ED 400 328 TM 025 733

AUTHOR Nissan, Susan; And Others
TITLE An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension. TOEFL Research Reports, 51.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-37
PUB DATE Feb 96
NOTE 52p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Classification; *Dialogs (Language); *Difficulty Level; English (Second Language); Item Banks; Language Tests; *Listening Comprehension Tests; Test Construction; Test Format; Test Items; *Vocabulary
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

One of the item types in the Listening Comprehension section of the Test of English as a Foreign Language (TOEFL) test is the dialogue. Because the dialogue item pool needs to have an appropriate balance of items at a range of difficulty levels, test developers have examined items at various difficulty levels in an attempt to identify their features. In this study, a classification system was created for certain item features, a sample of the current dialogue item pool was classified, and data analyses were conducted in an attempt to characterize the features of easy and difficult dialogue items. The results of the analyses indicate that, of the features studied, five were significant: (1) the presence of infrequent oral vocabulary; (2) the sentence pattern of the utterances in the stimulus; (3) the presence of negatives in the stimulus; (4) the necessity of making an inference to answer the item; and (5) the roles of the speakers in the stimulus. An appendix discusses the dialogue item and presents examples, and presents the Item Classification form. (Contains 3 figures, 14 tables, and 46 references.) (Author/SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 51
FEBRUARY 1996

An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension

Susan Nissan

Felicia DeVincenzi

K. Linda Tang

Educational Testing Service

An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension

Susan Nissan
Felicia DeVincenzi
K. Linda Tang

Educational Testing Service
Princeton, New Jersey

RR-95-37

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

Abstract

One of the item types in the Listening Comprehension Section of the TOEFL® test is the Dialogue. Because the Dialogue item pool needs to have an appropriate balance of items at a range of difficulty levels, test developers have examined items at various difficulty levels in an attempt to identify their features. In this study, the authors created a classification system for certain item features, classified a sample of the current Dialogue item pool, and conducted data analyses in an attempt to characterize the features of easy and difficult Dialogue items. The results of the analyses indicate that, of the features studied, five were significant: the presence of infrequent oral vocabulary, the sentence pattern of the utterances in the stimulus, the presence of negatives in the stimulus, the necessity of making an inference to answer the item, and the roles of the speakers in the stimulus.

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. The Council was formed through the cooperative effort of more than 30 public and private organizations 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. 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, 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. 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. Many projects require the cooperation of other institutions, however, 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 (1995-96) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
Carol Chapelle	Iowa State University
Fred Davidson	University of Illinois at Urbana-Champaign
Thom Hudson	University of Hawaii
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Acknowledgments

The authors thank the following people for their substantial contribution to the success of this project:

- Natalie Nordby Chen, Sara Jensen, Pamela Mollaun, Dan Tumposky, and Jean Yepes for participating in the rating of the items

- Natalie Nordby Chen, Susan Chyn, Dan Eignor, Roy Freedle, April Ginther, Rick Morgan, and Tina Wright for reviewing an earlier draft of this paper

- Kikumi Tatsuoka for her suggestions about the statistical analyses

- Judy DeChamplain for preparing the figures in this report

- Cindy McMahon, Shari DeFrank, and Maryanne Nieciecki for their secretarial assistance

Table of Contents

List of Tables

List of Figures

Introduction

Purpose

The purpose of this study was to determine which features of TOEFL Dialogue items contribute to item difficulty. Stimulus-related and item-related features were examined to determine whether these features, singly or in different combinations, have a significant impact on item difficulty.

Background

Numerous researchers have examined factors affecting the difficulty of various item types (e.g., Brown, 1988; Abraham & Chapelle, 1992; Freedle & Kostin, 1993). The length of stimulus and the amount of information the examinee needs to process to determine the correct answer are among the variables that contributed to item difficulty in cloze items (Brown, 1988; Abraham & Chapelle, 1993). Freedle and Kostin (1993) found that the presence of negation contributed significantly to the difficulty of reading comprehension items. However, these variables and others have not been examined in the context of multiple-choice listening items.

The Dialogue item is based on an aural stimulus. (See Appendix 1 for a description of the item type, its terminology, and some examples). In their papers from the TOEFL Invitational Conference on Communicative Competence, Bachman, Douglas, and Savignon made reference to the characteristics of the Dialogue item type with respect to the testing of communicative competence (Stansfield, 1986). In a 1985 TOEFL Research Report, Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro included a content analysis of the Dialogue. In two separate 1990 reports on TOEFL subtest functioning, Henning and Way studied the statistical characteristics of Dialogues. Henning's 1991 study examined the effect of response length, but not within the Dialogue item format. How Dialogue design characteristics contribute to item difficulty in the current TOEFL has not yet been formally researched.

Some preliminary work on this question has been done, however. According to a February 1991 TOEFL item-pool inventory, Dialogues with Deltas[1] above 11.0, the approximate mean Delta of the test, were in short supply. If the pool were not replenished, specifications of future tests would not be met. In order to produce more challenging Dialogues, TOEFL test developers inventoried the difficult Dialogue items (those having Deltas of 12.0 and above[2]) and identified characteristics that appeared to

[1] The Delta value is an item difficulty index computed using the proportion of examinees answering an item correctly.

[2] An item with a Delta value of 12.0 is one that was answered correctly by 60% of the population.

contribute to their difficulty. For example, in several Dialogues with higher Deltas, the script called for the speakers in the stimulus to make implications, use certain idiomatic expressions, or respond to each other obliquely. Since February 1991, test developers have been incorporating such characteristics into their Dialogue items, and pretest data have indicated that a relationship may indeed exist between certain design features and difficulty. Therefore, the authors attempted to confirm these preliminary indications about the relationship between design features and item difficulty with a more formal study.

Method

In this study the impact of 17 independent variables on item difficulty was explored. Item difficulty was defined in terms of Delta value, the dependent variable in the study.

Independent Variable Selection

To further the 1991 work on the TOEFL item pool, features that appeared to be common in Dialogues with higher Deltas were included as variables. Other variables were chosen to shed light on the utility of particular item-writing and test-assembly approaches with respect to the design features of the Dialogues and the relationship between these features and item difficulty.

The 17 independent variables are listed in Table 1.

Table 1 - The Independent Variables

1.	Number of content words in stimulus
2.	Presence of infrequent words in stimulus
3.	Presence of culture-specific words in stimulus
4.	Length of stimulus (in seconds)
5.	Utterance pattern
6.	Local coherence
7.	Negatives in stimulus
8.	Intonation
9.	Explicit/Implicit information tested by stem
10.	Degree of undirectedness of stem
11.	Negative in correct answer
12.	Negative in distractors
13.	Concrete object referred to in stimulus
14.	Role of speaker(s) crucial
15.	Location of speakers crucial
16.	Male or female second speaker
17.	Position of correct answer

Three kinds of variables were identified. The first is related to multiple-choice item design and TOEFL Section 1 test specifications. The *ETS Test Development Manual* (1992) provides specific points to keep in mind when developing multiple-choice items. These include surface characteristics of an item, such as the relative length and structure of the correct answer and distractors, option order, and complexity and clarity of stem wording. In addition, the *TOEFL Item Writing Guidelines for Listening Comprehension* (1994) establish additional program-specific guidelines for item writers. Although axiomatic approaches like these are common, evolving within the culture of multiple-choice test design, many have not been supported by empirical data (see Haladyna & Downing, 1989). In this study, Variable 3 (Culture-specific words), Variable 5 (Utterance pattern), Variable 7 (Negatives in stimulus), Variable 10 (Degree of undirectedness), Variable 11 (Negative in correct answer), Variable 12 (Negative in distractors), Variable 16 (Gender of speaker), and Variable 17 (Position of answer) were chosen to specifically examine some of the design characteristics treated in ETS or TOEFL item-writing guidelines.

The second kind of variable is related to processing. Researchers have argued that language processing requires both bottom-up and top-down strategies (Kamil, 1978; Rumelhart, 1980; Dechert, 1983; Anderson & Lynch, 1988). Variable 1 (Content words), Variable 2 (Word frequency), Variable 4 (Stimulus length), Variable 6 (Local coherence), Variable 8 (Intonation), and Variable 9 (Implicit/Explicit information) were selected with the semantic, syntactic, and prosodic aspects of bottom-up listening in mind -- considering that to process the items, the examinee will be relying on an internalized knowledge of the rules of spoken language. These variables are also related to a common view of listening as proposed by Rivers and Temperly (1978): "Listening is not a passive but an active process of constructing a message from a stream of sound with what one knows of the phonological, semantic, and syntactic potentialities of the language."

The third kind of variable is related to the context of the situation. For more than 20 years, theories of language have recognized the importance of the context in which communication occurs. Beginning with the work of Hymes (1972) and including Halliday (1978), Canale and Swain (1980), Canale (1983), and Bachman (1990), the ability to understand aspects of the social context in which communication occurs (including information about the roles of the participants, their relative status, their shared information, and the functions of their communications) has been considered a significant contributor to language competence.

Each Dialogue item is independent of the others; the context in which an exchange takes place is unrelated to the context of the next item. The physical setting, the roles and relationship of the speakers, and the purpose of their communication can vary with each item. Aside from the utterances of the two speakers, there are no clues as to the context of each situation. When listening to each exchange, the listener may attempt to determine the situation in which the two speakers are found, encountering numerous situations by the time the Dialogue subpart of the test is finished. Variable 13 (Concrete object), Variable 14 (Role of speakers), and Variable 15 (Location of speakers) were used to investigate the relationship between inferences involving context of situation and difficulty.

A more detailed description of each variable follows.

Variable 1 - Content Words. Variable 1 was the total number of content words in the two utterances in the stimulus. At first, the authors tried to identify a variable that would directly determine how many "information segments" there were in a stimulus. This is based on the assumption that in order to process the sound, the listener may somehow need to segment the sound -- "chunking the signal into syntactic groupings to reduce the load on memory" (Miller, 1956). Longer stimuli with more chunks of information were predicted to more heavily tax the memory and therefore be more difficult to process.

During the early stages of developing the classifications, establishing a system for operationally defining, and subsequently, counting, the potential "chunks" in a stimulus was challenging. Using the grammatical entities of phrases and clauses proved unsuccessful, as were measures of propositional relationships, mainly because the spoken language of the stimulus contained fragments that could not be classified and counted systematically using these concepts. Most propositional models have been devised for analysis of the connected discourse that occurs in written prose and were not helpful for quantifying the short speech samples in the Dialogues.

