DOCUMENT RESUME

ED 227 448 CS 007 024

AUTHOR Johnston, Peter; Afflerbach, Peter
TITLE Centrality and Reading Comprehension Test Questions.
PUB DATE Nov 82
NOTE 12p.; Paper presented at the Annual Meeting of the New York State Reading Association (16th, Kiamesha Lake, NY, November 2-5, 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Higher Education; *Item Analysis; *Reading Comprehension; *Reading Research; *Reading Tests; Standardized Tests; *Test Items; Test Reliability; Test Validity
IDENTIFIERS *Discrimination Index

ABSTRACT

A study examined the nature of the questions contained in two major standardized reading comprehension tests in terms of their centrality to the text. It was hypothesized that the use of a discrimination index for item selection would tend to favor relatively trivial questions. Half the reading selections from the Stanford Diagnostic Reading Test (SDRT), which gives more weight to item discrimination indexes, and the Metropolitan Reading Test (MRT), which does not, were randomly selected. The multiple choice questions that followed each reading selection were transformed into statements by adding the correct answer to the question stem. Thirty faculty and graduate students from a school of education read each of the 10 passages and were asked to rate each statement on a four-point scale as to whether it was central or peripheral to comprehension of the passage. Thirty-seven percent of the items from the SDRT were rated peripheral, while only 12% of the items from the MRT were so rated. When the mean for all statement ratings was figured, the statements from the SDRT were found to be significantly more peripheral than those from the MRT. Since the Stanford manual clearly states that discrimination indices played a large part in decisions about which items to include, the finding adds some credence to the suggestion that an emphasis on the discrimination index in test item selection will tend to favor more trivial items. (HTH)

CENTRALITY AND READING COMPREHENSION TEST QUESTIONS

Peter Johnston

Peter Afflerbach

State University of New York at Albany

Reading Department LCB-30

1400 Washington Avenue

Albany, NY 12222

CENTRALITY AND READING COMPREHENSION TEST QUESTIONS

Tests of reading comprehension generally present students with a series of brief passages, each followed by multiple choice questions. These questions perform a crucial function. It is the student's success or failure on them that tells us whether or not the student has comprehended the text. The development of these questions is thus of great importance. The purpose of this paper is to focus attention on certain characteristics of these questions which may limit the validity of standardized tests.

Recent theoretical investigations of reading comprehension (e.g. Omanson, 1979; Schank, 1975; Spiro, 1980) have proposed that the reader constructs a coherent network or causal chain in which the more central elements of the text are stored. Peripheral elements of text are stored in this chain or network only as they are related in some way to the central theme or chain. Thus, readers who have comprehended a text could be expected to recall the central elements or to adequately answer questions which relate to those elements. On the other hand, they would be considered less likely to have stored information which was peripheral to the main course of events or arguments. If one wished to know whether or not a person had comprehended a given text, then it would be more convincing to know that he was able to respond adequately to central questions than to find that he could adequately locate peripheral details. That is, central questions would represent a more valid measure of reading comprehension, as it is now understood through conceptual dependency theory, than would peripheral questions.

Tuinman(1979) has suggested that test questions tend to assess rather trivial information. He contends that since test item writers have to write many questions about a brief text, and often the questions must be of a specified type, trivial details will be emphasized. If Tuinman's claim is true, then our tests of reading comprehension would be of questionable validity. We would be testing information to which a good reader would normally devote little attention during reading.

Recently, Johnston (1981) made a similar claim, though on different grounds. He found that readers performed quite differently under two different testing conditions. When the text was available for the reader to refer back to, as is the case in current standardized reading comprehension tests, the questions which were most readily answered by students were ones which related to relatively trivial, or peripheral, information. Questions which addressed the more central information in the text were very poorly answered under these conditions. Consequently, differences between different students' scores were more strongly dependent upon students' performance on the more trivial questions. Furthermore, the students who had more prior knowledge relevant to the text topic generally scored higher than other students, and did especially well on these peripheral questions. Thus it was hypothesized that the use of a discrimination index for item selection would tend to favor relatively trivial questions.

