DOCUMENT RESUME

ED 385 581 TM 024 019

AUTHOR Freedle, Roy; Kostin, Irene
TITLE The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-13; TOEFL-RR-44
PUB DATE May 93
NOTE 56p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Construct Validity; *Difficulty Level; *Multiple Choice Tests; *Prediction; *Reading Comprehension; Reading Tests; Second Language Learning; Test Format; *Test Items
IDENTIFIERS Discourse; *Test of English as a Foreign Language; Variance (Statistical)

ABSTRACT

Prediction of the difficulty (equated delta) of a large sample (n=213) of reading comprehension items from the Test of English as a Foreign Language (TOEFL) was studied using main idea, inference, and supporting statement items. A related purpose was to examine whether text and text-related variables play a significant role in predicting item difficulty. It was hypothesized that multiple-choice reading comprehension tests are sensitive to many sentential and discourse variables and that many of these variables contribute significant independent variance in predicting item difficulty. The majority of the sentential and discourse variables identified in the literature were found to be significantly related to item difficulty within TOEFL's multiple choice format. A significant relationship was found between item difficulty and text and text-related variables, supporting the claim that multiple-choice items yield construct valid measures of comprehension. For the full sample of 213 items, with the equated delta the dependent variable, 33% of the item difficulty variance could be accounted for by 8 variables. Five tables present analysis details, and an appendix gives correlations and regressions for each of three item types. (Contains 43 references.) (SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 44
MAY 1993

The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items

Roy Freedle
Irene Kostin

ETS
Educational Testing Service

The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items

Roy Freedle and Irene Kostin

Educational Testing Service
Princeton, New Jersey

RR-93-13

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

The purpose of the current study is to predict the difficulty (equated delta) of a large sample (n=213) of TOEFL reading comprehension items. (Only main idea, inference, and supporting statement items were sampled.) A related purpose was to examine whether text and text-related variables play a significant role in predicting item difficulty; we argued that evidence favoring construct validity would require significant contributions from these particular predictor variables. In addition, details of item predictability were explored by evaluating two hypotheses: (1) that multiple-choice reading comprehension tests are sensitive to many sentential and discourse variables found to influence comprehension processes in the experimental literature, and (2) that many of the variables identified in the first hypothesis contribute significant independent variance in predicting item difficulty.

The great majority of sentential and discourse variables identified in our review of the experimental literature were found to be significantly related to item difficulty within TOEFL's multiple-choice format. Furthermore, contrary to predictions which we attributed to critics of multiple-choice tests, the pattern of correlational results showed that there is a significant relationship between item difficulty and the text and text-related variables. We took this as evidence supporting our claim that multiple-choice reading items yield construct valid measures of comprehension. That is, since critics have pointed out that reading items can often be correctly answered without reading of the text passage, this seems to imply that item variables (not text nor text-related variables) should be prominent predictors of reading item difficulty. Since the contrary relationship was found, we concluded that this provides evidence favoring construct validity. We found, further, in several stepwise linear regression analyses, that many of these text and text-related variables provide independent contributions in predicting reading item difficulty. This was interpreted as providing additional support for construct validity.

More specifically, apart from the correlational results, the following stepwise linear regressions results were obtained.

For the full sample of 213 items, and where equated delta (an index of item difficulty) is the dependent variable, we found 33 percent ($p < .0001$) of the variance of item difficulty could be accounted for by eight variables. All eight variables reflected significant and independent contributions due solely to text and text/item overlap variables. This result provided evidence favoring construct validity of the TOEFL reading comprehension items. We also conducted a separate analysis of a subset (n=98) of the full set of 213 items to examine the possible statistical effect of nesting in the original sample. (Nesting occurs when several items relating to the same passage are analyzed together; a non-nested subset is formed when only one item per passage is used.) Eleven variables accounted for 58 percent ($p < .0001$) of the variance of this non-nested sample. Ten of these 11 variables reflected significant and independent contribution of text and text/item overlap variables. Hence this subanalysis provided further support for construct validity. While both analyses provide evidence f voring construct validity of the TOEFL reading items, the differences in amount of variance accounted for suggests that nesting effects should be a concern in future studies predicting item difficulty.

Additional regression analyses explored the adequacy of our predictor variables to predict performance for candidates who were classified into five TOEFL ability levels based on their total TOEFL scores. These data were analyzed twice: once for the nested sample (n=213 items) and again for the non-nested (n=98) sample. The regression results indicated that for the full sample of 213 items, 39 percent of the variance ($p < .0001$) of reading item difficulty could be accounted for in the lowest ability group (the lowest-scoring 20 percent of the examinees); but only 14 percent of the variance ($p < .0001$) of the highest-scoring examinees (the top 20 percent of the examinees) could be accounted for. For the non-nested sample, the results showed that 61 percent ($p < .0001$) of the lowest- and 36 percent ($p < .0001$) of the highest-ability group could be accounted for. These results suggest that our set of predictor variables works best for the lowest-scoring TOEFL examinees. It again shows the importance of studying the effects of nested vs. non-nested data samples when conducting future work in this area.

All analyses cited above were conducted using the same set of predictor variables. In conclusion, we demonstrated the following: (1) reading item difficulty can be significantly predicted by variables similar to those reported in the experimental literature on language comprehension, (2) the TOEFL reading items examined here appear to be construct valid, (3) the variables studied predict better the lower ability examinees in comparison with higher ability examinees, and (4) the statistical problem of nesting needs to be taken into account in predicting reading item difficulty.

Acknowledgements

We want to thank Grant Henning, Gordon Hale, and Janice Scheuneman for helpful comments on earlier drafts of this paper. This work was funded by the TOEFL Research Committee.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and, in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

✣ ✣ ✣

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1992-93) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins (Chair)	Southern Illinois University at Carbondale
Linda Schinke-Llano	Millikin University
John Upshur	Concordia University

Table of Contents

Page

List of Tables

Introduction

Purpose of Current Study

The primary purpose of the current study is to predict reading item difficulty for three TOEFL reading item types: main subject or main idea items (henceforth we shall refer to this category more briefly as main idea items), inferences, and supporting idea items which together constitute about 75% of the reading items. To achieve this goal we need to identify a set of variables that earlier studies suggest should be predictive of comprehension difficulty. Confirming evidence that these earlier identified variables, as they apply to the passage content and structure, are predictive of TOEFL reading comprehension item difficulty, can be taken as evidence favoring the claim that the TOEFL reading section is in fact a measure of passage comprehension--that is, that multiple-choice tests of reading comprehension are construct valid. Such an outcome might lead to modifications in statements made recently by Royer (1990) as well as Katz, Lautenschlager, Blackburn, and Harris (1990) who have argued that multiple-choice reading tests are primarily tests of reasoning rather than passage comprehension--these arguments are presented in greater detail below.

Background Studies

Only a few studies appear to have focused on predicting language comprehension item difficulty using items from standardized ability tests (Drum, Calfee & Cook, 1981; Embretson & Wetzel, 1987; Freedle & Fellbaum, 1987; Freedle & Kostin, 1991; 1992). While not specifically focused on predicting language comprehension item difficulty, many other studies of language processing have isolated a wide variety of variables that influence comprehension difficulty with respect to decision time and recall measures. A few such studies of particular interest here are the study of negations (Carpenter & Just, 1975), the study of rhetorical structure (Grimes, 1975) and its effect on accuracy of prose recall (Meyer, 1975; Meyer & Freedle, 1984) and accuracy of prose comprehension (Hare, Rabinowitz & Schieble, 1989), the use of referential expressions in constructing meaning (Clark & Haviland, 1977), and the use of syntactic "frontings" (see details below) which appear to guide the interpretations of semantic relationships within and across paragraphs (see Freedle, Fine & Fellbaum, 1981; also see Stark, 1988). The particular manner in which these selected variables will be employed will become evident later in this report. Using this set of hypothetically relevant variables, the primary strategy employed in this work has been to try to capture the large- and small-scale structures of the reading passages, and their associated items, in order to best account for observed reading item difficulty in a multiple-choice testing context.

First we review those studies that predict item difficulty for language comprehension multiple-choice tests.

Drum, Calfee, and Cook (1981) predicted reading comprehension item difficulty using various surface structure variables and word frequency measures for the text, and several item variables which also depended on surface structure characteristics (e.g., number of words in the stem and options, number of words with more than one syllable, etc.). They reported good predictability using these simple surface variables; on average, they indicate that about 70 percent of the variance[1] of multiple-choice reading item difficulty was explained.

Embretson and Wetzel (1987) also studied the predictability of 75 reading item difficulties using a few of the surface variables studied by Drum et al. (1981). But, in addition, because of the brevity of their passages, Embretson and Wetzel (1987) were able to do a propositional analysis (see Kintsch & van Dijk, 1978) and add variables from this analysis, along with several other measures, as predictor variables. In particular they found that connective propositions were significant predictors. We believe that Meyer's (1975) top-level rhetorical structures, which we include in the present study, indirectly assess the presence of connectives (such as and, but, however, since because, etc.) since each of the rhetorical devices differently emphasizes these connectives. For example, a top-level causal structure tends to use connectives such as since and because. A list structure tends to use connectives such as and and then, while a comparative structure will often employ connectives such as however, yet, etc.

Freedle and Fellbaum (1987) found that lexical overlap helps to account for multiple-choice item difficulty in the TOEFL test of single sentence comprehension (i.e., TOEFL's listening comprehension section) such that item options that contain greater lexical overlap with the presented stimulus sentence tend to be the options that get selected by the test takers; however, this tendency was most prevalent among the lower ability examinees and was virtually absent from the higher ability examinees. It is interesting to question whether such a simple strategy of lexical overlap will play any substantial role in predicting reading comprehension item difficulty as a function of overall verbal ability.

Before we review the findings of Freedle and Kostin (1991; 1992) for predicting the difficulty level of SAT and GRE reading items, it is necessary to review the earlier experimental literature to understand why the set of predictor variables used in their studies was chosen. Following this brief review we shall take up in greater detail their findings.

A number of studies have dealt with variables that have been found to influence the difficulty of reading comprehension. Most of these additional variables were investigated in empirical studies which did not use multiple-choice methods to yield an index of comprehension difficulty. Instead, many

[1] While the Drum et al. (1981) study was innovative in analyzing the multiple-choice testing process into its constituent parts (i.e., determining the relative contribution of the item's stem, the item's correct and incorrect options, as well as the text variables to item difficulty), some of the study's analyses appeared to be flawed. Ten predictor variables were extracted from very small reading item samples (varying between 20 and 36 items) taken from seven children's reading tests. At most two or three predictors instead of 10 should have been extracted fro such small samples (see Cohen & Cohen, 1983); hence 70 percent of the item difficulty variance is probablv too large an estimate of the variance actually accounted for.

used dependent measures such as recall of passages or decision time to infer the influence that certain variables have on comprehension difficulty. (Examples of these variables can be found in the materials and method section of this report.)