The pause unit (Brazil, 1983) was also considered as a basis for quantifying this variable; but it proved unwieldy because the prosodic information upon which the pause unit is based is mostly supplied by stress, and stress is perceived through a combination of loudness, pause, and pitch movements. When the authors and the test specialists who carried out the classifications tried to analyze the stress patterns and count the pause units by listening to the Dialogues, they found discrepancies in how the different listeners perceived the stress patterns; therefore the raters could not be calibrated.

The authors then modified their approach by counting the content words (i.e., nouns, verbs, adjectives, and adverbs) in the stimulus as a way of estimating the relative information load of the Dialogues. This modification was related to the pause unit in that it recognized the importance of stress in relation to the information load of the utterance, since content words receive the stress in spoken English (Selkirk, 1984). In addition, the number of content words could be ascertained reliably by the raters in the study.

Variable 2 - Word Frequency. The frequency of the content words was considered an important variable to include in the study. The authors noted that items in the TOEFL Dialogue pool tended to be difficult when answering the item correctly was linked to a specific word or idiom. For the purposes of the study -- since the items are based on a spoken stimulus -- the frequency of the words in the stimulus was determined according to a frequency list of spoken English, and not on one based on analyses of written text. Numerous word frequency lists were considered; a list compiled by Berger (1977) was chosen because it was the only one with all of the following critical features:

- based entirely on conversations in the United States

- based primarily on adult conversations

- based partially on university student conversations

- included a large number of words in corpus (100,000)

- arranged words in alphabetical order

Items with content words in the stimulus that did not appear in Berger's word frequency list were presumed to be more difficult.

Variable 3 - Culture-Specific Vocabulary. The TOEFL Committee of Examiners has suggested that if items contain vocabulary that is culturally bound to the experience of university life in the United States and Canada (e.g., "Intro to Lit," "Freshman Comp"), they will be more difficult for TOEFL examinees. (See TOEFL Committee of Examiners meeting report of spring 1993.) The TOEFL Program has received letters from test center supervisors who relate that cultural vocabulary may be an undue source of difficulty for examinees who have not been to the United States (60% of the TOEFL examinees take the test overseas). In an attempt to address the issues raised by these perceptions, three item classifications were defined regarding cultural vocabulary:

Type A - No cultural vocabulary

Type B - Cultural vocabulary, the understanding of which is not critical to answering correctly

Type C - Cultural vocabulary, the understanding of which is critical to answering correctly

Vocabulary that is specific to North American campus life (e.g., "Chemistry 102") or to a particular field or topic (e.g., "pollution report") was considered to be culture-specific, as were nonmetric measurement terms. It was hypothesized that Type C items (with cultural vocabulary that needs to be understood to answer the item correctly) would be most difficult. It was also recognized that few Type C items would appear in the sample due to ongoing efforts by test developers to use culture-specific vocabulary sparingly in order to address the concern of the Committee of Examiners.

Variable 4 - Length of Stimulus. Because the examinee hears each Dialogue only once, it was presumed that the difficulty of the item would relate to the length of the stimulus. It was not clear, however, what the nature of the relationship would be. One could hypothesize that the longer items would be easier because examinees would be provided with more context, which would facilitate comprehension. On the other hand, one could also propose that the longer items would be more difficult, as examinees would have more language to retain and understand. The authors wanted to determine if the data from the sample could provide an indication of a relationship between utterance length and difficulty in Dialogues and the nature of what that relationship might be. The length was measured in seconds.

Variable 5 - Utterance Pattern. The relationship between the two utterances in the stimulus varies in Dialogue items; the most common is the question-statement format. The four patterns are:

Question-statement	(Q-S)
Question-question	(Q-Q)
Statement-question	(S-Q)
Statement-statement	(S-S)

The *TOEFL Listening Comprehension Test Specifications* (1994) state that there should be variety in the utterance patterns of the two speakers in a Dialogue. Because variety is a part of the test specifications, the authors were interested in investigating whether a relationship exists between the utterance patterns and item difficulty. Because the Q-S pattern appeared to be the most predictable kind of exchange (i.e., a question would likely be followed by its answer), it was assumed to be the easiest. Initial analyses of the four patterns, however, indicated that the means of the Q-S subgroup and the S-S subgroup were higher than the means of the Q-Q subgroup and the S-Q subgroup. In addition, in the data set, the cell sizes of the four subgroups were very different. Based on these initial results, Q-S and S-S items (both ending with a statement) were combined and assigned a classification of 1, and Q-Q and S-Q items (both ending with a question) were assigned a 0. The authors recognized that in combining the categories some information might have been lost, but it was hoped that the cell sizes of the combined categories would be more comparable and their results more meaningful.

Variable 6 - Local Coherence. In some items, the coherence between the first and second speakers' utterances is supported by explicit links, such as lexical links (e.g., lexical repetition) and structural links (e.g., anaphora). In other items, the connection is not explicit, but can be inferred. Consider the following exchange:

> (woman) There's a thunderstorm watch for this afternoon.
>
> (man) Oh, no! I was just about to work in the garden.

In this stimulus, there are no structural or lexical links between the utterances, giving no clues as to what the connection is between a thunderstorm watch and working in the garden. It was predicted that an item without structural or lexical links would be more difficult than an item with them.

Variables 7, 11, and 12 - Negatives. Freedle and Kostin (1993) reported that the number of negatives in an item contributes to item difficulty in reading comprehension items. In this study, negatives were analyzed separately in the stimulus, the correct answer, and the distractors. Negative markers (e.g., "no" and "not") were counted, as well as negative prefixes (e.g., "un-," "dis-"). Negative tags were also counted, even if their meaning was not negative. The rationale for counting tags with positive meaning as a negative is that for the nonnative listener, the presence of the negative form could be a source of confusion regardless of its meaning in an utterance. It was hypothesized that Dialogue items with negatives would be more difficult. Dialogues with no negative, one negative, and more than one negative in the stimulus were compared in Variable 7. Variables 11 and 12 looked at whether there was a negative in the correct answer or in the distractors, respectively.

Variable 8 - Intonation. Initially, the authors proposed assigning classifications that described the phonological characteristics of the Dialogue stimuli. These included the number of unstressed syllables relative to the total number of syllables and the number of intonation peaks or valleys that gave added cues to meaning. Because of the stress-timed nature of spoken English, the ratio of unstressed syllables to the total number of syllables in the stimulus was considered to be an important relationship affecting difficulty.

Given the scope of the current study, however, analysis of stress and intonation was not practical. The TOEFL item pool is stored in two places: the recording of each item is embedded in a longer recording of the test in which the item occurs. The script text and text of the stem and options reside in the TOEFL pool database. To collect the recordings of each Dialogue and create a master tape would require accessing excerpts from a great number of original recordings. Analyzing such a tape would require expertise, processes, and equipment that were not initially anticipated for this study.

In order to conduct initial investigations in this study, the recording scripts were studied, and items that had special script directions to the speaker were identified. These items had speaker directions indicating, for example, the tone of voice a speaker should use. The scripts included notations such as "enthusiastically," "worriedly," underlining of words to be spoken with special emphasis, and exclamation points. It was predicted that items with such directions would be easier than the others.

Variable 9 - Inference. Variable 9 identified items according to whether the information tested is explicitly or implicitly stated in the stimulus. The answer to an item that tests explicit information is often a paraphrase of what was stated in the stimulus. To answer an item that tests implicit information, it is often necessary to go beyond what is actually stated in the stimulus. Most of the Dialogues that tested inference had stems worded "What does the man/woman imply?" or "What does the man/woman imply about x?" (See Appendix 1, Example 3, for this kind of item.)

In many comprehension taxonomies, a distinction is usually made between skills requiring comprehension or recognition of explicitly stated information and the ability to draw an inference or to understand what is implied (Richards, 1983; Rost, 1990). Bloom's taxonomy (1956) distinguishes between the explicit "comprehension" level and the implicit "inferencing" level. In their propositional analysis, Kintsch and van Dijk (1978), who claim that their model applies to both reading and listening comprehension, discuss the notion of propositions that are implicitly included in a text. A text with an implicit proposition requires more complex processing than one in which the proposition is explicit. It was hypothesized that items testing implicit information would be more difficult than items testing explicit information.

Variable 10 - Degree of Undirectedness. Test development materials advise item writers to create "directed" stems, that is, questions that are worded specifically and clearly. The *ETS Test Development Manual* (1992) claims that "the test taker should not have to read the stem and options to figure out what the question is asking." If this acknowaxiom is valid, a Dialogue item with an undirected stem, for example, "What does the man mean?", might not give the test taker enough information as to what the question and answer are about, and therefore might be more difficult for examinees to answer than an item with a stem that directly refers to a particular point in the Dialogue, for example, "What does the man say about the computer program?" (The latter focuses the test taker's attention more specifically on the point tested.) It was also predicted that even if the stem were undirected, if all four options have the same grammatical subject, the item would be easier than when the options were not directed and had different subjects.

Three categories were identified:

Type A - Directed stem, directed options

Type B - Undirected stem, directed options

Type C - Undirected stem, undirected options

Variables 11 and 12 - See Variable 7 (pages 7-8).

Variable 13 - Concrete Object. In Variable 13, items were classified according to whether the speakers refer to a concrete object in their shared environment or experience. Listeners might need to recognize the existence of this object in the setting of the Dialogue to make sense of the stimulus.

Variable 14 - Role of Speakers. In Variable 14, raters judged whether the language of one of the speakers is linked to a specific role the speaker plays. For many Dialogues, the situations are somewhat similar; they tend to represent experiences common to young adults in the university setting (e.g., too much noise in the dormitory, problems with a lab experiment), and the speakers take on an anonymous "every student" role. In other cases, the speakers' exchange is of a very general nature and could be inferred to be spoken by practically anyone without misunderstanding the gist of the Dialogue or the speakers' intentions. For some items, however, the identity of the speakers diverges from the "every student" and "any person" roles. The language of the speakers and their communicative function is directly linked to some specialized role.

The following example exhibits a specialized role (and a probable location).

(man) I'm looking for a warm jacket.

(woman) We have some very nice ones marked down.

(narrator) What does the woman mean?

When processing this item, it would be helpful to assume that the woman is a sales clerk (and that the speakers are probably situated in a store that sells clothing). The authors posited that these types of items might require more inferences about the context of the situation than the general student-life items require.