This hypothesis was not, however, generated using standardized test

materials. Particularly, the texts were longer than those normally found in tests of reading comprehension. The hypothesis seemed readily tested, and rather important in terms of the validity of current assessment practice. The present study examined the nature of the questions contained in two major reading comprehension tests, in terms of their centrality to the text.

METHOD

Materials

The tests selected for study were the Metropolitan Reading Test (Intermediate, Form JS, 1978) and the Stanford Diagnostic Reading Test(Form A, Level III, 1974). The particular advantage in choosing these two tests was that the SDRT Manual specifically states that "... more weight was given to item discrimination indices in SDRT III "(p. 28). The Metropolitan manual, on the other hand, describes no fewer than eight different criteria which were used for item selection, including teacher questionaires and instructional objectives, and it makes no claim to have emphasized the use of discrimination indices.

In order to make a manageable task that could be done within an hour, half of the reading selections in each test were randomly selected, with their accompanying questions. The multiple choice questions which followed each reading selection were transformed into statements by adding the correct answer to the question stem.

For example:

TEST ITEM

1) The "turning point" of the American Revolution was the Battle of -
1 Yorktown 3 Long Island
2 Saratoga 4 Cowpens

was changed to:

The "turning point" of the American Revolution was the Battle of Saratoga.

Next, the researchers created some additional statements based on the same reading selection. These statements were randomly interspersed amongst the statements formed from the multiple choice questions in order to create a broader comparative base for the rating task to follow. All statements were accompanied by a 4-point rating scale which was anchored at either end by the terms "very peripheral"(1) and "very central"(4).

Subjects

Thirty faculty and graduate students in a School of Education participated in the rating task. None was aware of the hypothesis being tested.

Procedure

Subjects read each of the ten test passages. After reading each passage, subjects read and executed the following instructions:

Please rate the following statements in terms of the extent to which an understanding of the statement is central (in some absolute sense) to an understanding of the text.

	very peripheral			very central
The British surrendered at Yorktown.	1	2	3	4

Statement ratings with means at or below 2.00 were considered peripheral, and those with means at or above 3.00 were considered central to an understanding of the text. These criteria generally required at least 70% directional agreement between raters. Subjects were allowed to re-read the text as needed, as is common practice with standardized reading comprehension tests.

RESULTS AND DISCUSSION

Means and standard deviations were calculated for each statement and for each test. The ratings were examined both in terms of absolute numbers of central and peripheral items, and in terms of mean ratings. Of the items rated, eleven (37%) of the items from the Stanford were rated 2.00 or below (peripheral to an understanding of the text), while three (12%) of the items from the Metropolitan received such ratings. (See Figure 1)
Three (10%) of the items from the Stanford were rated as being central to an understanding of the text (above 3.00), while three (12%) of the items from the Metropolitan were so rated.

The mean for all statement ratings for the Stanford was 2.17 while the mean for the Metropolitan was 2.53. These means were compared with each other and with the expected normal mean (2.5) using simple t-- tests. The statements from the Stanford were found to be significantly more peripheral than those from the Metropolitan ($t_{53}=2.83$ $p<.05$), and from the expected mean ($t_{29}=2.82$, $p<.01$). The Metropolitan did not differ significantly from the expected mean ($t_{24}=.397$; $p>.05$). The two tests were clearly quite different in terms of the nature of the questions selected. The Stanford Diagnostic Reading Test

contained items which were generally rated more peripheral than those found in the Metropolitan, or than would be expected on the basis of chance selection of items. Since the Stanford manual clearly states that discrimination indices played a large part in decisions about which items to include, the finding adds some credence to the suggestion that an emphasis on the discrimination index in test item selection will tend to favor more trivial items over more central ones. The criterion makes sense from a statistical point of view, but not in terms of a theoretical model of how readers comprehend. Emphaisizing the use of a discrimination index may increase the reliability of a test. However, at the same time it may decrease the validity of the test by forcing the selection of inappropriate items. Indeed, dependence on such a criterion as the discrimination index will tend to produce tests which test not what is comprehended in normal reading, but what can be comprehended under specific circumstances.

