




**TOEFL**<sup>®</sup>

# Research Reports

*RR - 79*  
*July 2004*



Exploring Item Characteristics  
That Are Related to the  
Difficulty of TOEFL  
Dialogue Items

Irene Kostin

**Exploring Item Characteristics That Are Related to the  
Difficulty of TOEFL Dialogue Items**

Irene Kostin

ETS, Princeton, NJ

RR-04-11



*ETS is an Equal Opportunity/Affirmative Action Employer.*

Copyright © 2004 by ETS. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, Graduate Record Examinations, GRE, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service. The Test of English as a Foreign Language is a trademark of Educational Testing Service.

College Board is a registered trademark of the College Entrance Examination Board.

## **Abstract**

The purpose of this study is to explore the relationship between a set of item characteristics and the difficulty of TOEFL<sup>®</sup> dialogue items. Identifying characteristics that are related to item difficulty has the potential to improve the efficiency of the item-writing process. The study employed 365 TOEFL dialogue items, which were coded on 49 variables, including 5 significant variables reported in Nissan, DeVincenzi, and Tang (1996). Of the 5 significant variables in Nissan et al., 3 correlated significantly with item difficulty in this study. Another 11 met a critical probability criterion. These 11 included representatives from three broad categories of variables: 2 in the category of word-level factors, 1 in the category of discourse-level factors, and 8 in the category of task-processing factors. Multiple regression analyses indicate that the variables in this study account for about 40% of the variance in item difficulty.

Key words: English language learning, English as a second language (ESL), item difficulty, listening comprehension, test items, Test of English as a Foreign Language<sup>™</sup> (TOEFL<sup>®</sup>), Test of English for International Communication<sup>™</sup> (TOEIC<sup>®</sup>)

---

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 board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously 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 in consultation with the TOEFL Committee of Examiners. Its 12 members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, reviews and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL 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 and applied linguistics. 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 (2003-2004) members of the TOEFL Committee of Examiners are:

Micheline Chalhoub-Deville	University of Iowa
Lyle Bachman	University of California, Los Angeles
Deena Boraie	The American University in Cairo
Catherine Elder	Monash University
Glenn Fulcher	University of Dundee
William Grabe	Northern Arizona University
Keiko Koda	Carnegie Mellon University
Richard Luecht	University of North Carolina at Greensboro
Tim McNamara	The University of Melbourne
James E. Purpura	Teachers College, Columbia University
Terry Santos	Humboldt State University
Richard Young	University of Wisconsin-Madison

---

To obtain more information about the TOEFL programs and services, use one of the following:

**E-mail:** [toefl@ets.org](mailto:toefl@ets.org)

**Web site:** [www.ets.org/toefl](http://www.ets.org/toefl)

### **Acknowledgments**

I would like to thank Susan Nissan for providing valuable background information concerning TOEFL dialogue items and also for providing details concerning how the significant variables in her study of TOEFL dialogue items were coded. I would also like to thank Marc Tolo for doing the coding that was essential for determining the intercoder reliability for several of the variables in this study. Additionally, I would like to thank Fred Cline for carrying out complex statistical analyses for this study. Finally, I would also like to thank the reviewers of this report—Isaac Bejar, Neil Dorans, Dan Eignor, Catherine Elder, and Susan Nissan—for their helpful and informative comments and suggestions.

## Table of Contents

	Page
Introduction.....	1
Literature Review.....	1
Word-level Factors .....	2
Sentence-level Factors.....	3
Discourse- level Factors .....	4
Task-processing Factors .....	5
Method.....	6
Data.....	6
Variables Assessing Item Characteristics.....	6
The Coding .....	13
Results and Discussion .....	15
Conclusions and Implications.....	27
Future Studies.....	28
References.....	29
Appendixes	
A - Coding Instructions.....	32
B - Instructions for Coding Lexical Overlap .....	52

## List of Tables

	Page
Table 1. Intercoder Reliability Based on 60 TOEFL Dialogue Items From Two TOEFL Forms .....	14
Table 2. Correlation of Variables With Item Difficulty (Equated Delta).....	16
Table 3. Results of Stepwise Multiple Regression, With Only Significant Variables Remaining in the Equation .....	26





## **Introduction**

The purpose of this study is to explore the relationship between a set of item characteristics and the difficulty of TOEFL® dialogue items, an item type currently included in the Listening Comprehension Section of TOEFL. As part of this purpose, the study attempts to replicate the significant findings reported by Nissan, DeVincenzi, and Tang (1996). The study also investigates additional variables that were not included in the Nissan et al. study.