Variable 15 - Location of Speakers. In Variable 15, the items were classified according to whether the listeners needed to assume a particular physical location of the speakers to make sense of the stimulus.

Variable 16 - Speaker Gender. Since female voices are generally higher than male voices, their acoustic properties are believed to be more likely to present complications for recording and playback at the test administration sites. Related to this, concerns have been raised that the female voices on the recordings might be more difficult to understand than the male voices. Since Dialogues are designed so that the stem of an item usually focuses on the second speaker, investigating the relationship between item difficulty and the gender of the second speaker was possible.

Variable 17 - Position of Correct Answer. Changing the position of the correct answer was found to affect the IRT parameter estimates of Listening Comprehension items when all other variables are held constant (Golub-Smith, 1987). The authors were interested in further investigating key position and attempting to determine whether it contributed to item difficulty when other features were not held constant.

In the Dialogue item, the examinee has 12 seconds to choose an answer. The authors posited that when the answer is option A, the examinees may not need to read the other three options, and the item may be easier. Conversely, items with the answer as option D might be more difficult because it is necessary to read through and process all four options before selecting a correct answer. This prediction is based on what Rost (1990) has described as a "well-established principle in memory research," that intervening stages (in this case, processing each option) between an event (listening to the Dialogue) and recalling the event (recognizing the correct answer) distorts the examinee's memory of the original event.

Classification of Variables

The authors developed a classification system for the analysis of the variables. (See Appendix 2 for the classification form that was filled out by the raters.) Some variables were continuous (e.g., Variables 1 and 4) and some were categorical (e.g., Variables 5 and 6). Continuous variables were classified with the corresponding numbers (e.g., for Variable 1, the total number of content words). Categorical variables with just two classifications were assigned 0 or 1. Categorical variables with more than two classifications were assigned either a letter (A, B, C, or D) or a number (0, 1, or 2).

Two test specialists were trained in the classification system. They independently classified 25 TOEFL Dialogue items, which the authors selected from the TOEFL item pool. The items were selected so they represented a variety of the variables in the study and a range of difficulty levels. Each item was assigned a value for each of the variables. A group discussion was held during which the classifications were compared and discrepancies were recorded. As a result of the discussion, the classification system was revised somewhat and additional features were identified and their corresponding classifications were included in the study.

Three additional test specialists were trained in the classification system, and 225 items from 15 recently administered operational tests were classified. Operational items were used because both the recordings of the stimuli and the written script were available for analysis. The test specialists independently classified the items for all of the variables. Each item was classified by two different test specialists. For these initial ratings, correlations were calculated for several variables. For example, the interrater reliability for Variable 5 was .91, and for Variable 7 it was .82. A record was kept of the discrepancies. Items that received discrepant classifications by the two raters were then classified by a third rater and discussed. All discrepancies were resolved through discussion.

During the preliminary analyses, it became apparent that the sample consisted of too few items at the upper and lower ends of the difficulty scale. Thus, 33 additional items from the easy and difficult portions of the operational pool were classified. These items were classified only for the features that were found to be significant during the early analyses. With the addition of these items, the total number of Dialogues in the sample was 283.

Table 2 summarizes the classifications of the independent variables and the predictions made for each.

Table 2 - Classifications and Predictions

		Predictions	
Variables	**Easiest**	**<----->**	**Most difficult**
1. Number of content words	few	<----->	many
2. Infrequent words 0. All words on Berger's list 1. At least one word not on Berger's list	none (0)	<----->	some (1)
3. Culture-specific words 0. None 1. Some, not related to correct answer 2. Some, related to correct answer	none (0)	<----->	some (1,2)
4. Length of stimulus (in seconds)	short	<----->	long
5. Utterance pattern 0. Question-question or statement-question 1. Question-statement or statement-statement	QS-SS(1)	<----->	QQ-SQ(0)
6. Local coherence (structural or lexical link)	link (0)	<----->	no link (1)
7. Negatives in stimulus 0. None 1. One negative 2. More than one negative	none (0)	<----->	many (2)
8. Intonation	cues (0)	<----->	no cues (1)
9. Explicit/Implicit information	explicit (0)	<->	implicit (1)
10. Degree of undirectedness 0. Directed stem, directed options 1. Undirected stem, directed options 2. Undirected stem, undirected options	directed (0)	<----->	undirected (2)
11. Negative in correct answer	no (0)	<----->	yes (1)
12. Negative in distractors	no (0)	<----->	yes (1)
13. Concrete object referred to	no (0)	<----->	yes (1)
14. Role of speaker(s) crucial	no (0)	<----->	yes (1)
15. Location of speakers crucial	no (0)	<----->	yes (1)
16. Male or female speaker	male (0)	<----->	female (1)
17. Position of answer	A,B,C	<----->	D

Data Structure of the Dependent Variable

The dependent variable used in this study was Delta, which is an item difficulty index from classical test theory. Delta is a function of the proportion of correct responses in an item and, theoretically, is normally distributed with a mean of 13 and a standard deviation of 4. For the TOEFL Dialogue item pool, the mean Delta is about 10.8 and the standard deviation is about 1.5. The Delta frequency distribution for the 283 items used in this study is presented in Figure 1, and the descriptive statistics are summarized in Table 3.

As indicated in Table 3, the mean Delta and standard deviation of the 283 items used in this study were sufficiently close to the population mean Delta and standard deviation of the Dialogue item pool for the sample to be considered representative of the Dialogue item pool to which the results of this study were to be generalized. Figure 1 illustrates that the number of items at the two tails of the Delta distribution was small and that a large portion of items was concentrated in the middle Delta range, that is, more than 50% of the items were in the Delta range of 9.5 to 12.0, and less than 4% of the items had Deltas less than 8.0 or greater than 14.4. Because the effect of the independent variables on items with a Delta between 10.0 and 12.0 might not have been as large as on items with a Delta below 10.0 or above 12.0, the data structure might not have provided maximum power to detect the true impact of an independent variable on Delta (i.e., might not have been able to detect that a variable has an impact on Delta when a true impact exists).

Figure 1
Delta Distribution of the 283 Dialogue Items

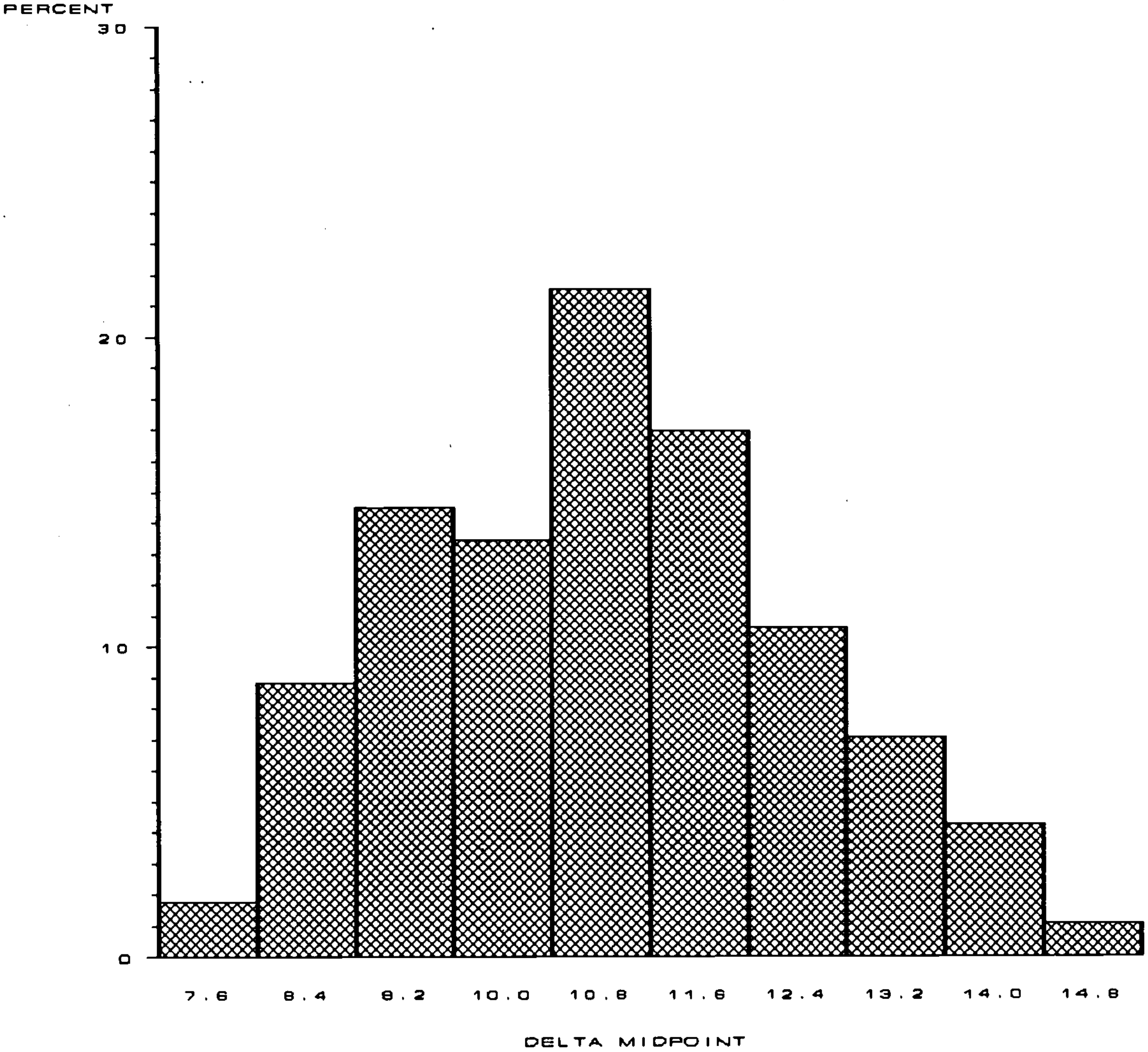


Table 3
Descriptive Statistics of Delta Based on 283 Items

Minimum	Maximum	Mean	Std Dev	Skewness	Kurtosis
7.60	15.0	10.89	1.56	0.22	-0.47

Analysis of the Continuous Variables

Of the 17 independent variables investigated in this study, Variables 1 (Content words) and 4 (Length of stimulus) were continuous and the other 15 variables were categorical. A multiple regression model was used to investigate the relationship between Delta and Variables 1 and 4. The regression model and the hypotheses tested that are related to this model are listed in Table 4.

