DOCUMENT RESUME

ED 368 784 TM 021 297

AUTHOR Huntley, Renee M.; Miller, Sherri
TITLE Are Reading Comprehension Tasks Affected by Line References in Test Items?
PUB DATE Apr 94
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Classification; Difficulty Level; High Schools; *High School Students; Racial Differences; *Reading Comprehension; Reading Tests; Sex Differences; Test Construction; Test Format; *Test Items
IDENTIFIERS ACT Assessment; *Line References (Tests)

ABSTRACT

Whether the shaping of test items can itself result in qualitative differences in examinees' comprehension of reading passages was studied using the Pearson-Johnson item classification system. The specific practice studied incorporated, within an item stem line, references that point the examinee to a specific location within a reading passage. Versions of 71 test items with and without line references were prepared and classified by the Pearson-Johnson system as textually explicit, textually implicit, or scriptally implicit. Each experimental unit was administered as part of the ACT Assessment to nearly 425 examinees. The practice of citing specific lines of text generally served to make items easier, although mainly for low-ability examinees, thus accounting for the consistently lower discrimination in the version with line references. The performance of males and females, or Blacks and Whites, however, was not differentially affected. That performance is affected by the shaping of test items has important implications for test construction. It remains to be studied whether adding a line reference gives an advantage to an examinee in a timed test. Four tables present study findings. (Contains 6 references.) (SLD)

Are Reading Comprehension Tasks Affected by Line References in Test Items?

Renee M. Huntley
Sherri Miller

American College Testing

Paper presented to the annual meeting of the
American Educational Research Association
New Orleans, April 1994

Are Reading Comprehension Tasks Affected by Line References in Test Items?

Objectives

The steady proliferation of research into reading processes has not only aided classroom instruction but also alerted the educational community to the validity problems in some reading tests. Validity issues raised include the "meaningfulness of the construct with which we are dealing" and "the degree to which our chosen assessment technique actually reflects ability in the specific skill which we claim it measures" (Johnston, 1983). Validity problems are further confounded by the indisputable fact that "the psychological reality of the systems of classification has not yet been established" (Johnston, 1983). Given this lack, it might be helpful to study whether the shaping of items can itself result in qualitative differences in examinees' comprehension of reading passages.

For the purposes of this study, the Pearson-Johnson item classification system (Pearson & Johnson, 1978) was used for several reasons: it has intuitive and ecological validity; it recognizes the differences in reading behavior effected by prior knowledge; and, especially, it deals with the source of the information not just as between the text and the reader's head but as regards location within the text.

A frequent practice in reading tests is the subject of this study--namely, incorporating within an item stem line references that point the examinee to a specific location within a reading passage. Having text readily available during a testing period, all by itself, affects the nature of what is being assessed (Johnston, 1984). So one could readily anticipate that citing specific lines of text would further affect the nature of examinee performance, especially in a timed testing context. Just exactly how performance is affected, whether item classifications are also affected,

and whether males or females, blacks or whites, and high- or low-ability examinees constitute the groups significantly affected are the issues this study was designed to illuminate.

Method

Two versions were constructed of 71 reading comprehension test items selected from among 13 passage sets (i.e., a reading passage with items generated from that passage to assess an examinee's reading comprehension). Version 1 of the item included a line reference in the item stem that directed the examinee's attention to a specific passage location; the stem of Version 2 omitted the line reference but otherwise was identical to the first version. The following pair of item stems illustrates the two versions:

> Which of the following countries is similar to the United States in maintaining both a trade deficit and also rapid economic growth (lines 17-19)? -- Version 1
>
> Which of the following countries is similar to the United States in maintaining both a trade deficit and also rapid economic growth? -- Version 2

Each item was classified according to the Pearson-Johnson system (1978) as being Textually Explicit (TE), Textually Implicit (TI), or Scriptally Implicit (SI). Both question and answer information are stated in a single sentence in the text in TE items. In TI items, on the other hand, question and answer information are stated in different sentences in the text, and thus the reader is required to combine separate pieces of information. For SI items, the reader is dependent on both information from the text and also on background knowledge.