The ability to predict the difficulty of TOEFL dialogue items could improve the efficiency of the item-writing process. Statistical specifications for TOEFL dialogue items as well as for other item types call for items with a relatively wide range of difficulties. When assembling a test, occasions arise where there are shortages of items at certain difficulty levels. For example, Nissan et al. (1996) reported an occasion where there was a shortage of difficult TOEFL dialogue items in the item pool such that, if the pool were not replenished, specifications of future tests would not be met. More recently, there has been a shortage of easier TOEFL dialogue items (Marc Tolo, personal communication, 2002). A knowledge of the characteristics that are associated with harder or easier items could help item writers produce items of the desired level of difficulty.

## **Literature Review**

The literature reviewed below will include studies not only in the area of listening comprehension but also in the area of reading comprehension. The inclusion of reading comprehension studies is based on findings in the literature of similarities between reading and listening. For example, Kintsch, Kozminsky, Streby, McKoon, and Keenan (1975) presented college students with paragraphs for reading and listening that were matched for number of propositions. The time allowed for reading was limited to that needed to present the paragraphs orally. The researchers found that the level of recall, measured by the number of propositions correctly recalled, was virtually identical for both methods of presentation. Kintsch et al. also reported that while paragraph length and number of different arguments contained in the paragraphs affected recall accuracy, these effects did not differ for reading versus listening. They concluded that the processes underlying reading and listening are probably similar. Studies by Kintsch and Kozminsky (1977) and Smiley, Oakley, Campione, and Brown (1977) also support

this conclusion. Other studies have reported high intercorrelations between reading and listening tests (see review by Sticht & James, 1984, pp. 293-317).

The Nissan et al. (1996) variables that this study is attempting to replicate are discussed in the appropriate sections below. These variables are: the presence of infrequent oral vocabulary discussed in the section on word-level factors, the presence of negatives in the dialogue discussed in the section on sentence-level factors, the sentence pattern of the utterances in the dialogue and the roles of the speakers in the dialogue discussed in the section on discourse-level factors, and the necessity of making an inference to answer the items discussed in the section on task-processing factors. In their study, Nissan et al. used equated delta as the measure of item difficulty; higher values on this measure are associated with more difficult items and lower values are associated with easier items. Also, several of the factors listed below are discussed in *TOEFL 2000 Listening Framework*, by Bejar, Douglas, Jamieson, Nissan, and Turner (2000).

### ***Word-level Factors***

Past research has shown that the meaning of an unfamiliar word can often be inferred from the linguistic context in which it is embedded (Miller, 1999). However, the sparse linguistic context in TOEFL dialogues (ranging from 8 to 53 words in the current study) probably makes it difficult to infer the meaning of an unknown word from context, so one might expect that vocabulary knowledge will have a significant effect on the difficulty of TOEFL dialogue items. Employing TOEFL dialogue items in their study, Nissan et al. (1996) reported findings supporting this hypothesis. Their measure of vocabulary knowledge was the presence of an infrequent vocabulary word in the dialogue. A dialogue was coded as having an infrequent word if it contained a word that was not on a word list of 100,000 common words (Berger, 1977), a list based entirely on conversations in the United States, primarily between adults and some between university students. Nissan et al. found that the presence of an infrequent word in the dialogue was positively associated with item difficulty. The findings of a study by Kelly (1991) demonstrate the importance of vocabulary knowledge to listening comprehension in situations where the linguistic context is somewhat greater than in the case of TOEFL dialogues. Advanced English language learners in France both transcribed and translated English passages (ranging from 82 to 121 words) that they listened to. Kelly categorized their errors as perceptual, lexical, or syntactical; he also rated the errors in regard to whether they resulted in minimal

comprehension failure or severe comprehension failure. Kelly reported that lexical errors, typically in response to unfamiliar vocabulary, accounted for most of the errors where comprehension was severely impaired.

Phonological variables also may affect item difficulty. Henrichson (1984), for example, reported that the difference in listening comprehension between native speakers of English and nonnative speakers was greater when they listened to spoken English employing sandhi-variation than when they listened to spoken English without sandhi-variation. This finding supports the hypothesis that sandhi-variation makes comprehension of spoken language more difficult for nonnative speakers of English. *Sandhi-variation* refers to “the phonological modification of grammatical forms which have been juxtaposed” (Crystal, 1980, p. 311). Examples of sandhi-variation are *gonna* for *going to*, *wanna* for *want to*, and *hasta* for *has to*.

### ***Sentence-level Factors***

Several researchers have hypothesized that syntactic complexity affects listening comprehension such that the more complex the syntax is in a text, the more difficult it is to comprehend (Anderson & Lynch, 1988; Rost, 1990). A few findings support this hypothesis. Nissan et al. (1996) reported that the presence of more than a single negative in TOEFL dialogues was positively associated with item difficulty. In a related finding, Freedle and Kostin (1999) reported that the number of negatives present in TOEFL mini-talk passages was positively related to item difficulty. Using the number of dependent clauses in a dialogue as a measure of syntactic complexity, Buck and Kostin (1999a), in a pilot study, found that this measure was positively related to the difficulty of dialogue items in the Test of English for International Communication™ (TOEIC®).