Table 4
Regression Model and Hypotheses

Model

$$\Delta = \beta_0 + \beta_1 * VAR1 + \beta_2 * VAR4 + \beta_3 * VAR1 * VAR4 + \epsilon.$$

Hypotheses

H_o: $\beta_3 = 0$: No interaction between Variables 1 and 4.
H_o: $\beta_1 = 0$: Variable 1 does not have a significant impact on Delta.
H_o: $\beta_2 = 0$: Variable 4 does not have a significant impact on Delta.

In addition, each of these variables was plotted against Delta to explore the potential curvilinear relationship between the variables and Delta.

Analysis of the Categorical Variables

For the 15 categorical variables, the *t* test and analysis of variance (ANOVA) methods were used to detect the impact on Delta. Because the data used in this study were from items in previously administered tests, the number of items in each variable combination was unequal. Table 5 illustrates this unbalanced data structure by using Variable 2 (Word frequency), Variable 5 (Utterance pattern), and Variable 9 (Explicit/implicit information) as examples.

Table 5 - Unbalanced Data Structure
Number of Items in Each Variable Combination

		VAR2=0	VAR2=1
VAR9=0	VAR5=0	16	17
	VAR5=1	94	54
VAR9=1	VAR5=0	12	7
	VAR5=1	43	40

It is important to note that when the data are not balanced, the *t* test and ANOVA *F* test are more sensitive to the violation of the assumption that the variances of the dependent variable will be equal for different independent variable combinations. When this assumption is violated, the Type I error rate is no longer accurate (Box, 1954; Cochran, 1947; Rogan & Keselman, 1977).

In addition to the unbalanced nature of the data, certain variable combinations were not present among the 283 items. For example, there were no items with Variable 5 = 0, Variable 9 = 1, and Variable 14 = 1. When a variable combination is missing, the corresponding population marginal means for the variables are not estimable in ANOVA models (Milliken & Johnson, 1984). Thus, for the above example, the marginal means for Variable 5 = 0, for Variable 9 = 1, and for Variable 14 = 2 were not estimable in the three-way ANOVA model. Milliken and Johnson also pointed out that when some treatment combinations are not observed, it is best to consider the experiment as a one-way experiment and analyze each variable independently.

Because the data structure was unbalanced and there were missing values for some variable combinations, the following steps were used to evaluate the impact of the independent variables on Delta:

Step 1: Test whether the homogeneity of variance assumption is held for each variable and variable combinations of interest.

Step 2: Analyze the impact of each variable on Delta independently, using a *t* test if the variable has two levels and using a one-way ANOVA method if the variable has more than two levels. (The homogeneity of variance assumption did not hold for Variable 16, Gender of speaker, and hence, an approximate *t* test was used for that variable.)

Step 3: After the significant variables are identified in step 2, explore the combinations of these variables using ANOVA methods. Estimate the population Delta means for the significant variables and variable combinations and construct confidence intervals for the impact of the significant variables and variable combinations on Delta.

Step 1: Test of Assumptions. Both the *t* test and the ANOVA *F* test assume (1) the homogeneity of variance for different levels of an independent variable or for independent variable combinations; (2) the Delta populations from which the samples were selected are normal; and (3) the random error components for each item are independent. The second assumption is considered not to be critical in practice because it has been shown that both the *t* test and the *F* test for ANOVA are robust to violations of the normality assumption (Box, Hunter, & Hunter, 1978; Montgomery, 1984). The third assumption can be considered to be met in this study because each item was independently classified. However, both the *t* test and the *F* test are sensitive to violations of homogeneity of variance when sample size in each variable combination is not equal (Keppel, 1982), which was the case in the present study.

A statistical method has been developed to test the homogeneity of variance hypotheses, H_o: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$, where *a* is the number of levels for a single variable or the number of variable combinations. In this study, when a = 2, the folded form of the *F* statistic, *F′*, was used to test the homogeneity of variance hypothesis (Montgomery, 1984). (The *F′* statistic is the ratio of the larger variance to the smaller variance. The null hypothesis would be rejected if $F' > F_{\alpha/2,n1-1,n2-1}$ or if $F' < F_{1-(\alpha/2),n1-1,n2-1}$, where α is the Type I error rate and n_1 and n_2 are the sample sizes for the two levels of the variable. The details for this test are discussed by Montgomery, 1984.)

When *a* > 2, Bartlett's test (1937) was used to test the homogeneity of variance hypothesis $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$. This test was derived by Bartlett specifically for situations in which the sample sizes are not equal. Bartlett's test statistic *U* was computed based on the log transformations of the variances. (The hypothesis of homogeneity of variance is rejected if $U > \chi^2_{\alpha,a-1}$, where α and *a* as defined above. The details for Bartlett's test are discussed in Milliken and Johnson, 1984.)

The significance level used in this study for both the folded form of the *F* test and for Bartlett's test is $\alpha = 0.01$, based on a recommendation by Milliken and Johnson (1984).

Step 2: Identify Significant Variables. For the variables that met the homogeneity of variance assumption, the regular *t* test was used to compare the mean Delta differences for variables having two levels. The hypothesis tested was H_o: $\mu_1 = \mu_2$. The one-way ANOVA was used to detect the effect when a variable had more than two levels and the hypothesis tested was H_o: $\mu_1 = \mu_2 = \ldots = \mu_a$, where *a* is the number of levels for the variable.

It was found that for Variable 16 (Speaker gender), the homogeneity of variance assumption was not met, and so for Variable 16, an approximate *t* statistic was used to test the impact of the variable on Delta. This statistic is not distributed as exactly as *t* distribution. However, the distribution of this statistic is well-approximated by the *t* distribution with degrees of freedom approximated by Satterthwaite's (1946) formula.

Five independent variables, Variable 2 (Word frequency), Variable 5 (Utterance pattern), Variable 7 (Negative in stimulus), Variable 9 (Explicit/implicit information), and Variable 14 (Role of speakers), were identified as having a significant impact on Delta. Among these variables, Variable 7 had three levels; the other four variables had two levels each. Duncan's (1955) multiple range test was used for Variable 7 to compare Delta means at the three levels. Duncan's test was chosen from the many multiple comparison tests available because it adequately controls the type I comparisonwise error rate, and because it is powerful for detecting differences between means when real differences exist (Montgomery, 1984).

Step 3: Explore Combinations of the Significant Variables and Confidence Intervals. After the five significant variables were identified, the effects of their combinations were investigated. Three-way ANOVA was conducted for Variable 2 (Word frequency), Variable 5 (Utterance pattern), and Variable 9 (Explicit/implicit information). Variable 7 (Negative in stimulus) and Variable 14 (Specialized role of speaker) were, however, very unbalanced. This lack of balance resulted in missing variable combinations. Therefore, Variables 7 and 14 were included in the two-way ANOVA, but not in the higher order ANOVA. The statistical model for Variables 2, 5, and 9 is summarized in Table 6.

Table 6
ANOVA Models

Three-Way ANOVA for Variables 2, 5, and 9

Model

$\text{Delta}_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$.

where, μ is the overall mean Delta, α_i, β_j, γ_k are the variable effects, i, j, k = 1 or 2, $(\alpha\beta)_{ij}$, $(\beta\gamma)_{jk}$, and $(\alpha\gamma)_{ik}$ are the pairwise interactions, $(\alpha\beta\gamma)_{ijk}$ is the three-way interaction, and ϵ_{ijkl} is the random error for an item, l=1,...283.

Two-Way ANOVA for Variables 7 and 14

Model

$\text{Delta}_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijl}$.

where, μ is the overall mean Delta; α_i, β_j are the variable effects, i = 1 or 2, j = 1 , 2, or 3 for Variable 7, and j = 1 or 2 for other variables; $(\alpha\beta)_{ij}$ is the pairwise interactions; and ϵ_{ijl} is the random error for an item, l = 1,...,283.

For the ANOVA tests, the Type III sum of squares was used to test an individual effect (Freund & Littell, 1981). The Type III sum of squares is interpreted as the reduction in the sum of squares due to the effect of interest after other effects have been included in the model. For example, when testing H_o: $(\alpha\beta)_{ij} = 0$ in the two-way model, the Type III sum of squares can be described by the reduction notation, $R(\alpha\beta|\mu\ \alpha\ \beta)$, which can be interpreted as: the reduction in Delta variation due to $\alpha\beta$ interaction after including the effects of α and β in the model.

Four combinations of Variables 2, 5, and 9 were formed to investigate the magnitude of the joint effects of these variables on Delta. These four combinations are summarized in Table 7. Because Variable 7 and Variable 14 were very unbalanced, the number of items in some combinations involving Variables 7 and 14 was too small to make the combination effect meaningful. Therefore, the effects of combining Variables 7 and 14 were not investigated.

Table 7 - Variable Combinations

(VAR2=0 VAR5=0 VAR9=0) vs.
(VAR2=1 VAR5=1 VAR9=1)

(VAR2=0 VAR5=0) vs.
(VAR2=1 VAR5=1)

(VAR2=0 VAR9=0) vs.
(VAR2=1 VAR9=1)

(VAR5=0 VAR9=0) vs.
(VAR5=1 VAR9=1)

With unbalanced data, the least square mean is considered to be the best estimate of the marginal means (Milliken & Johnson, 1984). Least square parameter estimates are obtained by minimizing the sum of squares of the residuals for the ANOVA model. For a two-way model, for example, the least square means for variable A can be described by $\bar{X}_{i.} = \hat{\mu} + \hat{\alpha}_i + \bar{\beta}_{.} + (\hat{\alpha}_i\hat{\bar{\beta}})$ and for variable B can be described by $\bar{X}_{.j} = \hat{\mu} + \bar{\alpha}_{.} + \hat{\beta}_j + (\bar{\alpha}_{.}\hat{\beta}_j)$, where α is the effect of variable A and β is the effect of variable B.

The differences between the least square means for variable combinations were estimated, and the 95% confidence intervals for the differences were computed. For the one-way ANOVA model, which was used to test the significant impact of each variable on Delta, the least square mean is identical to the weighted marginal mean. The Delta differences for each of the five significant variables were estimated, and the 95% confidence intervals for the differences were computed.

In summary, for the continuous variables, a regression method was used to evaluate the relationship between the independent variables and Delta. For the categorical variables, *t* tests and ANOVA *F* tests were used. The effects of various combinations of variables were also investigated with the ANOVA method.