Carpenter and Just (1975) found that sentence negations typically increase comprehension decision time in comparison with sentences without negations. (This suggests that the number of negations contained in TOEFL reading passages may also influence multiple-choice item difficulty.) Furthermore, one can inquire whether additional negations that are used in the item structure itself (either in the item stem and/or among the response options) may also separately contribute to comprehension difficulty over and above the contribution of text negations.

Abrahamsen and Shelton (1989) demonstrated improved comprehension of texts that were modified, in part, so that full noun phrases were substituted in place of referential expressions. This suggests that texts with many referential expressions may be more difficult than ones with few referential expressions. Again, for purposes of studying more broadly the effect of number of referential expressions on comprehension difficulty of multiple-choice tests, a separate count can also be made of referential expressions that occur in the item proper.

Hare et al. (1989) studied, in part, the effect of four Grimes' (1975) rhetorical organizers on the difficulty of identifying the main idea of passages--students either wrote out the main idea if it wasn't explicitly stated or underlined it if it was explicitly stated. They found a significant effect of rhetorical organization such that list type structures (see definitions and examples below) facilitated main idea identification whereas some non-list organizers made main idea information more difficult to locate. Meyer and Freedle (1984) examined the effect of Grimes' organizers on the ability of students to recall passages which contained the same semantic information except for their top level rhetorical organization. They found, like Hare et al. (1989), that list structures facilitated recall (for older subjects). However, they also reported that university students were best helped by comparative type organizations; this latter finding was not confirmed by Hare et al.

It seems likely that rhetorical organization will contribute to comprehension difficulty within a multiple-choice testing format; however, it is not clear, given the differences between Meyer and Freedle (1984) and the Hare et al. (1989) studies, whether we can say in advance which type of structure will be found to facilitate performance. Top level rhetorical structure meaningfully applies only to the text structure; a comparable entry for items is not feasible.

Freedle, Fine, and Fellbaum (1981) report differences in the use of "fronted" structures at sentence beginnings (and paragraph beginnings) as a function of the judged quality of student essays. Fronted structures included the following: (1) cleft structures ("It is true that she found the dog," where the initial "it" is a dummy variable having no referent), (2) marked topics consisting of several subtypes (a) opening prepositional phrases or adverbials ("In the dark, all is uncertain"; "Quickly near the lodge, the boat overturned") or (b) initial subordinate clauses ("Whenever the car stalled, John would sweat") and (3) combinations of coordinators and marked topics or cleft structures that begin independent clauses ("But, briefly, this didn't stop him"; "And, furthermore, it seems that is all one should say").

Freedle et al. (1981) showed that the better essays contained a significantly higher mean frequency of each of these fronted structures even after partialling out the effect of different lengths of essay as a function of ability level. They interpreted these fronted structures as authors' explicit markers for guiding readers to uncover the relationships that exist among independent clauses. It is not immediately clear whether differential use of all such structures would itself facilitate or inhibit comprehension of TOEFL passages. If we assume that the structures produced by the more able writers are structures that are more difficult to learn, then we can predict that the more frequently these fronted structures occur, the more difficult the text should be to understand. In support of this, Clark and Haviland (1977) suggest that at least cleft structures may be harder to understand than simple declarative sentences. Also Bever and Townsend (1979) found that when main clauses follow a subordinate clause, such sentences are more difficult to process than when main clauses occur in the sentence's initial position. This finding overlaps somewhat with frontings, since initial subordinate clauses would count as one type of fronting. By including a count of all such variables we can explicitly test the relevance of clefts and other fronted structures for their effect on comprehension difficulty in a multiple-choice testing context. This can be done separately for text as well as item content.

While Kieras (1985) specifically focused on the perception of main idea information in reading, his study will be seen as potentially relevant for all three item types treated in our study. Some examples of what we collectively have called main idea items are as follows.

a) Main subject of passage: "What does the passage mainly discuss?"
b) Main idea of passage: "What is the author's main point?"
c) Author's purpose: "What is the main purpose of the passage?"

Next we summarize Kieras' (1985) earlier work and then generalize to include inference and supporting idea items.

Kieras (1985) examined, in part, how students perceived the relative location of main idea information in short paragraphs. Using single paragraph passages extracted from technical manuals, he found that most students perceive main idea information as located early in the paragraph. A few thought the main idea occurred at or near the end of the paragraph. Students least often perceived information in the middle of the paragraph as a statement of the main idea. Kieras (1985) did not report the relative frequencies with which the main ideas actually occurred among the passages. Consequently it is difficult to know whether students tend to select the opening sentences of passages as containing the main idea because most of the passages placed the key idea in this place, or whether the students were simply reflecting a response bias to choose the opening sentences. Unless the main idea was equally represented by its location across the stimulus passages, the Kieras results are ambiguous.

However, the work of Hare et al. (1989) helps to clarify this issue. In one of their studies they systematically varied the main idea sentence in three locations: the opening sentence, the medial sentence, or the final sentence of a paragraph. The students underlined the sentence they thought was the main idea sentence. Correct identifications were greatest for initial occurrence of main idea sentences. One can infer from the Hare et al. results that two tendencies contribute to main idea correctness: opening sentences that do contain the main idea tend to be selected partly because of a prior bias to select early sentences, but also because students are attempting to understand the information in the text sentences.

One can generalize the Hare et al. (1989) work including the Kieras (1985) findings to demonstrate the possible relevance of locational effects concerning the way that students respond to multiple-choice items for multi-paragraph passages. If students tend to perceive early text information, especially information in the opening sentences of the first paragraph, as main idea information then when certain passages actually confirm this search strategy, such items should be easier than those that disconfirm it (where disconfirming main idea information would be information that occurs in the middle of a multi-paragraph text; it is disconfirming only because it fails to conform to the expectation that main idea information "should" be near the beginning of a passage). So, the relative ordering of difficulty should be: opening sentences that fit the main idea information as stated in the correct answer to a main idea item will be easiest (other things being equal), while main idea information that occurs near the middle of a text will be associated with the hardest main idea items.

Bhasin (1990) reported a study of main idea comprehension which suggested that bilingual Spanish-English students tend to focus on initial text information in helping them to select a response option in a multiple-choice comprehension test (Descriptive Tests of Language Skills).

Since we also intend to study inference as well as supporting idea items, we might inquire whether the Kieras (1985) and Hare et al. (1989) type findings about relative location of information in the passage for main idea items will also help account for item difficulty associated with these other two reading item types.

TOEFL's supporting idea items are of the following type: "According to the passage, x occurs when ..." It seems reasonable to expect that if the relevant supporting information occurs early in the passage, the item should tend to be easy. But if the relevant information is located near the middle of the passage, this should make such an item more difficult. If so, then this generalizes our interpretation of Kieras' (1985) results for main ideas to supporting idea items. We hypothesize that the surface location of relevant information influences the results. While one normally expects early text information to contain the relevant main idea, there is no corresponding expectation for supporting idea information. Nevertheless, the beginning of a passage may be especially salient even for supporting idea items, not because a prior expectation is confirmed or disconfirmed, but simply, because examinees may start their search for such information at the beginning of the passage.

A similar argument can be made for inference type items. Inference items usually have the following format: "It can be inferred from the passage that x ..." If the relevant text information needed to carry out the inference is located near the beginning of the passage, this might facilitate choosing the correct option. But if the relevant text information is in the middle, this might make the item more difficult.

Other variables that we can hypothesize will be of importance in affecting comprehension difficulty for multiple-choice tests are: vocabulary level (Graves, 1986), various measures of sentence complexity such as sentence length (Klare, 1974-75), passage length (Newsome & Gaite, 1971), paragraph length (Hites, 1950), number of paragraphs (Freedle, Fine, & Fellbaum, 1981) and abstractness of text (Paivio, 1986). In particular, longer sentence structures and longer and less frequently occurring words tend to make texts more difficult to understand, as can be inferred from their use in traditional readability formulas (see Graves, 1986); in addition, longer passages, longer paragraphs, and abstractness of texts also make _ ıssages more difficult to comprehend (see Newsome & Gaite [1971], Hites [1950], and Paivio [1986], respectively). Use of more paragraphs was positively correlated with the quality of written essays (Freedle, Fine, & Fellbaum, 1981); it remains to be demonstrated whether the number of paragraphs itself contributes to the difficulty of reading comprehension in a multiple-choice testing context.

Before we describe the findings of Freedle and Kostin (1991; 1992) it will be useful to collect the above broad review of variables, which are expected to influence reading comprehension item difficulty, into a single set. One can hypothesize that many of the variables listed, which are known to contribute to comprehension difficulty in non-multiple-choice testing formats (or to quality judgments of written essays), will be found to affect significantly comprehension measures as determined within a multiple-choice testing format. More succinctly:

Hypothesis 1. We expect the following variables to influence reading item difficulty significantly as determined within a multiple-choice testing format:

a. Negations: The greater the number of negations the more difficult the comprehension.
b. Referentials: The greater the number of referentials the more difficult the comprehension.
c. Rhetorical organizers: Based on past studies we predict that rhetorical organizers will significantly affect comprehension but we do not make a directional prediction.
d. Fronted structures: We predict the sum and each of the three fronted structures will make comprehension more difficult. The three fronted structures of interest are:
    1. Cleft structures.
    2. Marked topics.

3. Combinations (of coordinators and marked topics or coordinators with cleft structures).

e. Vocabulary: The more multisyllabic words used, the greater the comprehension difficulty.

f. Sentence length: The longer the sentence, the greater the comprehension difficulty.

g. Paragraph length: The longer the paragraph, the greater the comprehension difficulty.

h. Number of paragraphs: The more paragraphs, the greater the comprehension difficulty.

i. Abstractness of text: The more abstract the text, the greater the comprehension difficulty.

j. Location of relevant text information: Information located early in text will facilitate comprehension, whereas information located in the middle of a text will make comprehension more difficult.

k. Passage length: The longer the passage, the more difficult the comprehension.

l. Lexical overlap between text and options: The more lexical overlaps between the words in the correct option and the words in the text, the easier the item.

<u>The relevance of Hypothesis 1 to criticisms of multiple-choice reading tests as tests of passage comprehension</u>. Hypothesis 1, particularly as it applies to the coding of passage content, can be viewed as important to demonstrating the construct validity of a multiple-choice reading comprehension test, as we shall now endeavor to explain. Royer (1990) maintains that "There is evidence that standardized reading comprehension tests that utilize multiple-choice questions do <u>not</u> measure the comprehension of a given passage. Instead they seem to measure a reader's world knowledge and his or her ability to reason and think about the contents of a passage" (Royer, p. 162). Royer then cites work by Tuinman (1973-74), Drum et al. (1981), and Johnston (1984) to support this claim. Tuinman's work is similar to the findings of Katz et al. (1990) wherein multiple-choice reading items are correctly responded to above chance levels in the <u>absence</u> of the reading passage. This seems to imply that <u>item</u> structure and content alone is sufficient to guide an examinee's performance, while text structure and content may play a somewhat less important role than previously believed. This type of argument suggests how to test, in a more rigorous way, whether a multiple-choice test of reading comprehension is or is not construct valid. If one can show that variables that code for an <u>item</u>'s structure and content better correlate with reading test performance than do the variables that code for the <u>text</u>'s structure and content, then Royer's conclusion would seem to be correct. However, if one can show that text variables play a strong and significant role in multiple-choice reading comprehension test performance, then one can begin to call into question the argument that Royer appears to be making.