An experimental unit consisted of one reading selection with its accompanying items, among which were embedded several experimental items, ranging in number from 5 to 10 and distributed equally so that both versions were represented; no one experimental unit was

conspicuously marked by a particular item type.

Each experimental unit was administered as part of the ACT Assessment pretest procedures to approximately 425 examinees in randomly equivalent samples such that only one experimental pretest unit was administered to any one examinee, and no examinee was given both versions of the same experimental item. The administration of the experimental units occurred over two successive pretest administrations.

Differences in performance (p-values and biserials) between the two versions of the items were examined. It should be noted that biserials were derived by using the total 40-item test score on the ACT Assessment Reading Test, which was administered at the same time. Performance on the ACT Assessment Reading Test was also used to define the following examinee ability levels: the upper 27% of examinees constituted the high-ability level; the middle 46%, the middle-ability level; and the lower 27%, the low-ability. One-way analyses of variance were performed on the difficulty and discrimination values of the two item versions for the total group. To compare the discrimination values, the biserials were transformed to a z-value using a transformation developed by R. F. Tate (See Walker and Lev, 1953, p. 269-70), which is similar to Fisher's logarithmic transformation of r to z. Further analyses were conducted to study the performance of the two item versions by ability group and by item location. To compare differences in performance on the two versions and across the three item classifications between males and females, and blacks and whites, analysis of variance procedures were used separately for males/females and for blacks/whites. Item classifications for each version were compared judgmentally to determine whether item classification and, consequently, cognitive demands had changed.

Results

Table 1 below presents the mean difficulty and biserial values for the experimental items with and without line references by total group, gender, and race. The data in Table 1 is based on a total of 142 items (71 items each administered with a line reference and without a line reference). Of the approximately 425 examinees who took each item, about half the students were male and half female. Of these, approximately 40 examinees were African American, and approximately 300 were Caucasian.

Table 1
Means and Standard Deviations for both Difficulty (proportion answering the item correctly) and Discrimination (Biserial r) by Total Group, Gender, and Race

Group	Performance **WITH** line references (N=71 items)		Performance **WITHOUT** line references (N=71 items)	
	Mean Difficulty (SD)	Mean Discrimination (SD)	Mean Difficulty (SD)	Mean Discrimination (SD)
Total Group	.68 (.17)	.38 (.11)	.59 (.15)	.41 (.10)
Males	.68 (.17)	.41 (.12)	.59 (.15)	.43 (.11)
Females	.68 (.17)	.36 (.11)	.58 (.15)	.40 (.10)
African-American	.58 (.16)	.28 (.10)	.46 (.14)	.35 (.12)
Caucasian	.70 (.17)	.36 (.11)	.60 (.15)	.39 (.10)

2 x 2 analyses of variance were performed to examine the effects of line references on item difficulty and discrimination by gender (males versus females) and by race (African-

American versus Caucasian). No interactions were present for difficulty or for discrimination for gender or race variables: line references did not differentially affect males and females or blacks and whites. No significant main effects of line reference were observed for discrimination. However, the main effect of line reference on difficulty was significant in both analyses: items with line references are, on the average, easier than the same items without line references (by line reference and gender: $F_{1,136}=23.53$, $p<.001$; and by line reference and race $F_{1,136}=12.76$, $p<.001$). The mean difference in p-values between items with line references and items without line references across all groups was around .10. Two pairs of items deviated from this expected trend in difficulty: they were identified as being easier without a line reference.

In the case of one of these pairs, the first version referenced several paragraphs among which occurred a sentence whose wording was repeated in the keyed response. The second version cited lines of the text from within those paragraphs, but did not include the sentence echoed in the keyed response; these lines demanded of the examinee an interpretation or translation. Although the classification of this item did not overtly change, the cognitive task had subtly shifted.

In the other pair of items, the classification did change from a TE without a line reference to a TI with a line reference. The line reference pointed to a paragraph late in the text where a generalization was enunciated. The keyed response, however, echoed the wording of the first paragraph of the text. In some regards, the cognitive demands of this pair are similar to the other pair of deviant items but here, at least, the shift in classification does indeed match the shift in task. Both pairs of items, however, were deemed to be flawed not only because the contrast between the two versions was not sharply delineated, but also because the items hinged on mere

recognition of passage phrasing rather than on the construction of meaning by the examinee.