Results and Discussion

Insignificant Variables

Descriptive statistics for Variable 1 (Number of content words) and Variable 4 (Length of stimulus) are summarized in Table 8.

Table 8 - Descriptive Statistics for Variables 1 and 4

Variables	Mean Value	Standard Deviation	Range	
			Minimum	Maximum
1	10.15	3.14	4.0 words	20.0 words
4	7.36	1.85	2.79 words	15.42 words

Regression analysis indicated that the two continuous variables, Variables 1 and 4, did not have a significant impact on Delta. As illustrated in Table 9, the regression coefficients were not significantly different from 0. The R^2 value for the regression equation used (see Table 4) was 0.005, that is, only 0.5% of the variations in Delta could be accounted for by Variables 1 and 4 and their interactions. Because the regression analysis indicated that the relationship between Variable 1 and Delta and Variable 4 and Delta was weak and the regression coefficients were not significant, further validation of the regression model was not necessary.

Table 9 - Significance Level of Regression Coefficients for Variables 1 and 4

Parameter Estimates	*T* test for Ho: Parameter = 0	*P* Value
$\hat{\beta}_1 = 0.11$	1.06	0.29
$\hat{\beta}_2 = 0.10$	0.71	0.48
$\hat{\beta}_3 = -0.01$	-0.99	0.32

Table 10 lists the ten categorical variables that did not have a significant impact on Delta. For each variable, the number of items at each level of the variable, the mean and standard deviation of Delta at each level of the variable, and the statistical results testing the effect of the variable on Delta are summarized. The results of tests for violation of the assumption of homogeneity of variance for each variable are also presented in Table 10.

As indicated in the last column of Table 10, using $\alpha = 0.01$ as a significant level, the homogeneity of variance assumption was met for all the variables except Variable 16 (Speaker gender). Therefore, the *t* test was valid for the variables other than Variable 16 that have two levels, and the *F* test was valid for the variables that have more than two levels. For Variable 16, the modified *t* test was used because the homogeneity of variance assumption was not met for this variable.

The *P* values for the effects of the variables reported in Table 10 indicate that these variables did not have a significant impact on Delta based on the conventional significant level, $p = 0.05$.

Table 10 - Insignificant Categorical Variables

					Variable Effect			Homogenous Variance
Variable	Levels of Variable	N[a]	Mean	S.D.	Test Statistic	df	P value[b]	P value[c]
3	VAR3=0	224	10.84	1.32				
	VAR3=1	19	11.30	1.47	1.18	2/		
	VAR3=2	7	10.63	0.95		247	0.31	0.485
6	VAR6=0	222	10.92	1.58				
	VAR6=1	61	10.81	1.46	-0.45	281	0.66	0.473
8	VAR8=0	37	10.83	1.26				
	VAR8=1	213	10.88	1.33	0.02	248	0.85	0.726
10	VAR10=0	96	11.00	1.54				
	VAR10=1	77	10.86	1.63	0.33	2/		
	VAR10=2	110	10.82	1.53		280	0.72	0.814
11	VAR11=0	206	10.88	1.46				
	VAR11=1	77	10.94	1.81	0.28	281	0.78	0.016
12	VAR12=0	140	10.98	1.52				
	VAR12=1	143	10.81	1.59	0.90	281	0.37	0.583
13	VAR13=0	208	10.84	1.56				
	VAR13=1	75	11.04	1.55	0.96	281	0.34	0.936
15	VAR15=0	217	10.83	1.57				
	VAR15=1	66	11.09	1.51	1.17	281	0.24	0.703
16	VAR16=0	114	10.90	1.51				
	VAR16=1	136	10.85	1.14	0.28	207.1[d]	0.78[d]	0.002
17	VAR17=A	71	11.16	1.71				
	VAR17=B	70	10.86	1.31				
	VAR17=C	72	10.93	1.71	1.46	3/		
	VAR17=D	70	10.62	1.44		279	0.23	0.029

Notes: [a] The number of items classified for each variable is not equal.

[b] When the variable has two levels, the hypothesis tested was H_0: $\mu_1 = \mu_2$ and the test statistic was the t statistic. When the variable has more than two levels, the hypothesis tested was H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ and the test statistic was the F statistic.

[c] When the variable has two levels, the hypothesis tested was H_0: $\sigma_1^2 = \sigma_2^2$ and the test statistic was the folded form of F statistic. When the variable has more than two levels, the hypothesis tested was H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ and the test statistic was Bartlett's statistic.

[d] Modified t test and Satterthwaite's degrees of freedom approximation were used due to $\sigma_1^2 \neq \sigma_2^2$

For Variable 6 (Local coherence), Variable 8 (Intonation), Variable 10 (Undirectedness), Variable 11 (Negative in answer), Variable 12 (Negative in distractors), and Variable 16 (Speaker gender), the mean Deltas at different levels of the variables were very similar, which was a clear indication that these variables did not have a significant impact on Delta. For Variable 3 (Culture-bound vocabulary), the majority of the items (224 items, 89.6%) did not have cultural vocabulary (Variable 3 = 0). The number of items that had cultural vocabulary (Variable 3 = 1 or 2) was too small (26 items, 10.4%) to provide a stable estimate of the impact on Delta. Therefore, the data distribution for Variable 3 was too unbalanced to draw any meaningful statistical inferences. For Variables 13 and 15, a slight increase in Delta occurred when an object (Variable 13 = 1) or the physical location of the speakers (Variable 15 = 1) was involved in understanding the items. The magnitude of increases in Delta was not large enough to be statistically significant, however.

A discussion of each of the insignificant variables follows.

Variable 1 - Content Words. The results indicated that the number of content words, in isolation from other variables, could not be used to predict difficulty. It was interesting to discover that the number of content words varied in the sample from as many as 20 in some Dialogues to as few as 4 in others. It appears that for the TOEFL examinees, the frequency of the words in the item may be more relevant than the actual number of words in the item. The following item, which had quite a low Delta (8.7) and a rather large number of content words (all of which appeared on Berger's list and were thus considered frequent in this study), illustrates this phenomenon.

(woman) I can't believe how long these lines are. I wonder if it's because the bank is closed tomorrow.

(man) No, there's some problem with the computers. They have to process everything by hand.

(narrator) What does the man mean?

(A) The bank is short staffed.
(B) The bank is closed.
(C) The bank's computers aren't working.
(D) The bank should get some computers.

It is possible that when there are more content words in the stimulus, the context of the Dialogue may be more easily established by the examinee. Also, there may be more redundancies in the language of the stimulus. Either of these possibilities might contribute to a lower difficulty level of an item.

Variable 3 - Culture-Specific Vocabulary. Only seven items in the study were classified as requiring comprehension of culture-specific vocabulary in order to answer the item correctly. Moreover, these seven items had a mean Delta of 10.63, ranging from 9.2 to 11.5. TOEFL test developers are sensitive to the fact that examinees may not be familiar with certain aspects of North American culture and attempt to design the items with other clues (e.g., linguistic clues) in the stimulus. Consider the following example:

(woman)	How far is your walk to school?
(man)	Miles, it seems.
(narrator)	What does the man imply?

(A) The walk is pleasant with a friend.
(B) He doesn't mind the long walk.
(C) The walk isn't easy for him.
(D) He has joined the school track team.

Though the item includes a nonmetric measurement ("miles"), answering the item correctly does not require knowing exactly how long a mile is. The intonation and the expression "it seems" cue the fact that the walk is rather long and not so easy.

Variable 4 - Length of Stimulus. Because the examinee hears each Dialogue only once, it was presumed that the difficulty of the item would relate to the length of the stimulus and the memory load it represented for the examinee. The length of the Dialogues ranged from as short as 2.79 seconds to as long as 15.42 seconds. The data indicated that no direct correlation existed between difficulty and length, however. The authors suppose that the Dialogue stimuli, even at their longest, may be of an inadequate length to provide information about whether stimulus length contributes to difficulty for listeners at the relatively high level of proficiency for which TOEFL is geared.

Variable 6 - Local Coherence. The results indicated that Variable 6, the local coherence of the two utterances in the stimulus, was not significant. Perhaps this could be explained by considering Schlesinger's (1968) view that in comprehension one first relies on semantic expressions and resorts to syntactic processing only in doubtful cases. Rost (1990) states that "syntactic context plays a very limited role in constraining word recognition." It may be that the aural vocabulary of TOEFL examinees is sophisticated enough to make reliance on the understanding of specific structures and direct lexical links less critical.

Variable 8 - Intonation. The lack of special intonation directions in the script was not found to contribute significantly to difficulty in this study. The mean Deltas of items with and without special script directions were similar (10.88 and 10.83, respectively). However, because the number of items that did have script intonation directions was small (N=37) and the actual recordings were not analyzed for this variable, these results are uncompelling. A future research study that conducts thorough analyses of the recorded stimulus material -- examining such features as stress, intonation, rate of speech, and pauses -- might be worthwhile to add to the descriptive data already collected on Dialogue items.

Variable 10 - Degree of Directedness. This variable behaved in a manner contrary to what was expected. The most difficult items were those with directed stems and options, and the easiest were the undirected ones (though the differences were not found to be significant). Perhaps due to the general nature of the question "What does the woman mean?" and the high frequency at which the undirected stem occurs, the stem itself becomes devoid of meaning for examinees, and they merely locate the one option that is true according to the conversation. When the stem is more specific, it is often necessary to understand the stem completely in order to find its more specific answer.

Variables 13 and 15 - Objects and Location. Of the three variables examining context of situation, Variables 13, 14, and 15, the data showed only Variable 14 (Role of speaker) to be statistically significant. However, Table 10 indicates that there was a trend with respect to Variables 13 and 15, with the mean of the items that did not include these variables being 10.84 and 10.83, respectively, and rising .20 and .26 when the variable is present. These data, considered along with the data on Variable 14, suggest that certain characteristics of these context-linked items, other than speaker role, may warrant further research.

Variable 16 - Gender of Second Speaker. The variation of difficulty of the items with a male second speaker versus those with a female second speaker was not significant. The authors assume this is due to the quality controls in the process of TOEFL recording. The voices used on TOEFL recordings are professional speakers who are chosen for the clarity of their voices, the degree to which their pronunciation represents standard American English, and their ability to present appropriate vocal characterizations. A test specialist directs the recording sessions to ensure appropriate delivery, and the recordings are later listened to by another test developer before they are approved for use in a test administration.