Of course Katz et al. (1990) have also shown that a significant increase in correct responses occurs when passages are available to a control group. Hence it seems that Royer (1990) appears to have overgeneralized the importance of item structure exclusively in concluding that multiple-choice reading tests do <u>not</u> measure passage comprehension. That is, if multiple-choice tests of reading did not tap passage comprehension and were solely a reflection of outside knowledge and reasoning ability (as implied by information in the items alone), then the subsequent addition of the passage should have had no noticeable effect on reading item correctness. Since Katz et al. clearly showed a significant augmentation of item correctness when the passage was available, one must conclude that multiple-choice reading tests do measure passage comprehension and simultaneously tap other abilities such as reasoning.

Royer's (1990) citation of Drum et al. (1981) also concerns the claimed importance of item structure exclusively to reading comprehension item correctness. The plausibility of the incorrect option was the most important predictor in the Drum et al. study. They classified this plausibility as an item variable. However, we claim that incorrect option plausibility is more accurately classified as a <u>text/item overlap variable</u>, and is not just an item variable. That is, in order to decide whether an incorrect option is a plausible answer or not, Drum et al. used not only the item information but the text information as well--in their study an incorrect option that was contradicted by the text was not rated as a plausible one. Hence Drum et al.'s best predictor is one that necessarily implicates the reading of the text. This leads us to conclude that Royer's acceptance of Drum et al.'s classification scheme led him to use their results, incorrectly we feel, to support further his hypothesis that text comprehension does not play a crucial role in multiple-choice reading tests.

But suppose Royer's (1990) critique of multiple-choice tests is assumed to be correct. Then there is little reason to expect that the variables listed under Hypothesis 1 (<u>a</u> through <u>l</u> above, at least as it applies to the coding of the text) will be significantly related to multiple-choice reading test item difficulty. This should follow because, by (Royer's) hypothesis, multiple-choice tests are not tests of comprehension; hence variables, known to be related to comprehension difficulty (in the experimental literature), should not correlate with performance on multiple-choice reading comprehension tests. However, if Royer is incorrect, then there is good reason to suppose that most if not all of the variables listed under Hypothesis 1, at least as applied to the coding of the text, will be found to correlate significantly with reading item difficulty as obtained from multiple-choice testing.

If supporting evidence is found for Hypothesis 1, there is a second implication that is important to evaluate. There are few studies that assess the simultaneous influence of many variables on comprehension (Goodman, 1982). With the current TOEFL passages it should be possible to evaluate, via regression analyses, whether the twelve categories of variables of Hypothesis 1 contribute <u>independent</u> information in accounting for reading comprehension item difficulty. This leads us to our second hypothesis.

<u>Hypothesis 2</u>. Many of the twelve categories of variables provide <u>independent</u> predictive information in accounting for reading item difficulty.

Now we shall review the two studies of Freedle and Kostin (1991; 1992) that used a large sample of reading items for predicting multiple-choice item difficulty. (The reader should note that Freedle and Kostin [1991; 1992] evaluated only the first eleven categories of Hypothesis 1 above for both the SAT and GRE reading data.)

Using 110 SAT main idea reading items, Freedle and Kostin (1991) presented correlational evidence favoring six of the first eleven categories of Hypothesis 1. For Hypothesis 2, regression analyses of the SAT data indicated that five of the eleven categories provided independent predictive information concerning item difficulty.

For 244 GRE reading items involving main idea, inference and explicit statement items, the comparable results regarding Hypotheses 1 and 2 were as follows. Evidence favoring seven of the eleven categories listed under Hypothesis 1 was found when just the correlational evidence was examined. Pooling the regression results for each of the three item types, Freedle and Kostin (1992) found six of the 11 categories provided independent predictive information.

This brief review of the Freedle and Kostin (1991; 1992) findings suggests that there is evidence favoring both Hypotheses 1 and 2 for multiple-choice testing formats as evidenced by the SAT and GRE reading items. Furthermore, since most of the above significant categories were represented by <u>text</u> variables, Freedle and Kostin were able to conclude that evidence exists favoring construct validity of multiple-choice reading comprehension items. They concluded this because the difficulty of such multiple-choice reading items is more closely associated with text as opposed to item variables.

Materials and Method

The 213 reading comprehension items taken from 20 TOEFL forms constitute the total item sample. One hundred reading passages were represented. There were five passages per test form. The five passages in each test form cover the following subject matters: arts, humanities, social sciences, life sciences, and physical sciences. Only <u>main idea</u> (n=59), <u>inference</u> (n=61), and <u>supporting ideas</u> (n=93) items were selected for study. Other item types, such as author's tone and author's organization, occur infrequently and were not scored. We also did not sample items that use a format in which different combinations of three elements constitute the list of options as in (a) only I is corre(, (b) only I and II are correct, (c) I and III are correct, (d) II and III are correct, (e) none are correct. We also excluded special items which featured a capitalized NOT or LEAST in the item stem.

The data for each item difficulty measure (equated delta) were based on approximately 2,000 examinees; these examinees were randomly selected from a much larger pool of examinees who responded to each TOEFL test form. The equated delta value slightly adjusts the difficulty of each item across forms so that items can be meaningfully compared across groups of people taking different test forms. The adjustment stems from the fact that the examinees who respond to a particular test form differ slightly in overall ability level from those responding to other test forms. The delta of each test form is adjusted so that it has a mean of 13.0 and a standard deviation (S.D.) of 4.0.

The five ability levels are determined on a form by form basis. The distribution of TOEFL scores is divided evenly into five parts so that the lowest 20 percent represents the lowest ability group, the next lowest group represents people who received somewhat higher scores, while the highest ability group represents the 20 percent of the examinees who received the highest scores. A large percentage of examinees (in excess of 45 percent) in the highest group happen to be German, while a similarly high percentage (in excess of 35 percent) of the lowest group happen to be Arabic--see Alderman and Holland, 1981.

Two main data bases were used for most of the analyses below. The first data base consisted of the following: 213 items (59 main idea items, 61 inferences, and 93 supporting ideas); for any given passage at least one item was used in this analysis; for some passages more than one item was associated with each passage. However, no passage contained more than one main idea item, one inference item, and one supporting idea item--that is, no passage contained two inferences or two supporting ideas. This large sample of 213 items therefore can be described as a "nested" data sample inasmuch as there is not just one item per passage represented. Our second major data base consisted of 98 items and 98 passages (this represents a subset of the 213 item set); it contained 32 main ideas, 33 inferences, and 33 supporting idea items. Since each passage was represented by just one item, this data base is described as a non-nested sample.

Most of the independent variables listed below were motivated by the literature review presented above. These, along with a few additional variables (e.g., number of rhetorical questions in the passage, type of passage subject matter, lexical coherence across text paragraphs), had been used in our earlier studies (Freedle & Kostin, 1991; 1992) but were not specifically involved in evaluating Hypotheses 1 and 2 described above.

In addition to the literature review, another factor involved in selecting many of the predictor variables was the ease of automatically scoring the variable. For example, regarding estimating vocabulary difficulty, it is relatively easy to program a computer to score number of syllables and somewhat more difficult to score, say, the affective connotations of a word (e.g., whether it evokes negative or positive emotions); to do this automatically would require constructing a large table of subjectively rated 'ords. This does not deny the possible importance of word connotations, especially since unpublished work for SAT reading items (by author R.F.) indicates that a significant relationship does exist between item difficulty and the number of words in the item having negative emotional connotations. But the goal of choosing primarily easily automated variables led to scoring number of syllables rather than alternative methods when estimating the possible effects of vocabulary on item difficulty prediction.

Independent Variables for Representing Text and Item Information

Item Variables

Item type
v1 --Main idea
v1a--Z-1: Main subject of the passage.
v1b--Z-2: Main idea of the passage.
v1c--Z-3: Author's purpose.
v2 --Inference
v3 --Supporting idea

Variables for item's stem
v4 --Words in stem: Number of words in stem (the item question).
v5 --Fragment stem: Use of full question or sentence fragment.
v6 --Negative stem: Use of negation (e.g., use of "no," "never," "neither," "none," "no one," etc.; in addition, prefixed words such as "uncover," "impossible," "disheartened," and suffixed words such as "relentless" were also counted as instances of negation).
v7 --Fronted stem: Use of fronting (e.g., use of any phrases or clauses preceding the subject of the main independent clause, or use of clefts--for details see description below for v46 to v50).
v8 --Reference stem: Sum of referentials to text, other parts of stem, or options. (See below for definitions under text variables v55 to v57.)

Variables for item's correct option
v9 --Answer position: Ordinal position of correct answer.
v10--Words correct: Number of words in correct option.
v11--Negative correct: Use of negation(s) in correct option.
v12--Fronting correct: Use of fronting(s) in correct option.
v13--Reference correct: Use of referential(s) in correct option.

Variables for item's incorrect options
v14--Words in incorrects: Number of words summed over all incorrect options.
v15--Negative incorrects: Use of negation(s) summed over incorrect options.
v16--Fronted incorrects: Use of fronting(s) summed over incorrect options.
v17--Reference incorrects: Use of referential(s) summed over incorrect options.

Text Variables

Vocabulary variable for text
v18--Vocabulary: Number of words with three or more syllables for the first 100 words of the passage (estimates vocabulary difficulty--see Gunning, 1964).

Concreteness/abstractness of text

v19--Concreteness: Determines whether main idea of text and its development is concerned with concrete or abstract entities (1=abstract to 5=concrete).

Subject matter variables of text

v20--Physical science

v21--Life science

v22--Natural science: Combined v20 and v21 into a single natural science variable.

v23--Social science: Subjects such as anthropology, economics, sociology, political science.

v24--Humanities: Subjects such as history, philosophy, etc.

v25--Arts: Fine arts, architecture, literature, and music.

v26--Natural science excerpt: Represents an "excerpt of natural science" (that is, this could be a section of a scientific study).

v27--About natural science: Represents a passage "about natural science" (this is not an excerpt from a scientific study but is a commentary that concerns the topic of science).

For v20 through v25, the classification of subject matters was based on TOEFL's subject matter classifications.

