	7.3[h]	5.8[h]
	P_{90}	29	28	33.5	30	34	27.5
	P_{75}	24	24	32	24	30	23
	P_{50}	19	18	27	21	23	20
	P_{25}	16	15	23	16	17	16
	P_{10}	14	11	17.5	12	16	12.5
End of Testing	N	63	54	61	59	100	84
	M	40.1[j]	38.8[j]	50.2[k,n]	44.1[k,n]	46.4[l]	42.4[l]
	SD	6.8	6.2	7.6 [m]	11.0 [m]	7.9	8.9
	P_{90}	44	44	54	54	54	54
	P_{75}	44	44	54	54	54	50
	P_{50}	44	42	54	46	49.5	43.5
	P_{25}	36	35	51	38	38	37
	P_{10}	32	29	38	26	33	31

Note. [a] $t_{104} = -0.3256$, $p = .7454$; [b] $t_{116} = 5.5169$, $p < .0001$; [c] $t_{175} = 4.1616$, $p < .0001$; [d] $F_{96,79} = 1.49$, $p = .0678$; [e] $t_{175.0} = 4.2415$, $p = .0001$; [e] $t_{103} = .9755$, $p = .3316$; [f] $t_{116} = 5.1079$, $p < .0001$; [g] $t_{175} = 4.1825$, $p < .0001$; [h] $F_{96,79} = 1.60$, $p = .0308$; [i] $t_{174.7} = 4.2773$, $p = .0001$; [j] $t_{115} = 1.0283$, $p = .3060$; [k] $t_{116} = 3.5555$, $p = .0005$; [l] $t_{182} = 3.1906$, $p = .0017$; [m] $F_{58, 60} = 2.09$, $p = .0051$; [n] $t_{102.8} = 3.5344$, $p = .0006$.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Effects of Different Test-Taking Conditions on Reading Comprehension Test Performance

Author(s): N. Scott Bishop & David A. Frisbie

Corporate Source:

Publication Date: April, 1999

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 the identified document, please CHECK ONE of the following options and sign the release below.

__X__ Check here for Level 1 Release, permitting reproduction and dissemination in microfiche and other ERIC archival media (e.g. electronic)and paper copy.

or

_____ Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

or

_____ Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

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 this document as indicated above. Reproduction from the ERIC microfiche or electronic 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.

Signature: 
Printed Name: N. Scott Bishop
Address: Iowa Testing Programs
346 Linquist Center
University of Iowa
Iowa City, IA 52246

Position: Research Assistant
Organization: University of Iowa
Telephone Number: (319)339-4382
Date: 7/12/99