Variable 17 - Position of Correct Answer. The results showed that the position of the correct answer is not a predictor of difficulty when other variables are not controlled. The mean Deltas indicated that items with the correct answer in the "A" position are slightly more difficult than other items. This is the opposite of the predicted result, but the difference was not significant.

Significant Variables

Five variables, Variable 2 (Infrequent vocabulary), Variable 5 (Utterance pattern), Variable 7 (Negative in stimulus), Variable 9 (Implicit versus explicit information tested), and Variable 14 (Role of speaker), were identified as having a significant impact on Delta using either a *t* test or an *F* test. For each of these five variables, the number of items at each level of the variable, the mean and standard deviation of Delta at each level of the variable, and the statistical results testing the impact of the variable on Delta are summarized in Table 11. The results of testing homogeneity variance for each variable are also presented in Table 11.

Using p = 0.01 as a significance level, the homogeneity of variance assumption was not violated for any of the variables listed in Table 11. Therefore, the regular *t* test was used for the variables that have two levels, and the *F* test was used for variables that have more than two levels. The *P* values for the effect of these variables reported in Table 11 indicate that all five of these variables have a significant impact on Delta, that is, $p < 0.03$.

Table 11 - Significant Variables

					Variable Effect			Homogenous Variance
Variable	Levels of Variable	N[a]	Mean	S.D.	Test Statistic	df	*P* value[b]	*P* value[c]
2	VAR2=0	165	10.72	1.55				
	VAR2=1	118	11.14	1.54	2.29	281	0.023	0.982
5	VAR5=0	52	10.31	1.52				
	VAR5=1	231	11.02	1.54	3.02	281	0.003	0.946
7	VAR7=0	153	10.75[d]	1.58				
	VAR7=1	106	10.90[d]	1.45	4.28	2/		
	VAR7=2	24	11.74	1.69		280	0.015	0.490
9	VAR9=0	181	10.62	1.48				
	VAR9=1	102	11.38	1.58	4.07	281	0.0001	0.459
14	VAR14=0	258	10.82	1.54				
	VAR14=1	25	11.64	1.54	2.52	281	0.012	1.000

Notes: [a] The number of variables classified for each variable is not equal.

[b] When the variable has two levels, the hypothesis tested was H_o: $\mu_1 = \mu_2$ and the test statistic was the *t* statistic. When the variable has three levels, the hypothesis tested was H_o: $\mu_1 = \mu_2 = \mu_3$ and the test statistic was the folded form of *F* statistic.

[c] When the variable has two levels, the hypothesis tested was H_o: $\sigma_1^2 = \sigma_2^2$ and the test statistic was the *F* statistic. When the variable has three levels, the hypothesis tested was H_o: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and the test statistic was Bartlett's statistic.

[d] Not significantly different based on Duncan's multiple range test.

A discussion of each of the significant variables follows.

Variable 2 - Infrequent Vocabulary. The presence of words not on Berger's list was a significant variable. The mean Delta of items with words not on this list was 11.14; the mean Delta of items containing only words on this list was 10.72; the difference was significant at the 0.02 level. This supports the hypothesis that infrequent vocabulary is an important contributor to item difficulty.

This variable proved to be significant in spite of several limitations. Once the actual classifications were made, it was noticed that certain high frequency campus words, like "semester" and "textbook," were not on Berger's list, though 25,000 words were supposedly from campus conversations. In a future analysis of difficulty, it may be worthwhile to run a frequency count on the actual words in TOEFL pool items or previously disclosed TOEFL tests and examine infrequent words in these lists as a variable for predicting difficulty.

Another limitation is that using a frequency list based on single-word counts does not capture idiomatic expressions. For example, if an item has an expression like "wearing a sweater inside out," the expression "inside out" would not be counted as infrequent according to Berger's list, since the individual words "inside" and "out" are both on the list. An ideal list for future research should also include analyses of usage and collocations.

Variable 5 - Utterance Pattern. The results indicated that items with a statement in the second utterance (Q-S and S-S = 1) were significantly more difficult than those with a question (Q-Q and S-Q = 0). To further investigate the reason for this unexpected result, the authors examined items in the Q-S (1) group that had a Delta greater than 12.0. In these items, many of the distractors seemed to be challenging and attractive because they required understanding of the first speaker's utterance and expressed plausible responses to the first speaker's question. In addition, the second speaker in many of these items implied a "yes" or "no" response (rather than explicitly stating one) and provided additional information. The following item illustrates these features:

(woman) Has the technician called about the repairs yet?

(man) When he does, I'll have you talk to him.

(narrator) What does the man imply?

(A) He's already spoken to the technician.
(B) The woman should make the repairs herself.
(C) The woman should explain what needs to be repaired.
(D) The technician has already arrived.

If one understood the woman's question in this item, options A and D would be quite plausible responses (though not correct interpretations of the man's answer). In fact, the item analysis indicated that A and D were much more attractive than option B. The design of the item required comprehension of the entire exchange; partial comprehension and reliance on an expected answer would lead one astray. Designing distractors with plausible responses to the first speaker's utterance appears to be a reasonable way to create difficult items. A future study may consider examining predictability of responses, that is, how common or predictable the message of the second speaker is.

Variable 7 - Negative in Stimulus. In the case of Variable 7, Duncan's (1955) multiple range test was used to compare the mean Delta for three conditions: no negatives in the stimulus (Variable 7 = 0), only one negative in the stimulus (Variable 7 = 1), and more than one negative in the stimulus (Variable 7 = 2). The results reported in Table 10 indicated that there was no significant difference in mean Deltas when there were no negatives and when there was only one negative in the stimulus. When the number of negatives increased to more than one, however, the mean Delta value became significantly higher.

It is interesting to note that the presence of a negative in the stimulus contributed significantly to difficulty, but the presence of a negative in the options did not (see Tables 10 and 11). Moreover, the presence of two negatives in the stimulus was much more difficult (mean = 11.74) than one negative (mean = 10.90) or no negative (mean = 10.75). Consider the following example:

(man) I see you haven't even opened Carl's letter.

(woman) Nor do I intend to.

(narrator) What does the woman mean?

The Delta of this item was 13.7. Because there were two negatives in the stimulus ("haven't" and "Nor"), test takers had to understand not only that the letter hasn't been read, but also that it probably won't be read. The word "Nor" in the second speaker's utterance apparently contributed to the difficulty in interpreting and answering this item correctly.

Variable 9 - Implicit Versus Explicit Information Tested. The data examined in this study support the supposition that Dialogues testing implied information might be more difficult than those testing explicit information. One could hypothesize that this is due to some inherent difference in the item type associated with testing inferencing. As was stated earlier, most taxonomies of listening comprehension make some type of distinction between the processing of explicit and implicit information.

Another interpretation could be that method effects are involved with testing interpretive inferencing with multiple-choice items. Carston (1989) explains that, unlike in a formal logical proof, where all premises are explicitly given, interpretive inferring depends on some premises in a speech event that are implicit and must be supplied by the listener. In the case of interpreting the Dialogue item, the inference must be drawn partly on the linguistic clues in the Dialogue and partly on the examinee's interpretation of the communicative situation.

Because of the general nature of many stems (e.g., "What does the woman/man imply?" "What can be inferred from the conversation?"), there may be a wide range of logical inferences a proficient listener could draw for any one item. In some items, the utterances contain multiple propositions. This means that although a number of logical inferences could be drawn, the multiple-choice format allows for only one answer. To answer an item with the question "What does the man imply?" the examinee might think of a reasonable proposition that stems from any proposition in the man's statement. In the multiple-choice format, however, the correct answer is only one of the possible propositions. Sometimes the correct answer may be several steps removed from the examinee's proposition (see the item about the technician in the Variable 5 discussion). An examinee whose proposition does not match that of the test developer needs to discard or revise his or her initial inference and then select another one from the four choices. The examinee whose inference is reasonable but not the particular inference represented in the correct answer may be carrying out an additional inferencing activity -- first drawing an inference independently and then adjusting the inference based on the choices given.

During the item development process, reviewers check to see that the intended answer to an item matches their initial inferences. Care is taken to develop items with limited possibilities for multiple inferences. In the future, research that indicates how closely test developer inferences match candidate inferences might shed more light on the results of this study and inform future item development efforts. It would also be helpful to develop a template for inference items that limits the line of logical reasoning the examinee needs to follow in order to answer the item and therefore narrows the range of reasonable inferences a candidate would draw.

Variable 14 - Role of Speaker(s). Of the three variables examining context of situation, the data showed Variable 14 to be significant. When the language of one of the speakers was linked to a specific role the speaker plays and the role was not one of a casual acquaintance or classmate, the item was significantly more difficult.

These results could be interpreted in several ways. Perhaps the examinees lack the sociolinguistic competence that would allow them to make rapid inferences about situations with respect to speaker roles. (At the start of the study, it had been posited that these items might require more accurate inferences about the context of the situation than the general student-life items require.) Examinees may be unfamiliar with the special roles in the items and lack the schema that would enable them to understand the situation.

Another possible interpretation is that the variation in roles in the Dialogues is a test method factor. If most of the Dialogues in the test have speakers in an anonymous, "every student" role, the listener may expect all of the speakers to have such roles; when this expectation is violated, the examinee may be taken off guard and performance might then be negatively affected. Consequently, the item difficulty would be higher for that item.

Combinations of Significant Variables

After the five significant variables were identified, analyses were conducted using the ANOVA method to explore how combinations of these variables would affect difficulty. The results of the analysis of variance are summarized in Table 12.

Table 12 - Analysis of Variance--Variables 2, 5, and 9 (Dependent Variable--Delta)

Source of Variation	DF	Sum of Squares	Mean Square	F	*P* value
MODEL	7	79.12	11.30	5.14	0.0001
ERROR	275	604.83	2.20		
CORRECTED TOTAL	282	683.95			
		Type III SS			
VAR2	1	8.14	8.14	3.70	0.0554
VAR5	1	21.61	21.61	9.83	0.0019
VAR9	1	14.74	14.74	6.70	0.0102
VAR2*VAR5	1	0.96	0.96	0.44	0.5087
VAR2*VAR9	1	0.61	0.61	0.28	0.5995
VAR5*VAR9	1	1.14	1.14	0.52	0.4729
VAR2*VAR5*VAR9	1	5.17	5.17	2.35	0.1263

Variable 5 (Utterance pattern) and Variable 9 (Implicit/explicit information) each had a significant impact on Delta with $p = 0.0019$ and $p = 0.0102$, respectively. This finding is consistent with the results from the *t* tests reported in Table 11. For Variable 2 (Infrequent vocabulary), the significance level based on the Type III sum of squares is marginal ($p = 0.0554$), which indicates that when Variable 5 and Variable 9 and all the interaction effects were included in the ANOVA model, adding Variable 2 to the model produced only a moderately strong impact on Delta. The impact of Variable 2 on Delta was not as strong as the impact of Variables 5 and 9. The ANOVA results reported in Table 12 indicate that all of the interactions among Variables 2, 5, and 9 were insignificant. Therefore, each of these three variables had an independent impact on Delta.