Type of rhetorical organization

v28--Argument: Rhetorical presentation (i.e., author favors one of several points of view presented in text; occasionally other viewpoints may be only implied).

v29--List/describe: This Grimes' (1975) rhetorical organizer interrelates a collection of elements in a text that are related in some unspecified manner; a basis of a list "... ranges from a group of attributes of the same character, event, or idea, to a group related by simultaneity to a group related by time sequence" (Meyer, 1985, p. 270). "Describe" relates a topic to more information about it. We felt this was sufficiently similar to "list" to warrant scoring them as members of the same category.

v30--Cause: This is another Grimes' (1975) rhetorical organizer. "Causation shows a causal relationship between ideas where one idea is the antecedent or cause and the other is a consequent or effect. The relation is often referred to as the condition, result or purpose with one argument serving as the antecedent and the other as the consequent. The arguments are before and after in time and causally related." (Meyer, 1985, p. 271).

v31--Compare: Yet another Grimes' (1975) rhetorical organizer. The comparison relation points out differences and similarities between two or more topics.

v32--Problem/solution: This is defined as follows: "... similar to causation in that the problem is before in time and an antecedent for the solution. However, in addition there must be some overlap in topic content between the problem and solution; that is, at least part of the solution must match one cause of the problem. The ... problem and solution ... are equally weighted and occur at the same level in the content structure." (Meyer, 1985, p. 272).

Coherence of lexical concepts over whole text

v33--Coherence: This involves judging whether opening concepts of the first sentence occur throughout the text paragraphs. 3= maximum lexical coherence to 0 = no obvious lexical overlap.

Lengths of various text segments

v34--Paragraphs: Number of passage paragraphs.

v35--Text words: Number of words in passage.

v36--Text sentences: Number of text sentences.

v37--First paragraph words: Number of words in first paragraph.

v38--Longest paragraph words: Number of words in longest paragraph.

v39--First paragraph sentences: Number of sentences in first paragraph.

v40--Longest paragraph sentences: Number of sentences in longest paragraph.

v41--Independent clauses: Number of independent clauses in total text.

v42--Text sentence words: Average number of words per text sentence.

v43--Text paragraph words: Average number of words per paragraph.

v44--First paragraph sentence length: Average length of sentences in first paragraph.

v45--Longest paragraph sentence length: Average length of sentences in longest paragraph.

Occurrence of different text "frontings"

V46 through v50 distinguishes several types and combinations of "frontings." Some examples follow. Use of theme-marking: In the background, the scenery changed. Fortunately, the man escaped. Use of coordination: But, the car rocked. Use of clefts (deferred foci): It is the case that George is short. There are cases that defy reason. (It and there function as dummy elements without a referent.) Use of combinations: And, near the chair, the toy fell. Longest run of frontings: Number of successive independent clauses which begin with fronted information: e.g., "The man laughed. Then, he frowned. And when he turned, he fell." This example of three independent clauses has two successive sentences with fronted material; hence its run length is "2."

v46--Percent fronted text clauses

v47--Frequency fronted text clauses

v48--Frequency combinations of fronted text structures

v49--Frequency of text clefts: This is sometimes referred to as deferred foci that is one type of fronting.

v50--Longest fronted run: Number of consecutively fronted text clauses.

Text questions

v51--Text questions: Number of rhetorical questions in text.

Text special punctuations

v52--Semicolons: Number of semicolons used in text.

v53--Colons: Number of colons used in text.

v54--Dashes: Number of dashes used in text.

Text referentials

v55--Reference within text clauses: Frequency of within-clause referentials of all text clauses, e.g., "When George fell, *he* was hurt."

v56--Reference across text clauses: Frequency of across-clause referentials, e.g., "George fell. *That* hurt."

v57--Frequency special reference: Reference outside text, e.g., "*One* might feel sorry for George."

v58--Reference sums: Sum of v55, v56, and v57.

Text negations

v59--Text negatives: Number of negations in text.

Text/Item Overlap Variables

Overlap variables that apply to every item type

v60--Number of fronts in key overlapping sentence

v61--Number of referentials (within clauses) in key overlapping sentence

v62--Number of referentials (across clauses) in key overlapping sentence

v63--Number of referentials (outside clauses) in key overlapping sentence

v64--Number of negations used in key overlapping sentence

v65--Number of independent clauses in key sentence or sentences

Text/item overlap variables applicable only to main idea information

mm1--Main idea first sentence: Main idea information in first sentence of text.

mm2--Main idea middle text: Main idea information in near middle of passage.

mm3--All early main idea locations: Sum of the following: the main idea occurs in the first sentence, and/or the main idea occurs in the second sentence, and/or the main idea occurs later in the first short paragraph of 75 words or less.

mm4--First line lexical match: Ordinal position of the earliest word on the first line that overlaps with a content word in the correct answer of a main idea item.

mm5--Related words plus mm4: Same as mm4 but includes lexically related words.

mm6--First line lexical match for incorrects: If there is no lexical overlap on first line for correct option but there is for one or more of the incorrect options, or, if there is an overlap on the first line for the correct but the overlap for the incorrect comes in an earlier ordinal position than the correct option overlap.

Text/item overlap variables applicable to inferences

ii1--Unique word same sentence: Stem sends you to unique word in text and relevant information is in same sentence.

ii2--Information in last sentence: Relevant information is in last sentence of text.

ii3--Information middle of text: Relevant information is located more in middle of text.

ii4--Words before critical information: Number of words in passage you have to read before the sentence containing the relevant information begins.

ii5--Words in relevant paragraph: Number of words in paragraph in which the relevant information is located.

ii6--Information middle relevant paragraph: Relevant information is in the middle of a paragraph rather than the first or last sentence of that paragraph.

ii7--Number lexically matched words: Number of words in correct answer that overlap with words in key text sentence.

ii8--Related words plus ii7: Same as ii7 but includes lexically related words.

ii9--Number words in key text sentence: Number of words in key text sentence containing the relevant inference information.

ii10--Percent lexically matched words: Percent words in correct answer that overlap with words in key text sentence.

ii11--Related words plus ii10: Same as ii10 but includes lexically related words.

Text/item overlap variables applicable to supporting idea items

ss1--Unique word, same sentence: Stem sends you to unique word in text and relevant information is in same sentence.

ss2--Key word occurs in multiple places: Stem suggests a particular topic, but that topic is mentioned in more than one sentence in passage.

ss3--Information is in middle of text: Relevant information located in middle of text.

ss4--Words before critical information: Number of words in passage you have to read before the sentence containing the relevant information begins.

ss5--Words in relevant paragraph: Number of words in paragraph in which the relevant information is located.

ss6--Information middle relevant paragraph: Relevant information is in middle of paragraph rather than the first or last sentence of that paragraph.

ss7--Number lexically matched words: Number of words in correct answer thaː overlap with words in key text sentence.

ss8--Related words plus ss7: Same as ss7 but includes lexically related words.

ss9--Number of words in key text sentence: Number of words in key text sentence containing the relevant supporting idea information.

ss10--Percent lexically matched words: Percent words in correct answer that overlap with words in key text sentence.

ss11--Related words plus ss10: Same as ss10 but includes lexically related words.

Dependent Variables

v66--Item difficulty: Item equated delta (referred to as just "delta").

v67--z-score for lowest ability group

v68--z-score for 2nd lowest ability group

v69--z-score for middle ability group

v70--z-score for 2nd highest ability group

v71--z-score for highest ability group

In scoring items, the structure and content of item stems, correct options, and incorrect options were recorded using the 20 variables listed above (three of these 20 being the code for main idea sub-item types; an additional variable, v1, represents the collective set of main idea items which is intended to replace the three main idea subtypes should these three subtypes prove to be of equal difficulty). Another set of 42 variables listed above was scored for capturing the text (i.e., passage) information. A final set of 34 variables also listed above represents the overlap of text and item information: six of these apply to all three item types; another six apply only to main idea items which overlap the text in special ways (see descriptions above), another 11 represent inference items interacting with the text and a final 11 represent supporting idea items interacting with the text.

The dependent variable v66 is an item's equated delta (an item's difficulty that converts percent corrects per test form to a common scale with mean 13.0 and S.D. of 4). See above for a more detailed desc-iption of equated delta.

The dependent variables v67, v68, v69, v70, and v71 are the z-score transformations of the percent pass scores for each ability group. (The percent pass scores are available from each item statistics card.)

Deletion of Variables Due to Colinearity Among Predictor Variables or Low Frequency of Occurrence

Intercorrelations among all predictor variables were examined for colinearity (which is defined as variables correlating .80 or more--see Nie, Hull, Jenkins. Steinbrenner, & Bent, 1975). Where there were clusters of variables correlating .80 or more, several analyses were run, until each specific variable within a cluster was located that accounted for the most variance. These specific variables were then retained for the final set of analyses, which are reported in the results section. The final list is presented in Table 1. This led to the deletion of the following variables: v10, v17, v22, v37, v39, v41, v42, v43, v45, mm1, mm4, ii7, ii10, ss7, and ss10. Because of low frequencies of occurrence (defined as two or fewer occurrences in either or both of our samples of items where n=213 and n=98); the following variables were deleted: v12, v16, and v51. We also deleted the three subtypes of main idea items: v1a, v1b, and v1c. An ANOVA evaluated whether these three subtypes were significantly different in difficulty. The results showed that they are not different: F(2,56) = 1.50, p = .23. Hence, there is no need in any of the analyses of main idea items to distinguish further among the three different main idea subtypes.

Table 1 (page 33) shows the variables that were used in the final analyses.

Relationship between the Scored Variables and the Categories Listed under Hypothesis 1

a. Negations: The variables relevant to this category are v6, v11, v15, v59, and v64.
b. Referentials: The variables relevant to this category are v8, v13, v55, v56, v57, v58, v61, v62, and v63.
c. Rhetorical organizers: The variables relevant to this category are v28 through v32.
d. Fronted structures: The variables grouped under this category are v7, v46 through v50, and v60.
e. Vocabulary: v18.
f. Sentence length: The variables grouped under this category are v44, ii9, and ss9.
g. Paragraph length: This category includes v38, v40, ii5,and ss5.
h. Number of paragraphs: v34.
i. Abstractness of text: v19.
j. Location of relevant text information: mm2, mm3, ii2, ii3, ii6, ss3, and ss6.
k. Passage length: v35, v36, ii4, and ss4.
l. Lexical overlap between text and options: mm5, mm6, ii8, ii11, ss8, and ss11.

The final set of variables (as shown in Table 1) includes 13 item variables, 34 text variables, and 28 text/item overlap variables.

Reliability of variables requiring subjective judgment. While many of our predictor variables are arrived at objectively (e.g., by counting the number of words in a passage), the following required some degree of subjective judgment: coherence, referentials, negations, frontings, Grimes' rhetorical predicates, location of relevant text information for answering an item, abstractness/concreteness, and about natural science vs. natural science excerpt. The following percentage agreement was obtained for two raters using a sample size of 35 cases:

Coherence = 74 percent agreement
Referentials = 92 percent agreement
Negations = 96 percent agreement
Frontings = 93 percent agreement
Rhetorical predicates = 89 percent agreement
Location of relevant text = 84 percent agreement
Abstractness/concreteness = 87 percent agreement
About natural science vs. natural science excerpt = 97 percent agreement

In general it is clear that these subjective measures yield high reliabilities.