The difficulty of the two item versions by ability group was also examined. Table 2 presents the mean difficulty of the two versions for the high- and the low-ability groups.

Table 2
Means and Standard Deviations for Difficulty for
Line Reference Type by Ability Level

Ability Group	Mean Difficulty (SD) WITH line references (N=71 items)	Mean Difficulty (SD) WITHOUT line references (N=71 items)
High Ability	.84 (.14)	.79 (.14)
Low Ability	.53 (.20)	.42 (.14)

An analysis of variance procedure was performed on the mean difficulty by item version for both the low- and high-ability examinees. No significant difference in difficulty was found for the high-ability examinees between the two versions. However, a significant difference was found for the low-ability examinees ($F_{1,134}$=16.27; p<.0001). For them, items with a line reference were, on the average, .11 points easier than items without a line reference. Thus, the difference in difficulty for items with and without line references is more pronounced in the low-ability group than in the high-ability group.

Also examined was the effect of line references on the difficulty and discrimination of an item according to its location within the passage set (beginning section - items 1 through 14; end section - items 15 through 20). Table 3 presents difficulty and discrimination statistics for items with and without line references by item location.

Table 3
Means and Standard Deviations for Difficulty and Discrimination for Line Reference Type by Item Location

Location	Performance WITH line references		Performance WITHOUT line references	
	Mean Difficulty (SD)	Mean Discrimination (SD)	Mean Difficulty (SD)	Mean Discrimination (SD)
Beginning Section (N=117 items)	.69 (.18)	.38 (.11)	.59 (.16)	.40 (.10)
End Section (N=25 items)	.63 (.14)	.39 (.09)	.56 (.14)	.44 (.08)

Analyses of variance were performed on the mean difficulty and mean discrimination by item version (i.e., with line reference or without line reference) for items located in the beginning section and in the end section. No significant differences were found for discrimination for items in either sections. No significant difference was found in difficulty between item versions for the end-section items. However, a significant difference in difficulty was found between item versions for beginning-section items, items in position 1 through 14 ($F_{1,116}$=9.76; $p<.01$). Items with a line reference in this location were, on the average, .10 points easier than items without a line reference in this location. The difference in difficulty for the two item versions appears to have been influenced by the item's location. This finding suggests that the test was somewhat speeded and that examinees were randomly guessing on items at the end of the passage set.

Regarding the item classifications, two questions were of interest: Did the addition of line references to an item stem result in a change in the classification of the item? Did a group of examinees--blacks or whites, males or females--perform better or less well on a particular item classification (TE, TI, or SI)?

Of the 71 pairs of experimental items, only 6 pairs manifested changes in classification. Of these 6, one has been discussed above in the context of the expected trend in difficulty of the paired items. Of the remaining 5 pairs, one pair had clearly been misclassified from the beginning. Even so, the line-reference version not only was easier but also revealed the same shift in cognitive demand observed in most of the line-reference versions and previously discussed. In the remaining 4 pairs of items, the classification shift reflected a shift in cognitive demand, but these items were deemed flawed because the line references in the stem pointed to passage text that was exactly echoed in the keyed response and thus what was being measured was far more superficial than had been the original intent of these experimental test units.

Whether or not an item contained a line reference in its stem, was any differential behavior on the basis of the item classification observed? Table 4 shows the mean difficulty and discrimination values for item classification by gender, race, and total group.

Table 4
Means and Standard Deviations for Difficulty and Discrimination
for Item Classification by Group

Group/Item Classification	Performance on Item	
	Mean Difficulty (SD)	Mean Discrimination (SD)
Total Group		
TE (119 items)	.66 (.15)	.40 (.11)
TI (17 items)	.50 (.15)	.34 (.14)
SI (6 items)	.55 (.18)	.45 (.19)
Blacks		
TE	.54 (.16)	.32 (.15)
TI	.40 (.10)	.29 (.11)
SI	.38 (.10)	.30 (.12)
Whites		
TE	.67 (.13)	.37 (.10)
TI	.52 (.14)	.33 (.14)
SI	.57 (.12)	.43 (.17)
Females		
TE	.65 (.12)	.38 (.11)
TI	.49 (.11)	.33 (.10)
SI	.51 (.18)	.45 (.14)
Males		
TE	.66 (.14)	.42 (.11)
TI	.51 (.12)	.36 (.11)
SI	.58 (.16)	.44 (.15)