Three-way ANOVAs with other variables were not conducted for Variable 7 (Negative in stimulus) and Variable 14 (Role of speaker) because of the unbalanced data structure of these two variables. Instead, the pairwise interactions between Variable 7 and the other variables, and Variable 14 and the other variables, were investigated using seven two-way ANOVA tests.

Confidence Intervals for Significant Variables and Variable Combinations

The least square mean differences for five variable combinations and the confidence intervals for the differences based on three-way ANOVAs are presented in Table 13. The mean differences for each of the five significant variables and the confidence interval for the differences based on a one-way ANOVA model are presented in Table 14. For the one-way model, the weighted means and the least square means are equivalent.

Table 13 - Difference Between Least Square Means for Delta and the Confidence Interval of Difference for Delta Under Different Variable Combinations

Variable	N	LS Mean Difference	95% CI for Difference
(VAR2=0 VAR5=0 VAR9=0) vs. (VAR2=1 VAR5=1 VAR9=1)	16 40	1.48	(0.63, 2.33)
(VAR2=0 VAR5=0) vs. (VAR2=1 VAR5=1)	28 94	1.22	(0.59, 1.85)
(VAR2=0 VAR9=0) vs. (VAR2=1 VAR9=1)	110 47	1.09	(0.38, 1.80)
(VAR5=0 VAR9=0) vs. (VAR5=1 VAR9=1)	33 83	1.38	(0.78, 1.98)

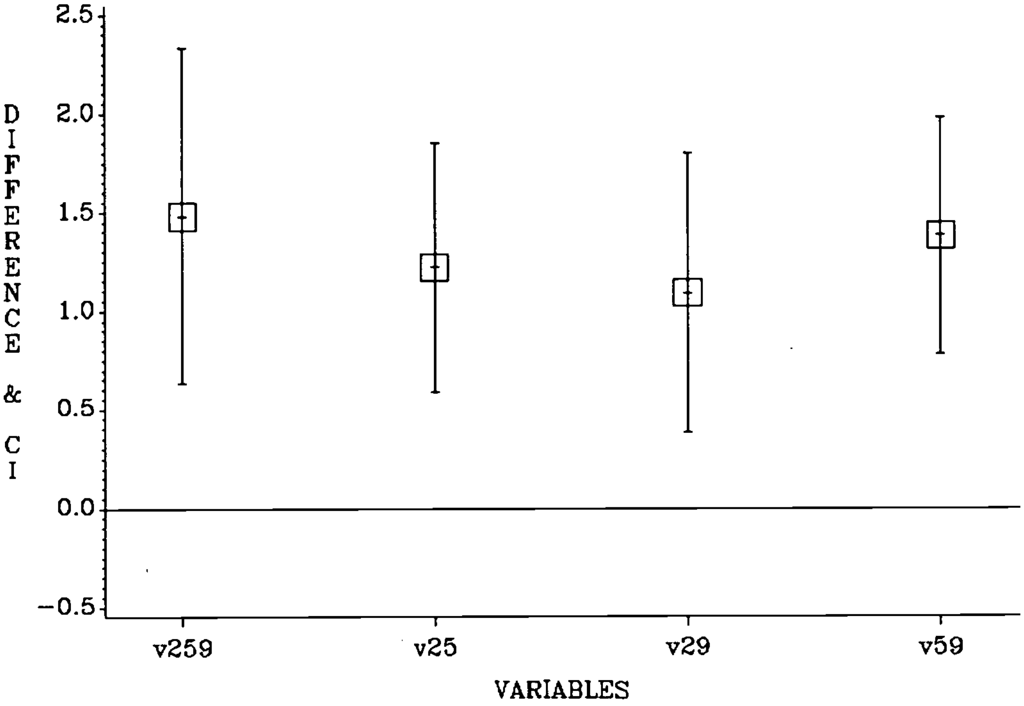
Table 14 - Difference Between Delta Means and the Confidence Interval of Difference for Delta for the Significant Variables

Variable	N	Mean Difference	95% CI for Difference
VAR2=0 vs. VAR2=1	165/118	0.42	(0.05, 0.79)
VAR5=0 vs. VAR5=1	52/231	0.71	(0.25, 1.17)
VAR7=0 vs. VAR7=2	153/24	0.99	(0.33, 1.65)
VAR7=1 vs. VAR7=2	106/24	0.84	(0.16, 1.52)
VAR9=0 vs. VAR9=1	181/102	0.76	(0.39, 1.13)
VAR14=0 vs. VAR14=1	258/25	0.82	(0.19, 1.45)

The means presented in Tables 14 and 15 indicate that the impact on Delta of combinations of variables was stronger than the impact of any individual variable. In fact, the mean Delta differences were greater than one for all the combinations of variables, and Delta differences were less than one for all the individual variables. Note that the combination of Variable 2 (Word frequency), Variable 5 (Utterance pattern), and Variable 9 (Inference) had the greatest impact on Delta. Tables 14 and 15 also report the confidence interval for each variable and for several combinations of variables. The mean Delta difference and the width of the confidence intervals are also illustrated in Figures 2 and 3. The boxes are the sizes of the mean Delta differences between the two levels of each variable or each variable combination, and the vertical lines encompass the confidence intervals for the differences. The confidence interval for Variables 2, 5, and 9 (illustrated in Figure 4) was narrower than for the other variables or variable combinations, which indicates that the impact of these three variables on Delta was more precisely estimated than it was for the other variables or variable combinations.

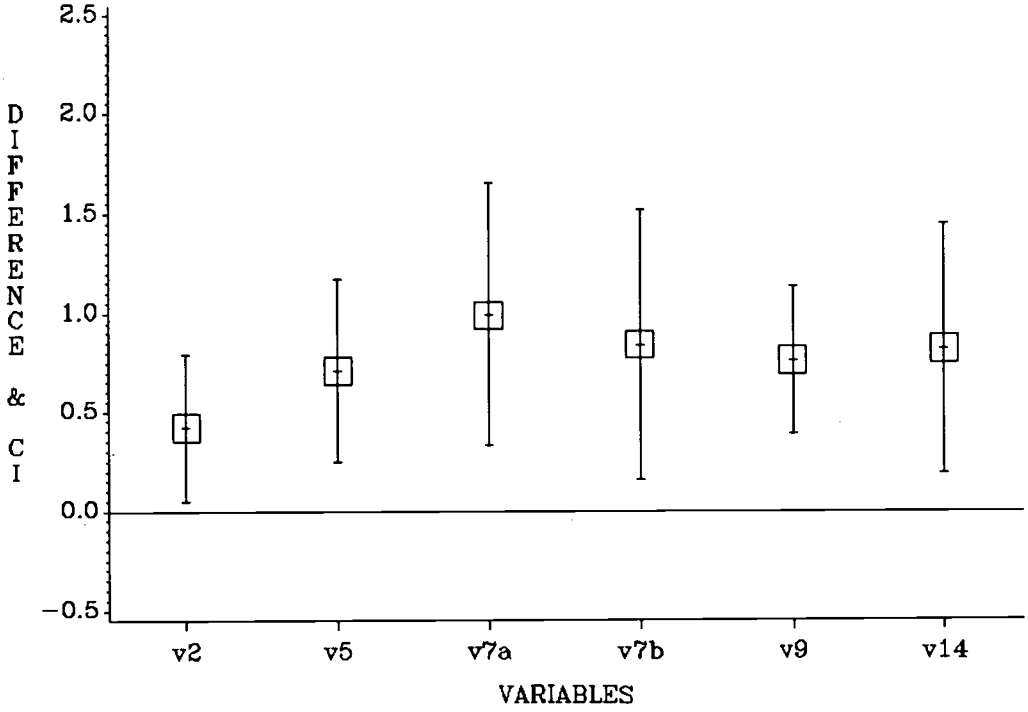
The confidence intervals in Tables 13 and 14 and Figures 2 and 3 provide useful information to item writers regarding the effects of these variables on difficulty. For example, an item with Variables 2, 5, and 9 = 1 can be predicted to have a Delta that will most likely be from .63 to 2.33 Delta points higher than an item with Variables 2, 5, and 9 = 0. In practice, this implies that an item containing infrequent vocabulary, an utterance pattern ending in a statement, and a question about the speaker's implication is more likely to have a higher Delta than an item without these features.

Figure 2
Mean Delta Differences and the Confidence Intervals
for Joint Variable Effects



Note: V259 indicates Variables 2, 5, and 9. The other labels can be interpreted in the same fashion.

Figure 3
Mean Delta Differences and the Confidence Intervals
for Individual Variable Effects



Note: V7a indicates the mean Delta difference between VAR7=2 and VAR7=0, and V7b is the mean Delta difference between VAR7=2 and VAR7=1.

Implications

The results obtained from this study indicate that five variables, Variable 2 (Word frequency), Variable 5 (Utterance pattern), Variable 7 (Negative in stimulus), Variable 9 (Explicit/implicit information), and Variable 14 (Role of speaker), had a significant impact on Delta. The joint impact of these variables on Delta was stronger than the impact of the variables considered individually.

Because the Delta distribution for the 283 items was clustered in the middle Delta range, and because the relationship between Deltas for the middle Delta range items and the independent variables was weaker than it was for the lower or higher Delta items, low R^2 values were observed for this study. (For example, the R^2 for the three-way ANOVA involving Variables 2, 5, and 9 was 0.12, that is, only 12% of the variation on Delta can be explained by these three variables.) Other sources of variations, which may also have affected item performance, were not controlled in this study, such as the examinee sampling and the voices on the recording. These variations may also have contributed to a low R^2. This study has achieved its major goal, however, which was to identify the variables that have a significant impact on item difficulty.