Results and Discussion

It is necessary to determine whether each of the three item types needs to be analyzed separately.

ANOVAs to Determine Significance of Three Reading Item Type Effects and MANOVAs to Determine Their Possible Interactions with Predictor Variables

We used a one-way ANOVA to determine whether the three reading item types (main ideas, inferences, and supporting ideas) significantly differ in difficulty. They do not differ significantly, $F(2, 210) = 1.67$, $p = .19$.

We also conducted a series of MANOVAs to help us determine whether there are significant interactions between predictor variables and the three item types. Fifty of the 75 variables (see Table 1, variables v4 through v65) represented item and text variables and a few text/item overlap variables that applied to every item type. Analyses of these 50 predictor variables (those for which an interaction analysis was meaningful) showed that only three of the variables (v50, v56, and v58) yielded a significant interaction with the three reading item types. Since three significant interactions are expected based on chance alone we can conclude that there is little statistical support for conducting separate correlation or regression analyses of each of the three item types. Hence, below we present analyses using all item types together either using a nested sample (n=213) or a non-nested sample (n=98).

Table 2 (page 36) presents data that help to identify those variables that are statistically significant in predicting reading item difficulty. In Table 2 we see that 32 different variables--in either or both of the two samples presented in the table--yield a significant correlation with item difficulty (equated delta). First, we will use portions of Table 2 to assess the apparent adequacy of Hypothesis 1 for each of the 12 categories listed under the hypothesis.

We are primarily interested here in whether the text and the text/item overlap variables satisfy the categories of Hypothesis 1. Again, our interest is due to our interpretation of Royer's (1990) critique that suggests that text (and, presumably, text/item overlap) variables should not yield significant category effects. Because of Royer we also point out when significant category effects hold for the item variables as well.

Correlates of the Difficulty of Reading Items as Determined by the Categories of Hypothesis 1--Based on Table 2 Results

a. Consistent with our Hypothesis 1a, correlations with several measures involving negations were significant: v64, the text/item overlap negations (the number of negations in that part of the text which is crucial to identifying the correct option) was significant. That is, the more negations present in the text overlap section, the harder the item. However, for the item variables we see that v11 (negations in the correct option) and v15 (negations in the incorrect options) also contribute significantly to item difficulty, such that the more negations in the correct and/or incorrect options the harder the item. (The broader measure of negations which were used throughout the passage--v59--was not significant and hence does not appear in Table 2.)

b. Correlations involving several referential variables are significant. Variables v56 and v58 are significant. V56 refers to frequency across clause referentials while v58 refers to the sum of all referentials (v55+v56+v57); both are text variables. There is one significant text/item overlap variable: v61. The more referential pronouns within the overlap text clauses the harder the item tends to be (v61).

c. In line with our general prediction we see that two text rhetorical organizers (v29 and v32) are significantly correlated with item difficulty. (Rhetorical organizers were not applied to overlap nor item variables.)

d. Consistent with prediction, the percent and number of fronted structures in the text as measured by variables v46 and v47, respectively, was related to item difficulty. The more fronted text structures present, the more difficult the item.

e. Vocabulary (v18) contributes to item difficulty. The more polysyllabic text words used (here, words having three or more syllables per the first 100 text words), the harder the items associated with such texts tend to be. (Vocabulary text/item overlap and vocabulary item scores were not coded.)

f. The measure of text sentence length contributes to item difficulty (v44, average words per sentence for first paragraph). Also an overlap variable ii9 (number of words in key text sentence containing relevant inference information) contributes to item difficulty.

g. Two variables--v38, ii5--relate to text paragraph length effects. V38 (number of words in longest paragraph) and ii5 (number of words in the paragraph containing relevant inference information) influence item difficulty such that the longer the paragraph the more difficult the item. (This concept does not apply to items.)

h. Number of paragraphs. There are no relevant text results for this category. (There is no equivalent for text/item overlap and item variables.)

i. As predicted, the concreteness (v19) of the text showed a significant effect. (Concreteness of text makes these items easier.) (Neither text/item overlap nor item score for concreteness were coded.)

j. As predicted, the following three text location variables are significantly correlated in the expected direction with reading difficulty: mm2 (main idea information is in the middle of passage), ii3 (relevant information is in middle of passage), and ii6 (relevant inference information is in the middle of a text paragraph rather than in the first or last sentence of the paragraph). Note: category j applies only to text/item overlap variables; it does not apply to the pure text or pure item variables.

k. Passage length v35 (number of words in passage) has a significant effect on item difficulty. This shows that items associated with long passages are more difficult. Variable ii4 (number of text words before the relevant text information is encountered for an inference item) makes inference items more difficult.

l. Lexical overlap variables significantly influence item difficulty. As predicted, the following variables make items easier: ii8 and ii11 (number and percent of words, respectively, in correct answer that overlap with words in the key text sentence including lexically related words for inference items); ss8 and ss11 (number and percent of words, respectively, in correct answer that overlap with words in the key text sentence including lexically related words for supporting idea items). In addition, the variable mm5 (ordinal position of the earliest word on the first line that overlaps with a content word in the correct answer of a main idea item includes lexically related words) makes items harder, such that the later the ordinal position, the harder the item. (Note: Category l applies only to text/item overlap variables; it does not apply to the pure text or pure item variables).

In summary, for text variables there were 10 possible categories that could have influenced item difficulty (category j and l were not relevant); of these 10 categories, eight were significantly related to item difficulty (negatives and number of paragraphs were not significantly related).

For the *text/item* overlap variables there were eight possible categories that could have influenced item difficulty, and all but one (the frontings) received some support. For *items* there were three possible categories, (categories *a*, *b*, and *d*) and only one (negations) received some support.

Overall, the correlational results suggest that those variables found to influence comprehension in the experimental literature also appear to influence our multiple-choice data whether we examine just the text/item overlap scores (which deal primarily with single sentences) or the total text scores.

There are a few additional variables in Table 2 that proved to be significant predictors of item difficulty but were not specifically covered by Hypothesis 1. These are:

v14 (greater number of words in incorrect options makes items harder).
v23 (social science content makes items harder).
v24 (humanities content makes items easier).
v52 (the greater the number of semicolons in the passage the harder the item).
ss1 (if a word in the stem corresponds to a unique word in the passage, and the relevant supporting idea information is contained in the same passage sentence as the unique word, the item is easier).

As mentioned, none of these latter variables were included in our category list under Hypothesis 1 but are mentioned here for the sake of completeness.

Based on the variables that are significant, one might be tempted to conclude without further analyses that the correlational results--see Table 2 --appear to support the *construct validity* of the TOEFL reading section. This conclusion would seem to hold whether we examine just the text/item overlap variables or examine just the text variables. Therefore, these correlational results alone might at first appear to call into question some of Royer's criticisms of multiple-choice tests of reading as being primarily tests of reasoning rather than comprehension. However, this result should be considered highly tentative because some of the variables are significantly intercorrelated and furthermore are rather low in absolute magnitude (i.e., while 24 variables are listed as significant in the nested sample only two correlations exceed an absolute magnitude of .20). In light of this, before any firmer conclusion can be reached concerning construct validity, a regression technique is the appropriate method needed to examine this issue--see the results below.

We have just noted above that, while the pattern of significant correlations for the *nested* sample was interpreted as possible evidence favoring construct validity, the actual magnitudes of the significant correlations were in many cases rather small. The *non-nested* sample reveals somewhat larger correlations than the nested sample (i.e., 15 correlations are at or in excess of an absolute value of .20 but with only one in excess of .30). This is an improvement, but caution is still advisable in attempting to draw any firm conclusions at this point regarding Royer's (1990) criticism of multiple-choice tests of reading. The magnitude of the multiple-R (see the regression results below) is the more appropriate place to draw firmer conclusions regarding construct validity for the non-nested sample.

Regression Analyses

Criteria for Admitting Variables into the Stepwise Regressions. For all stepwise regressions, the following criteria were used for admitting variables into the regression. All variables listed in Table 1 were available for possible selection. Each new variable that was admitted into the solution had to yield a significant individual t value ($p < .05$), and, in addition, for the final solution, the new t values for all previously admitted variables had to be significant. If the next variable admitted showed a non-significant t, then the previous solution was considered the final one.

The Overall Predictability of Item Difficulty and Evaluation of Hypothesis 2

Stepwise regression analysis of 213 reading items. As we see from Table 3 there are eight significant predictors of the difficulty of reading items (equated deltas). The overall $F(8,204)= 12.49$, $p < .0001$; the multiple-R = .57 which accounts for 32.9 percent of the variance. The significant variables in the order they emerge from the regression analysis are:

ss11 (percent of words in correct option that match words in key text sentence for supporting idea items, including lexically related words),
ss9 (the number of words in the key text sentence for the supporting idea items),
ii3 (the information for inference items is in the middle of the passage),
mm5 (ordinal position of lexically matched word(s) on first text line for main idea correct option, including lexically related words\,
mm3 (main idea is in the general beginning of the passage),
v24 (subject matter consists of humanities),
v32 (rhetorical organization is problem/solution), and
v28 (author of passage presents an argumentative stance).

As pointed out earlier, while the absolute magnitude of the correlations in Table 2 were somewhat low, this in itself need not interfere with obtaining a fairly robust multiple-R.

Evaluation of Hypothesis 2 for the nested sample. We note that these significant predictors agree with the following four categories of Hypothesis 1:

rhetorical organization, (v28 and v32),
sentence length (ss9),
location of relevant information (ii3 and mm3), and
lexical overlap (ss11 and mm5).

We note that, in addition, subject matter (v24) also was significant even though this was not listed under Hypothesis 1.

Only four categories of the 12 provide independent information concerning item difficulty. To evaluate whether the presence of additional categories of the 12 might have been forthcoming had a non-nested data sample been used (where only one item per passage was used and where the three item types were approximately equally represented) we conducted the following additional analyses.

<u>Regression analysis for non-nested data sample (n=98): Possible implications for Hypothesis 2</u>. It is possible that because the data in the combined item sample represent what statisticians call a nesting effect (wherein several items are associated with the same passage), a more robust result concerning the categories in regard to Hypothesis 2 might be possible, if we were to construct and analyze a non-nested data sample (a sample in which one item would be associated with one passage). To examine this, we constructed a special non-nested sample (n=98) from the larger sample (n=213) consisting of one item per passage; approximately one-third of this sample consisted of main idea items (n=32), one-third of inference items (n=33), and one-third of supporting idea items (n=33).