Neither of the tests had more central than peripheral items. The distribution of rating means depict one positively skewed and one normal distribution (Figure 1). What might be considered an optimal distribution? If one believes that to find out whether people comprehended something it is better to ask questions tapping information which is central to the text, then the more desirable distribution of test items would be negatively skewed, with a high mean centrality rating. Such a distribution would suggest a more valid test in terms of construct and face validity, since few would argue that responses to trivial questions represent an adequate indication of comprehension. It would be especially difficult to defend such a measure from

a theoretical standpoint.

The Stanford Users Manual notes that the use of discrimination indices will lead to scores that are "...more reliable for poor readers.." (p.28) While this consideration may provide better reliability, it may be at the cost of considerable loss in validity. We need to decide if we wish to assess whether or not the student comprehended or if we wish to rank order students on some more or less related dimension. Does a test of reading comprehension which consists of items with a peripheral bias give a true indication of comprehension?

One solution to this dilemma may be to consider using reading comprehension tests in which the text is not available to refer back to while answering questions. According to Johnston's (1981) findings, when the text is unavailable for lookbacks, the questions which are better discriminators are those which are central to an understanding of the text. Informal reading inventories are normally of this form, but their item development is relatively crude in comparison with the painstaking efforts of the developers of the Stanford Diagnostic Reading Tests and the Metropolitan. It would not be difficult to develop a group administered test which prevented lookbacks and stressed acquisition of central information. Certainly, the data suggest that such a test would provide a more valid measure of reading comprehension.

SUMMARY

This study suggests several considerations we need to make concerning

standardized reading comprehension tests. Johnston (1983) has noted that "...if we define comprehension as the forming of a coherent model of the text, then we are likely to be most interested in the reader storing central aspects of the text". We believe that the best way to measure whether a student has stored central aspects, or "gotten the gist", is to ask questions which concern those things in particular. Asking peripheral questions, while possibly increasing reliability of certain scores, may also reduce the validity of the test because it does not give appropriate weight to questions that deal with central information. Discrimination indices may help in the construction of tests which provide better differentiation between students. However, we may be unable to appropriately interpret results if perfomance on specifically peripheral questions have a large effect on the aggregate test score.

REFERENCES

Johnston, P.H. Assessment of reading comprehension: A cognitive basis. International Reading Association Research Monograph, Newark, Delaware, 1983.

Johnston, P.H. Prior knowledge and reading comprehension test bias. Unpublished doctoral dissertation. University of Illinois, 1981.

Karlsen, B., Madden, R. & Gardner, E.F. Stanford Diagnostic Reading Test Manual, (Level III). New York:Harcourt, Brace & Jovanovich, 1974.

Omanson, R.C. The narrative analysis. Unpublished doctoral dissertation, University of Minnesota, 1979.

Prescott, G.A., Balow, I.H., Hogan, T.P. & Farr, R.C. Metropolitan Achievement Test Manual (Intermediate, Form JS). New York:The Psychological Corporation, 1978.

Schank, R.C. The structure of episodes in memory. In D.G. Bobrow & A. Collins (Eds.) Representation and understanding:Studies in cognitive science. New York:Academic Press, 1975.

Spiro, R.J. Schema theory and reading comprehension:New directions (Tech. Rep. No. 191). Urbana:University of Illinois, Center for the Study of Reading, December, 1980.

Tuinman, J.J. Reading is recognition--When reading is not reasoning. In J.C. Harste & R.R. Carey (Eds.) New perspectives on comprehension. (Monograph in Language and Reading Studies No. 3). Bloomington:Indiana University, 1979. Pp. 38-48.

Figure 1

CENTRAL/PERIPHERAL RATINGS OF STANDARDIZED TEST QUESTIONS