No direct link was observed between 12 of the variables and item difficulty. In one sense, this result is counterintuitive because, during the pretesting efforts of 1991 and 1992, many of these variables were included in pretest items in an effort to increase the number of difficult items in the TOEFL pool, and the resulting pretest data illustrated the success of the effort.

There are two possible explanations for these results: either the variables identified as insignificant do not contribute to difficulty, or the variables identified as insignificant do contribute to difficulty, but their contribution cannot be directly determined through this method of investigation. Within this sample of items the variables appeared together in uncontrolled combinations, allowing some variables to cancel the effects of others.

Because the items in the study were sampled from the existing TOEFL pool and were not created specifically for this research, the resulting data structure of the sample was unbalanced. For example, Variable 13 was present in only 75 items and was absent in 208 (see Table 10). This lack of balance, which occurred with several other variables as well, reduced the statistical power needed to detect the impact of these variables on Delta.

It is understandable that no single characteristic of a language item contributes definitively to the difficulty of the item. Theories of communicative competence suggest that language as communication is a complex phenomenon that calls on experiential knowledge, personal schema, and several competencies that each communicator draws on at different rates, at different times, and in different situations (see Bachman, 1990; Canale & Swain, 1980). Each Dialogue item could be interpreted as one of these situations, and every examinee could be considered a different communicator. In her discussion of online processing, Carston (1989) explains that "every word has phonetic, syntactic, and semantic properties and that ... hearers begin their analysis at each of these levels as soon as they can." How each individual carries out this analysis remains unclear. In addition, no TOEFL Dialogue item is heard in isolation; each examinee is listening to a series of many different communicative events.

Besides the complexity of the language construct itself, the multiple-choice item design contributes additional complexities. As previously mentioned, preliminary item analyses of candidate responses are used to confirm or refute the test developer's intuition, but to date -- partly because of the complexity of language processing mentioned above and partly because of the complexity of item design -- test developers and researchers have not determined why a particular examinee chooses or avoids a particular option.

The most controlled method for examining how a particular item characteristic contributes to difficulty would be to purposely create and administer items with a certain characteristic along with items that are exactly the same except for that characteristic, that is, to keep constant all other characteristics of the item. Under these conditions, if the items with the characteristic that contributes to difficulty were more difficult than the items without the characteristic, the results of the current study would be confirmed. In order to gain substantive information about the variables that were shown to be significant, research with experimental items is recommended. In a future study, it would be important for the values of the variables to have more balanced cell sizes.

Because items that test interpretive inference appear regularly in comprehension tests, numerous language testing programs would benefit from further investigation of this significant variable. A study of the comparison between multiple-choice answers to inference questions and examinees' own constructed responses to the same questions would greatly benefit the language testing field.

The data from the current study will inform language test development in several ways. As the confidence intervals from Table 14 showed, item writers can manipulate the occurrence of the significant variables in an item and more accurately predict an item's difficulty before pretesting.

Test developers might explore the possibility of modifying the TOEFL Test Specifications to control for the number of items in a test form that contain the significant variables. For example, it might be worthwhile to consider limiting the number of items where the speakers have specialized roles. (Currently, test assembly specifications call for general relevance to adult conversation settings.) An initial question would be to investigate how often this kind of item actually appears in TOEFL tests. It would also be important to determine why these items are difficult: Is it because they require higher proficiency in listening comprehension, or is there an extra processing step (of understanding who the speakers are) contributing to the difficulty? If it is the latter, then perhaps these items measure something in addition to listening comprehension.

For TOEFL, an automated test assembly program is used by test developers in assembling operational forms according to IRT parameters and content specifications. The effectiveness of this computer-based assembly tool depends on the extent to which the item content classifications in the data-base reflect the important features of items. At present, Dialogues already have a computer classification indicating whether they test explicit or implicit information; classifications for the other variables found to be significant in this study could be added.

The results of this study can also be considered once item development for TOEFL 2000 is under way. Clearer understanding with respect to the relationship between discourse features of listening stimuli and difficulty, as well as insights into possible method effects associated with certain item features can be taken into account when different tasks and texts are considered and trialed for use in TOEFL 2000.

Replication of this study using other Dialogue pools is strongly recommended. Other test programs (e.g., the Test of English for International Communication) that include the Dialogue as part of their test design can adopt a similar Dialogue classification scheme for use in a formal analysis of their own item pools.

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal, 76*(iv).

Anderson, A., & Lynch, T. (1988). *Listening*. Oxford: Oxford University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing* (pp. 81-100). Oxford: Oxford University Press.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Journal of Royal Statistical Society*, Supplement 4, 137-147.

Berger, K. W. (1977). *The most common 100,000 words used in conversations*. Kent, Ohio: Herald Publishing House.

Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook 1: Cognitive domain*. New York: McKay.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Annals Mathematical Statistics, 25*, 290-303.

Brazil, D. (1983). Intonation and discourse: Some principles and procedures. *Text, 3*(1), 39-70.

Brown, J. D. (1988). What makes a cloze item difficult? *University of Hawaii Working Papers in ESL, 7*(2).

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Carston, R. (1989). Language and cognition. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge survey. III Language: Psychological and biological aspects* (pp. 38-68). Cambridge: Cambridge University Press.

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics, 3*, 22-38.

Dechert, H. (1983). How a story is done in a second language. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication*. London: Longman.

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrics, 11*, 1-42.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper* (TOEFL Research Report 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service (1994). Item writing guidelines for the Listening Comprehension Section of the TOEFL (Internal ETS document).

Educational Testing Service (1994). TOEFL Listening Comprehension Test Specifications (Internal ETS document).

Educational Testing Service (Spring 1993). TOEFL Committee of Examiners Report of Meeting (Internal ETS document).

Educational Testing Service (1992). Test Development Manual (Internal ETS document).

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL Reading Comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. Unpublished research report (in progress).

Freund, R. J., & Littell, R. (1981). *SAS for linear models*. Cary, NC: SAS Institute Inc.

Golub-Smith, M. (1987). *A study of the effects of item option rearrangement on the listening comprehension section of the Test of English as a Foreign Language* (TOEFL Research Report 24). Princeton, NJ: Educational Testing Service.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1).

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Henning, G. (1990). TOEFL subtest functioning. Unpublished report to the TOEFL Committee of Examiners, Princeton, NJ.

Henning, G. (1991). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL Listening Comprehension items performance* (TOEFL Research Report 33). Princeton, NJ: Educational Testing Service.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected Readings*. Harmondsworth, Great Britain: Penguin.

Kamil, M. L. (1978). Models of reading. In S. Pflaum-Connor (Ed.), *Aspects of reading education*. Berkeley, CA: McCutchan.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd edition). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5).

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, *15*, 661-675.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data, Volume I: Designed experiments*. New York: Van Nostrand Reinhold Company.

Montgomery, D. C. (1984). *Design and analysis of experiments* (2nd edition). New York: John Wiley & Sons.

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(a), 219-239.

Rivers, W., & Temperley, M. S. (1978). *A practical guide to teaching of English as a second or foreign language*. New York: Oxford University Press.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA *F* test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. *American Educational Research Journal*, *14*, 493-498.

Rost, M. (1990). *Listening in language learning*. London: Longman.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Erlbaum.

Satterthwaite, F. W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

Schlesinger, I. M. (1968). *Sentence structure and the reading process*. The Hague: Mouton.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Stansfield, C. W. (Ed.). (1986). *Toward communicative competence testing: Proceedings of the Second TOEFL Invitational Conference* (TOEFL Research Report 21). Princeton, NJ: Educational Testing Service.

Way, W. (1990). *TOEFL 2000*. Paper presented at the meeting of the TOEFL Committee of Examiners, Princeton, NJ.

Appendix 1 - The Dialogue Item

In the current TOEFL test design, Dialogue items appear in Listening Comprehension (Section 1). For each item, the test taker hears a short conversation, called the stimulus, between two people. The stimulus is between 5 and 20 seconds long. Then a narrator asks a question about what was said. (This question is called the stem.) The test taker has 12 seconds to read four possible responses (options) in the test book, select the correct answer to the question, and mark it on the answer sheet.

Example Dialogues

1. Easy item (Delta = 9.8)

(woman) How's Mary been doing in her new job?
(man) I don't know. I only see her every now and then.

(narrator) What does the man say about Mary?

(A) She has been working night and day.
(B) She hasn't been on the job very long.
(C) He doesn't have much contact with her.
(D) He's on his way to see her now.

2. Medium-difficulty item (Delta = 11.5)

(woman) Did you hear how Mark did in the chess tournament?
(man) Yeah. What a pity!

(narrator) What does the man mean?

(A) He is ashamed of Mark.
(B) He heard Mark was the winner.
(C) He is sorry Mark did poorly in the tournament.
(D) He is surprised Mark was in the chess tournament.

3. Difficult item (Delta = 13.0)

(woman) What did you think of the new doctor at the infirmary?
(man) You mean Dr. Randolf? He was away attending a conference.

(narrator) What does the man imply?

(A) The doctor wasn't well.
(B) He didn't see the new doctor.
(C) The doctor was going to see him anyway.
(D) He went to a conference with Dr. Randolf.

Appendix 2 - Item Classification Form

ITEM #____

_______ 1. Number of content words in stimulus

_______ 2. Number of words in stimulus not on Berger's list

_______ 3. Cultural vocabulary
0. None
1. Some, not related to correct answer
2. Some, related to correct answer

_______ 4. Length of stimulus (in seconds)

_______ 5. Utterance pattern
A. Question-statement
B. Question-question
C. Statement-question
D. Statement-statement

_______ 6. Local coherence
0. Structural or lexical link
1. No link

_______ 7. Number of negatives in stimulus

_______ 8. Intonation cues
0. Cues
1. No cues

_______ 9. Explicit/Implicit information
0. Explicit
1. Implicit

_______ 10. Degree of undirectedness
0. Directed stem, directed options
1. Undirected stem, directed options
2. Undirected stem, undirected options

_______ 11. Negative in correct answer
0. No negative
1. Negative

_______ 12. Negative in distractors
0. No negative
1. Negative

_______ 13. Concrete object
0. No object
1. Object

_______ 14. Special role
0. No role
1. Role

_______ 15. Physical location
0. Not relevant
1. Relevant

_______ 16. Gender of second speaker
0. Male
1. Female

_______ 17. Position of correct answer (A, B, C, or D)

Cover Printed on Recycled Paper

57906-09064 • Y26M.5 • 275708 • Printed in U.S.A.