As we can see from Table 3 (page 38), the *F* for the multiple regression for this non-nested sample of 98 items (and 98 passages) was as follows: *F*(11,86) = 10.70, *p* < .0001. The multiple-R equals .76 accounting for 57.8 percent of the variance of the equated deltas of reading items. While the regression analysis of the nested sample of 213 items accounted for 32.9 percent of the variance, here the non-nested sample of 98 items accounts for 57.8 percent of the variance. This suggests that the issue of nesting may be importantly altering our ability to account for item difficulty.

For the non-nested sample, the following significant variables, in the order they emerged from the regression analysis, were found to predict equated delta:

ss11 (supporting idea--percent of words in correct answer that overlap with words in key text including lexically related words),
ss9 (number of words in key text sentence containing relevant supporting idea information),
ii8 (inferences--number of words in correct answer that overlap with words in key text including lexically similar words),
v23 (subject matter is social science),
v56 (referentials across clauses),
v44 (average sentence length of first paragraph),
v29 (rhetorical organizer is list),
v11 (number of negations in correct answer of item),
ii5 (length of paragraph in which relevant information is located for inference items),
ss4 (supporting ideas--number of words that have to be read before relevant information begins for the supporting idea items),
mm5 (ordinal position of lexically matched words on first text line including lexically related words for the main idea correct option).

Regarding the 12 categories of Hypothesis 2, these 11 significant predictors for the non-nested sample include the following seven categories that provide independent predictive information:

K: lexical overlap--ss11, ii8, mm5;
F: sentence length--ss9, v44;
G: paragraph length--ii5;
C: rhetorical organizer--v29;
A: negations--v11;
B: referentials--v56;
L: passage length--ss4.

Thus seven categories out of 12 provide independent variance. In addition we note that a subject matter variable (v23) which was not included in Hypothesis 1 as a category was a significant predictor.

These particular results do suggest that the issue of nesting might significantly alter the degree to which Hypothesis 2 appears to be confirmed; the nested sample yielded four categories out of 12 in support of Hypothesis 2, while the non-nested sample yielded seven categories out of 12.

Cross-validation with another non-nested sample of TOEFL items. It was possible to construct another non-nested sample of items (n=72) from the larger sample of items (n=213) by substituting a different reading item for each passage than was selected for the n=98 item sample--however this could not be done for some passages which had only one original item associated with them. This, plus the constraint that the three reading item types be equally represented, resulted in a sample of 72 new items, one item per passage (24 main idea items, 24 inference items and 24 supporting idea items). This new non-nested sample can provide us with useful information. We can use the predictors from the n=98 sample and see if they can predict this new n=72 sample.

Using the 11 predictors from the n=98 sample, we find that the multiple-R = .55; this accounts for 29.8 percent of the variance of the smaller (n=72) sample and yields an $\underline{F}(11,60) = 2.31$, $\underline{p} = .02$.

We conclude that there is some evidence that we can get significant cross-validation across the two non-nested samples.

Separate Regressions for the Five Ability Groups

Stepwise regressions based on 213 items (nested sample). Of the 19 significant predictor variables listed in Table 4 (page 40), 11 are significant for at least two of the ability groups. We note that for the lower ability examinees, more variance is accounted for than for the higher ability examinees. We also note that, over the span of the five ability levels, on average, less variance has been accounted for than was true for the analysis which used all five ability groups together (with equated delta as the criterion--see results in Table 4 for n=213). That is, the variance accounted for varies from 14.3 percent to 39.1 percent. On average, these results are somewhat lower than the 32.9 percent achieved for all ability groups combined. It is possible that this overall reduction in variance accounted for may be due to the restricted range of scores associated with each ability level.

Stepwise regressions based on 98 items (non-nested sample). Of the eighteen significant predictor variables listed in Table 5 (page 42), seven are significant for at least two of the ability groups. Again, we see that in the sample of lower ability examinees there is more variance accounted for than in the sample of higher ability examinees. Once again, there is some suggestion overall that the variance accounted for by the five ability groups is somewhat lower than that reported for all five ability groups combined (that is where equated delta was the criterion--see Table 5, n=98). In particular the variance ranges from 32.1 percent to 60.7 percent for the five ability groups, whereas for the combined group (with equated delta) the variance was 57.8 percent. Again, it is possible that the restriction in score range associated with each ability group may be responsible for this overall attenuation of predictability.

Implication of Stepwise Regression Results for Construct Validity

Numerous stepwise regression analyses have been presented and found to provide some support for Hypotheses 1 and 2. There are additional implications of these results. In particular, we have interpreted Royer's (1990) critique of multiple-choice reading tests as implying that text and text/item overlap variables should play a minor role in predicting reading item difficulty, with his added implication that item variables should play a major role. By and large, all of our individual stepwise regressions suggest just the opposite conclusion: item variables play a very minor role while text and text-associated variables play by far the major role in accounting for reading item difficulty. Hence it appears that a strict reading of Royer's construct validity argument does not receive support.

Additional Data Analyses

Related data--involving the stepwise regression analysis of each of three reading item types along with their associated significant zero-order correlations--are presented in the Appendix.

Conclusion

In this study we have been interested primarily in determining how well the difficulty of reading items can be accounted for by a set of predictors that reflect the contribution of text structure, item structure, and the joint effect of both the text and item structure. We found that a substantial amount of the variance can be accounted for by a relatively small set of predictors; the variance accounted for ranged from 33 percent up to 61 percent for predicting equated delta, depending upon the particular analysis undertaken.

In predicting the performance of each of five ability groups we found that for the nested sample of 213 items, the percent variance of reading item difficulty accounted for ranged from 39 percent for the lowest ability group to 14 percent for the highest ability group; these figures were generally higher when the non-nested (n=98) sample was examined--there the variance accounted for was 61 percent for the lowest group and 36 percent for the highest ability group. These particular results indicate that our variables appear to be more sensitive to the performance of lower than higher ability examinees. Such a result agrees with that reported by Freedle and Fellbaum (1987) for TOEFL listening comprehension tests; Freedle and Fellbaum were better able to account for lower than higher ability examinees in predicting listening item performance.

Within this broader concern we have also focused upon a small set of hypotheses so as to come to terms more clearly with a number of claims that have been made in the literature concerning reading comprehension and the adequacy of reading comprehension tests per se. In particular, Royer (1990) and Katz et al. (1990) have questioned whether multiple-choice reading tests can be considered appropriate tests of passage comprehension in light of the fact that item content alone (in the absence of the reading passage) can be demonstrated to lead to correct answers well above chance levels of guessing. In addition, Goodman (1982) has pointed out that many of the experimental studies of comprehension have focused on just one or two variables at a time; he questions whether these separate studies, taken together, necessarily build up our understanding of how full comprehension of text takes place.

In response to these several concerns, we framed the prediction of the difficulty of reading items around two hypotheses meant to put into clearer perspective the viability of multiple-choice reading comprehension tests, here exemplified by the TOEFL reading passages and their associated items. Since many of the scored variables deal with text content similar to those of concern in especially the experimental literature, we reasoned that the successful prediction of the difficulty of reading items would allow us to draw several important conclusions.

The first hypothesis asserts that multiple-choice items will be sensitive to a similar set of variables as have been found to be important in studying comprehension processes in the experimental literature. The correlational evidence generally supported most of the categories detailed under Hypothesis 1 for the text and text-related (i.e., text/item overlap) variables. We interpreted that to mean that multiple-choice response formats yield similar results to those found in the more controlled experimental studies. We further interpret the results of our various regression analyses as broadly supporting the assertion that multiple-choice reading tests can be demonstrated to have construct validity. Furthermore, many of our current results agree with similar analyses carried out for SAT and GRE reading comprehension tests (see Freedle & Kostin, 1991; 1992). Hence we feel Royer's (1990) statement that multiple-choice tests do not measure passage comprehension can be called into question.

A second hypothesis asserts that many of the significant variables will be found to influence reading item difficulty jointly. Our stepwise regression analyses indicate that there is often considerable evidence that many of the different categories of variables studied in Hypothesis 2 do in fact jointly account for reading item difficulty. This result was further interpreted as a positive response to Goodman's (1982) inquiry as to whether the joint operation of many of the variables that had been studied in restricted experimental settings would necessarily increase our understanding of the factors influencing reading comprehension difficulty. In several cases our results appear to suggest that many of the different categories of variables do provide independent predictive information; hence, the few variables studied across disparate studies in fact jointly combine to increase our understanding of what influences comprehension difficulty. A related set of analyses using a large number of SAT and GRE reading items (Freedle & Kostin, 1991; 1992) further confirms the viability of this demonstration.

In short, we find considerable evidence that multiple-choice tests of reading comprehension yield results that are quite consistent with those obtained from controlled experimental studies dealing with language comprehension. More importantly, because of the relatively large size of our data base, the results also provide evidence that many variables affecting comprehension can be shown to contribute independent predictive information in determining reading item difficulty. In conclusion, we have found that a significant amount of the item difficulty variance can be accounted for by a relatively small number of variables for the three reading item types studied (main ideas, inferences, and supporting idea items).

Future work. We have commented on our better ability to predict the data of the low- as compared with the high-ability examinees. We would like to deepen our understanding of what particular strategies the high-ability examinees in particular are using. One technique that might prove useful in this regard is that of thinking aloud (e.g., see Freedle, Kostin, & Schwartz, 1987). By having examinees explain to us what they are thinking when they read the passage and respond to the options in each reading item, it should be possible, for example, to discover what special strategies the high ability examinees are using, and to develop new scoring categories that better reflect those strategies. This should improve our ability to predict the item difficulty values of high ability examinees.