Analysis of variance procedures were performed on difficulty and discrimination by item classification and gender, and by item classification and race. Both analyses showed significant main effects for difficulty by item classification (by gender, $F_{2,126}=14.81$; $p<.0001$ and by race, $F_{2,126}=12.07$; $p<.0001$). Tukey follow-ups revealed that TE items are easier than both TI and SI items for all groups studied, but caution must be exercised here because there were very few TI and SI items included in the study, a lack that could be remedied, perhaps, in a future study. No significant interaction effects were found when comparing blacks and whites, and males and

females, on the difficulty of items within the different item classifications. Finally, no significant differences were found for discrimination values.

Discussion

The result that performance is affected by the shaping of the item stems in this study has important implications for test constructors, whether classroom teachers or standardized testers. In this instance, the practice of citing specific lines of text in the item stems generally tends to make items easier, but mainly for low-ability examinees, thus accounting for the consistently lower discrimination in the version with line references. However, the performance of males or females and blacks or whites is not differentially affected.

The practice also subtly changes the nature of the cognitive demands on the examinee even when the classification of the item ostensibly remains unchanged. Given that systems of classification are tenuous at best, it is not surprising that a TE item with a line reference in its stem merely measures the examinee's ability to follow a direction in locating information in the text rather than measure the examinee's ability to recall information that just happens to be at a certain location in the text, as would be the case in a TE item without a line reference. If the item stem manipulation subtly changes the task, and this change is neither acknowledged nor accounted for, validity issues are raised regarding both the comprehension task (Cunningham & Moore, 1993) and also the reading process as is commonly understood today (Johnson & Afflerbach, 1982). Furthermore, when we locate the information in the text *for* the examinee, we may be depriving ourselves of possibly "useful sources of information" about the reader's strategies (Johnston, 1983). Clearly, editorial practices, like the one described in this study, have an impact on the testing context.

Are there any situations where line references can legitimately be included in an item stem? There is at least one: when an item deals with the meaning of vocabulary in the context of a particular sentence in the passage, a line citation is obligatory since the word of interest could reoccur in several other locations within the passage, where its meaning might vary. However, when the item deals with a larger linguistic unit (e.g., a phrase or clause), line references should be avoided if what is being measured is central to the meaning of the passage and the intent is to measure the examinee's ability to construct meaning and retain it. Finally, deliberately adding line references to item stems in order to make items easier, an editing practice sometimes used by test constructors, should be discouraged because such editing changes the nature of what is being measured.

Regarding the nature of what is being measured, items that lend themselves to line references too often appear to measure picky and basically trivial information in a reading passage. It bears repeating that the recall of trivial information is not typically deemed intrinsic to the reading process and is another practice that should be discouraged.

One aspect of speededness not examined in this study concerns whether adding a line reference gives an advantage to an examinee in a timed test. A follow-up study is not recommended because such test items mainly measure trivial information and mainly affect the performance of only low-ability examinees. For now, at least, we can reasonably conclude that the validity of what reading tests measure will be enhanced if we mainly avoid the practice of adding line references to item stems except where absolutely necessary and only when we are fully informed about the implications.

Bibliography

Cunningham, J. W. & Moore, D. W. (1993). The contribution of understanding academic vocabulary to answering comprehension questions. *Journal of Reading Behavior*, 25(2), 171-180.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 19(2), 219-239.

Johnston, P. H. (1983). *Reading comprehension: a cognitive basis.* Newark, Delaware: International Reading Association.

Johnston, P. & Afflerbach, P. *Centrality and reading comprehension test questions.* Paper presented at the Annual Meeting of the New York State Reading Association (16th, Kiamesha Lake, NY, November 2-5, 1982).

Pearson, P. D. & Johnson, D. (1978). *Teaching reading comprehension.* New York: Holt, Rinehart & Winston.

Walker, H. M. & Lev, J. (1953). *Statistical inference.* New York: Holt, Rinehart & Winston.