APPENDIX: Correlations and Regressions for Each of Three TOEFL Reading Item Types

TABLE A Correlations of Significant Item, Text, and Text/Item Variables with Equated Delta for Three TOEFL Reading Item Types

		Significant Correlation of Delta with Three Reading Item Types		
		(n=59 items)	(n=61)	(n=93)
Variable		Main Idea	Inference	Supporting
Variables Apply to All 3 Item Types				
	Item Variables			
v8	Stem: sum of referentials		.25++[a]	
v11	Correct: negations		.27++	
v13	Correct: referentials		.20+	
	Text Variables			
v18	Vocabulary		.23++	.25***
v19	Concreteness			-.18++
v20	Physical science	-.33***		
v24	Humanities			-.21**
v26	Natural science excerpt	-.31**		
v29	Grimes:list			-.21**
v34	No. paragraphs	.30**		
v35	No. words	.21+	.27**	
v36	No. sentences	.30**		
v38	No. words in longest paragraph		.37***	
v40	No. sentences in longest paragraph		.24++	
v44	Aver. wds/sentence/first paragraph		.34***	.17++
v46	Percent fronted, total text		.32***	
v47	Freq. fronted, total text	.28**	.25++	
v50	No. longest run fronted clauses	.24++	.32***	
v56	Freq. across clause refer.	.50***		
v58	Sum all referentials	.40***		
	Text/Item Overlap Variables That Apply to Every Item Type			
v60	Overlaps: no. fronts		.26**	
v61	Overlaps: within clause	.26**	.20+	
v63	Overlaps: no. refer. outside		-.31***	
v64	Overlaps: no. negations	.32**		
	Text/Item Overlap Variables That Apply to Some Item Types			
mm2	Main idea info. middle of text	.41***	.25++	.20++
mm3	Main idea info. in first, second and/or first short paragraph	-.36***		
mm5	First line lexical match plus related words	.47***	NA	NA
mm6	Sum of first line incorrect overlap scores	.25++		

Table A (Continued)

ii3	Info. in middle of text (Inferences & explicits only)	NA	.30**	
ii5	Length of info. paragraph (Inferences & explicits only)	NA	.28**	
ii13	Info. in last short paragraph (Inferences & explicits only)	NA	-.34***	
ss8	Number lexically matched words plus related words	NA		-.29***
ss9	No. wds in key sentence	NA		.24**
ss11	Percent lexically matched words plus lexically related words	NA		-.51***

[a] A positive correlation for delta means the variable makes the items harder. *** = signif. at p < .01, 2-tailed; ** = signif. at p <.05, 2-tailed; * = p <.06, 2-tailed; ++ = p <.05, 1-tailed; + = p < .06, 1-tailed. NA = not applicable. If a variable was not significant for the 2-tailed test but appeared as one of the variables listed under Hypothesis 1 where direction was predicted, we applied a 1-tailed test. Also if a variable was not significant at the 2-tailed test and it was significant for our earlier SAT and/or GRE data (Freedle & Kostin, 1991; 1992), we again applied a 1-tailed test.

Regression Results for Each of Three Reading Item Types

We mentioned in the body of this report that for the nested (n=213) and non-nested (n=98) samples the correlations tended to be rather small in magnitude. It should be noted here that each of three item type samples is a non-nested sample. Interestingly enough, Table A shows that 33 variables yield an absolute value of .20 or more for one or more item types (18 of these are in excess of .30). This is a substantial increase in the number of correlations of fairly robust size. While the sample sizes here are too small to justify focusing our main analyses on each reading item type, such data do suggest that in some cases evidence favoring construct validity can be deduced even at the correlational level.

Given these larger correlations, one might think that multiple correlations based on these correlations might well prove to be substantially larger than those reported above for the nested and non-nested samples. This is not necessarily the case as we will now demonstrate.

For main idea items (n=59) the multiple-R was .782 accounting for 61.1 percent of the variance. The overall F value was $F(7,51) = 11.43$, $p < .0001$.

For inference items (n=61) the multiple-R was .618 accounting for 38.2 percent of the variance. The overall F value was $\underline{F}(4,56) = 8.64$, $\underline{p} < .0001$.

For supporting idea items (n=93) the multiple-R was .642 accounting for 41.2 percent of the variance. The overall F value was $\underline{F}(5,87) = 12.2$, $\underline{p} < .0001$.

The percentage variance accounted for is therefore roughly equivalent to what was found in the main report (i.e., for the non-nested sample [n=98] it was 57.8 percent while for the nested [n=213] it was 32.9 percent). Hence the absolute magnitude of the zero-order correlations is not always an indicator that the multiple-R based on them will necessarily be substantially larger. This, in turn, implies that the small magnitudes of the correlations reported above (in Table 2) do not in themselves carry any negative implications concerning our ability to account for a substantial amount of the variance in predicting item difficulty in a multiple regression analysis--it is the fact that many of the correlations are significant that is important and the fact that they provide additive information concerning the criterion of item difficulty.

TABLE 1 **List of Variables Used in Analyses**

Independent Variables

Item Variables

Item type
v1 --Main idea
v2 --Inference
v3 --Supporting idea

Variables for item's stem
v4 --Words in stem
v5 --Fragment stem
v6 --Negative stem
v7 --Fronted stem
v8 --Reference stem

Variables for item's correct option
v9 --Answer position
v11--Negative correct
v13--Reference correct

Variables for item's incorrect options
v14--Words in incorrects
v15--Negative incorrects

Text Variables

Vocabulary variable for text
v18--Vocabulary

Concreteness/abstractness of text
v19--Concreteness

Subject matter variables of text
v20--Physical science
v21--Life science
v23--Social science
v24--Humanities
v25--Arts
v26--Natural science excerpt
v27--About natural science

Type of rhetorical organization
v28--Argument
v29--List/describe
v30--Cause
v31--Compare
v32--Problem/solution

Table 1 (continued)

Coherence of lexical concepts over whole text
v33--Coherence

Lengths of various text segments
v34--Paragraphs
v35--Text words
v36--Text sentences
v38--Longest paragraph words
v40--Longest paragraph sentences
v44--First paragraph sentence length

Occurrence of different text "frontings"
v46--Percent fronted text clauses
v47--Frequency fronted text clauses
v48--Frequency combinations of fronted text structures
v49--Frequency of text clefts
v50--Longest fronted run

Text special punctuations
v52--Semicolons
v53--Colons
v54--Dashes

Text referentials
v55--Reference within text clauses
v56--Reference across text clauses
v57--Frequency special reference
v58--Reference sums

Text negations
v59--Text negatives

Text/Item Overlap Variables

Special text/item overlap variables that apply to all item types
v60--Number of fronts in key overlapping sentence
v61--Number of referentials (within clauses) in key overlapping sentence
v62--Number of referentials (across clauses) in key overlapping sentence
v63--Number of referentials (outside clauses) in key overlapping sentence
v64--Number of negations used in key overlapping sentence
v65--Number of independent clauses in key sentence or sentences

Text/item overlap variables applicable to main idea information
mm2--Main idea middle text
mm3--All early main idea locations

Table 1 (continued)

mm5 --First line lexical match

mm6 --First line lexical match for incorrects

Text/item overlap variables applicable to inferences

ii1 --Unique word same sentence

ii2 --Information in last sentence

ii3 --Information middle of text

ii4 --Words before critical information

ii5 --Words in relevant paragraph

ii6 --Information middle relevant paragraph

ii8 --Number lexically matched words

ii9 --Number words in key text sentence

ii11--Percent lexically matched words

Text/item variables applicable to supporting idea items

ss1 --Unique word, same sentence

ss2 --Key word occurs in multiple places

ss3 --Information is in middle of text

ss4 --Words before critical information

ss5 --Words in relevant paragraph

ss6 --Information middle relevant paragraph

ss8 --Number lexically matched words

ss9 --Number of words in key text sentence

ss11--Percent lexically matched words

TABLE 2 Correlations of Significant Variables with Item Difficulty (Equated Delta)

Significant Variables		Nested Sample (n=213)	Non-nested Sample (n=98)
Item Variables			
v11	Negative in correct options	.13++	.16
v14	Words in incorrect options	.14**	.23**
v15	Negative in incorrect options	.11++	.12
Text Variables			
v18	Vocabulary	.17***	.20**
v19	Concreteness	-.11++	-.13
v23	Social science	.14**	.24**
v24	Humanities	-.15**	-.17++
v29	List/describe	-.17***	-.23**
v32	Problem/solution	.12	.20**
v35	Text words	.13++	.19++
v38	Longest paragraph words	.11++	.15
v44	First paragraph sentence length	.14**	.21**
v46	Percent fronted text clauses	.15**	.15
v47	Frequency fronted text clauses	.12++	.16
v52	Semicolons	.06	.22**
v56	Reference across text clauses	.10	.27***
v58	Reference sums	.06	.20**
Text/Item Overlap Variables			
v61	Overlap reference, within clauses	.09	.20**
v64	Overlap negatives	.11++	-.04
mm2	Main idea middle text	.17***	.21**
mm5	First line lexical match plus related words	.16**	.26***
ii3	Information middle of text	.18***	.10
ii4	Words before critical information	.13++	.05
ii5	Words in relevant paragraph	.17***	.10
ii6	Information middle relevant paragraph	.12++	.01
ii8	Number lexically matched words plus related words	.01	-.21**
ii9	Number words in key text sentence	.12++	-.01
ii11	Percent lexically matched words plus related words	-.02	-.19++
ss1	Unique word same sentence	-.13++	-.23**
ss8	Number lexically matched words plus related words	-.24***	-.19++
ss11	Percent lexically matched words plus related words	-.35***	-.37***

Table 2 (continued)

*** means significant at p < .01, 2-tailed
** means significant at p < .05, 2-tailed
++ means significant at p < .05, 1-tailed

If a variable was not significant for the 2-tailed test but direction was predicted we applied a 1-tailed test.

A positive correlation indicates that the presence (or more) of the variable makes the item harder while a negative correlation means that the presence (or more) of the variable makes the item easier.

TABLE 3 **Multiple Regression Results for Two TOEFL Reading Item Samples**[a,b]

	(n=213)		(n=98)	
Variable	Beta Weight	p Value	Beta Weight	p Value
ss11	-.49	.0001	-.51	.0001
ss4			.26	.0031
ss9	.37	.0001	.25	.0386
ii3	.21	.0018		
ii5			.33	.0007
ii8			-.18	.0244
mm5	.30	.0001	.28	.0042
mm3	-.21	.0031		
v23			.18	.0137
v24	-.20	.0012		
v29			-.17	.0242
v32	.18	.0033		
v56			.25	.0012
v11			.15	.0448
v28	.13	.0306		
v44			.21	.0085

[a] ss11 Percent of words in the correct answer that overlap with content words in the key text sentence, including lexically related words, for supporting idea items.

ss4 Number of text words before the relevant information begins for supporting idea items.

ss9 The number of words in the key text sentence for supporting idea items.

ii3 The relevant information is in the middle of the passage for inference items.

ii5 The length of the paragraph containing relevant inference information.

ii8 The number of words in the correct answer that overlap with content words in the key text sentence; includes lexically related words; for inference items.

mm5 The ordinal position of the earliest word in the first text line that overlaps with the content word in the correct answer for main idea items.

mm3 Main idea topic is in the general beginning of the passage.

v23 Represents social science content.

v24 Represents humanity content.

v29 Passage has a rhetorical organizer using list structure.

v32 Passage has a rhetorical organizer using a problem/solution structure.

Table 3 (continued)

v56 Frequency of across-clause referentials.
v11 Negations in correct answer.
v28 Passage has an argumentative structure.
v44 Average sentence length of first paragraph.

[b] n = 213 (nested sample)
F(8,204) = 12.49, *p* <.0001, multiple-R = .57, R squared = .329.

n=98 (non-nested sample)
F(11,86) = 10.70, *p* <.0001, multiple-R = .76, and R squared = .578.

TABLE 4 **Multiple Regression Analyses for Each Ability Group (for _z_-scores) Using the Nested Sample of 213 Items**

	Ability Level									
	Lowest[a,c]		2nd Low		Middle		2nd High		Highest	
Variable[b]	Beta	p	Beta	p	Beta	p	Beta	p	Beta	p
ii3	-.17	.01	-.18	.01	-.17	.02				
ss9	-.26	.00	-.36	.00	-.33	.00	-.26	.00		
ss11	.51	.00	.49	.00	.47	.00	.45	.00	.28	.00
mm3	.25	.00	.25	.00	.23	.00				
mm5	-.21	.00	-.27	.00	-.27	.00				
ii11	.17	.01								
v35	-.14	.02								
v4	-.13	.04								
v18	-.16	.01								
v38					-.10	.05	-.16	.01		
v24	.12	.04	.18	.00	.16	.01				
v28	-.14	.01	-.13	.03	-.13	.03				
v29	.14	.02					.14	.03		
v32			-.17	.01	-.16	.01				
v46	-.16	.01								
v15									-.14	.03
v61									-.17	.01
v44	-.12	.04							-.17	.01
v56	-.15	.01								

[a] The reader should note that the ability group _z_-scores differ in sign from equated delta so that here a positive beta weight is associated with variables making items easier (whereas for equated delta they are associated with making items harder).

[b] ii3 Inference information is located in middle of text passage.

ss9 Number of words in the key text sentence for supporting ideas.

ss11 Percent of words in correct answer that overlap with key text sentence, including lexically related words, for the supporting idea items.

mm3 Main idea information is located in general beginning of the passage.

mm5 Relating to main idea, it is the ordinal position of the earliest word on the first text line that overlaps with the correct answer content word.

ii11 Percent of words in correct answer that overlap with key text sentence, including lexically related words, for the inference items.

Table 4 (continued)

v35 Number of words in the passage.
v4 Number of words in the item stem.
v18 Text vocabulary (number of words with three or more syllables).
v38 Number of words in the longest paragraph.
v24 Represents humanities content.
v28 Represents an argumentative passage.
v29 Represents a list rhetorical organization of passage.
v32 Represents a problem/solution rhetorical organization of passage.
v46 Percent fronted clauses of the passage.
v15 Negations in the incorrect options.
v61 Number of referentials within text clauses from the text section that overlaps with the relevant material for getting the correct option.
v44 Average sentence length of the first paragraph.
v56 Across-clause referentials.

[c] The multiple-R for the low ability is .63 accounting for 39.1 percent of the variance. $F(15,197) = 10.72$, $p < .0001$.

The multiple-R for the second lowest group is .56 accounting for 31.4 percent of the variance. $F(8,204) = 11.69$, $p < .0001$.

The multiple-R for the middle group is .54 accounting for 29.7 percent of the variance. $F(9,203) = 9.52$, $p < .0001$.

The multiple-R for the second highest group is .41 accounting for 17.1 percent of the variance. $F(4,208) = 10.74$, $p < .0001$.

The multiple-R for the highest group is .38 accounting for 14.3 percent of the variance. $F(4,208) = 8.65$, $p < .0001$.

TABLE 5 **Multiple Regressions for Five Ability Groups Using the Non-Nested Sample of 98 Items**

	Ability Level									
	Lowest[b]		2nd Low		Middle		2nd High		Highest	
Variable[a]	Beta	p	Beta	p	Beta	p	Beta	p	Beta	p
mm2	-.15	.05								
mm3			.25	.00						
mm5									-.17	.06
ii11	.25	.00								
ii5	-.39	.00								
ii8	.19	.01	.21	.01	.21	.02	.21	.01		
ss4	-.22	.02	-.18	.04			-.20	.02		
ss9					-.18	.05				
ss11	.56	.00	.58	.00	.51	.00	.48	.00	.30	.00
ss6	-.24	.02								
v23	-.19	.01	-.25	.00	-.21	.02	-.25	.00		
v29	.21	.00	.19	.02					.27	.00
v56	-.24	.00	-.22	.01	-.22	.01	-.24	.00		
v58									-.21	.02
v15									-.21	.02
v61							-.19	.02		
v63	.15	.05								
v44	-.31	.00	-.28	.00	-.20	.04	-.29	.00	-.34	.00

[a] mm2 Main idea information is in middle of passage.

mm3 Main idea information is in the general beginning of the passage.

mm5 Relating to main idea, it is the ordinal position of the earliest word on the first text line that overlaps with a correct answer content word.

ii11 Percent of words in the correct answer that overlap with words in the key text sentence, including lexically related words; applies to inference items.

ii5 Length of the informative paragraph containing relevant correct answer inference item information.

ii8 Number of words in the correct answer that overlap with words in the key text sentence including lexically similar words; applies to inference items.

ss6 Relevant information is in middle of paragraph, rather than first or second sentence of paragraph.

ss4 Number of text words before the relevant information begins.

ss9 Number of words in the key text sentence relating to the correct answer of the supporting idea item.

ss11 Percent of words in the correct answer that overlap with words in the key text sentence, including lexically related words; applies to supporting idea items.

Table 5 (continued)

v23 Social science content of passage.
v29 List as a rhetorical organization of the passage.
v56 Frequency of across-clause referentials in the text.
v58 Sum of all text referentials.
v15 Negations in the incorrect options.
v61 Referentials within clauses of the portion of the passage that is relevant to getting the answer correct.
v63 Number of outsider referentials in the relevant overlapping text.
v44 Average sentence length of the first paragraph.

[b] Multiple-R for the lowest ability is .78 accounting for 60.7 percent of the variance. *F*(12,85) = 13.42, *p* < .0001.

Multiple-R for the 2nd lowest ability is .70 accounting for 48.6 percent of the variance. *F*(8,89) = 10.51, *p* < .0001.

Multiple-R for the middle group is .59 accounting for 35.3 percent of the variance. *F*(6,91) = 8.26, *p* < .0001.

Multiple-R for the 2nd highest group is .67 accounting for 44.3 percent of the variance. *F*(7,90) = 10.22, *p* < .0001.

Multiple-R for the highest ability group is .60 accounting for 35.7 percent of the variance. *F*(6,91) = 8.42, *p* < .0001.

References

Abelson, R. P., & Black, J. B. (1986). Introduction. In J. A. Galambos, R. P. Abelson, & J. B. Black (Eds.), Knowledge structures. Hillsdale, NJ: Erlbaum.

Abrahamsen, E., & Shelton, K. (1989). Reading comprehension in adolescence with learning disabilities: semantic and syntactic effects. Journal of Learning Disabilities, 22, 569-572.

Alderman, D., & Holland, P. (1981). Item performance across native language groups on the Test of English as a Foreign Language (ETS Research Report RR-81-16). Princeton, NJ: Educational Testing Service.

Bever, T. G., & Townsend, D. (1979). Perceptual mechanisms and formal properties of main and subordinate clauses. In W. Cooper & E. Walker (Eds.), Sentence processing. Hillsdale, NJ: Erlbaum.

Bhasir, J. I. (1990). The demands of main idea tasks in reading comprehension tests and the strategic responses of bilingual poor comprehenders. Unpublished dissertation, Columbia University Teachers College, New York.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82, 45-73.

Clark, H. H., & Haviland, S. (1977). Comprehension and the given-new contract. In R. Freedle (Ed.), Discourse production and comprehension. Norwood, NJ: Ablex.

Cohen, J., & Cohen, P. (1983). Applied multiple regression: Correlational analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-514.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Ervin-Tripp, E. (1964). An analysis of the interaction of language, topic and listener. American Anthropologist, 66, 86-102.

Freedle, R., & Duran, R. (1979). Sociolinguistic approaches to dialogue with suggested applications to cognitive science. In R. Freedle (Ed.), New directions in discourse processing. Norwood, N.J.: Ablex.

Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Eds.), Cognitive and linguistic analyses of test performance. Norwood, NJ: Ablex.

Freedle, R., Fine, J., & Fellbaum, C. (1981). Predictors of good and bad essays. Paper presented at the annual Georgetown University Roundtable on languages and linguistics, Washington, D.C.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages (ETS Research Report RR-91-29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1992). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences, and explicit statements (ETS Research Report RR-91-59). Princeton, NJ: Educational Testing Service.

Freedle, R., Kostin, I., & Schwartz, L. (1987). A comparison of strategies used by Black and White students in solving SAT verbal analogies using a thinking aloud method and a matched percentage-correct design (ETS Research Report RR-87-48). Princeton, NJ: Educational Testing Service.

Goodman, K. (1982). Language and literacy: The selected writings of Kenneth S. Goodman (Vols 1 & 2). F. Gollasch (Ed.). Boston: Routledge & Kegan Paul.

Graves, M. (1986). Vocabulary learning and instruction. In E. Rothkopf (Ed.), Review of research in education, Vol. 13. Washington, D. C.: American Educational Research Association.

Grimes, J. (1975). The thread of discourse. The Hague: Mouton.

Gunning, R. (1964). How to take the fog out of writing. Chicago: Dartnell.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. Journal of Educational Psychology, 79, 220-227.

Hare, V., Rabinowitz, M., & Schieble, K. (1989). Text effects on main idea comprehension. Reading Research Quarterly, 24, 72-88.

Hites, R. W. (1950). The relation of readability and format to retention in communication. Unpublished doctoral dissertation, Ohio State University, Ohio.

Hymes, D. (1962). The ethnography of speaking. In T. Gladwin & W. Sturtevant (Eds.), Anthropology and human behavior. Washington, D.C.: Anthropological Society of Washington.

Johnston, P. (1983). Reading comprehension assessment: A cognitive basis. Newark, NJ: The International Reading Association.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Boston: Allyn and Bacon.

Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. Psychological Science, 1, 122-127.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. In B. Britton & J. Black (Eds.), Understanding expository text. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Klare, G. (1974-1975). Assessing readability. Reading Research Quarterly, 10, 62-102.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111-151.

Meyer, B. (1975). The organization of prose and its effects on memory. Amsterdam: North-Holland.

Meyer, B. (1985). Prose analysis: Purposes, procedures, and problems. In B. Britton & J. Black (Eds.), Understanding expository text. Hillsdale, NJ: Erlbaum.

Meyer, B., & Freedle, R. (1984). The effects of different discourse types on recall. American Educational Research Journal, 21, 121-143.

Newsome, R. S., & Gaite, J. H. (1971). Prose learning: Effects of pretesting and reduction of passage length. Psychology Reports, 28, 128-129.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., & Bent, D. (1975). Statistical packages for the social sciences. 2nd edition. NY: McGraw-Hill.

Paivio, A. (1971). Imagery and verbal processes. NY: Holt, Rinehart & Winston.

Royer, J. (1990). The sentence verification technique: A new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), Cognitive assessment of language and math outcomes. Norwood, NJ: Ablex.

Stark, H. (1988). What do paragraph markings do? Discourse Processes, 11, 275-303.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning. Hillsdale, NJ: Erlbaum.

Tuinman, J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

Cover Printed on Recycled Paper

57906-07588 • Y43M.5 • 275589 • Printed in U.S.A.