In the area of reading, Abrahamsen and Shelton (1989) demonstrated improved comprehension of texts that were modified, in part, so that full noun phrases were substituted in place of referential expressions such as pronouns. This improvement in comprehension is hypothesized to have occurred because, in the modified condition, the test takers no longer had to figure out what the referentials were referring to. Consistent with this finding, Buck and Kostin (1999a) found that the presence of within-text referentials in TOEIC dialogues was positively related to item difficulty.

### ***Discourse-level Factors***

In the area of reading comprehension, several studies have shown that familiarity with the topic of a text facilitates text comprehension (McNamara, Kintsch, Songer, & Kintsch, 1996; Recht & Leslie, 1988; Spilich, Vesonder, Chiesi, & Voss, 1979). Using data from the TOEFL reading section, Hale (1988) reported results consistent with these findings: While the size of the effect was small, Hale found that students in two major field groups, the humanities/social sciences and the biological/physical sciences, performed better on passages related to their own groups than on other passages. Employing an immediate retrospective verbal report procedure, Yi'an (1998) investigated the comprehension processes involved when Chinese test takers, who were studying English as a foreign language, responded to multiple-choice questions about a recorded English language radio interview they had listened to; the protocols from this study showed that these test takers frequently used their background knowledge about the topic of the interview when responding to the multiple-choice questions.

Some findings regarding TOEFL listening items can be interpreted as illustrating the effect of background knowledge on comprehension. Nissan et al. (1996) reported that when the language of one of the speakers in a TOEFL dialogue was linked to a specific role the speaker played and the role was not one of a casual acquaintance or classmate, the items associated with such dialogues were significantly more difficult than items without this feature. (A more detailed description of this variable can be found in Appendix A, p. 39 of this report.) The authors hypothesized that such items may be more difficult because the test takers may be unfamiliar with the specific roles enacted in these dialogues. Freedle and Kostin (1999) reported that items associated with TOEFL mini-talks that dealt with academic subject matter such as science or the humanities were more difficult than items associated with mini-talk passages that had nonacademic subject matter. It is possible that differential familiarity with these different topics played a role here, too, in accounting for the relationship to item difficulty.

Nissan et al. (1996) reported an additional finding regarding the relationship between discourse characteristics of the text and item difficulty. They found that the utterance pattern in TOEFL dialogues was significantly related to item difficulty: For TOEFL dialogues composed of two utterances, they found that items associated with dialogues having a statement in the second utterance were significantly more difficult than items associated with dialogues having a question in the second utterance.

Several researchers have studied the effects on listening comprehension of different kinds of redundancy in a text. For second language listeners at lower and intermediate levels of ability, redundancy in a text in the form of repeated nouns seems to be more effective in facilitating listening comprehension than other restatement devices, such as use of synonyms (Chaudron, 1995). On the other hand, in a study by Chiang and Dunkel (1992), elaboration of information, repeating segments of the text, or paraphrasing information only facilitated the comprehension of high listening proficient second language test takers. According to Chiang and Dunkel, the lack of adequate vocabulary prevented the lower-level test takers from taking advantage of the kinds of redundant information used in their study.

### ***Task-processing Factors***

Task-processing factors typically involve an interaction between features of the text and features of the item.

One task-processing factor that has been found to influence listening item difficulty is whether or not an item requires the examinee to make an inference beyond what is explicitly stated in the text. Nissan et al. (1996) reported that TOEFL dialogue items that required an inference (i.e., items that tested implicit information) were significantly more difficult than items that tested comprehension of explicit information.

Lexical overlap between words in the text and words in an item's options has been found to affect listening item difficulty. Freedle and Fellbaum (1987) found that the greater the amount of lexical overlap between words in the correct option and words in a single stimulus sentence (an item type in the TOEFL Listening Section prior to 1995), the easier the item. In their pilot study of TOEIC dialogue items, Buck and Kostin (1999a) similarly found that easier items were characterized by a greater amount of lexical overlap between words in the dialogue and words in the correct option. They further found that if there was a greater degree of lexical overlap between words in the dialogue and words in the incorrect options as compared to the correct option, the item tended to be more difficult.

Studies in the field of reading comprehension have found that information from the most recent clause in a sentence is more accessible than information from an earlier clause (Gernsbacher, 1990). One possible implication in regard to listening stimuli such as dialogues is that the last clause of a dialogue is the one best retained in memory. Consistent with this, Buck

and Kostin (1999a) reported that when the information directly relevant to responding correctly to an item came at the end of a TOEIC dialogue, frequently coinciding with the last clause, the item tended to be easy. Furthermore, if there was lexical overlap between a word in the correct option and a word that came at the end of a TOEIC dialogue, the item also tended to be easy.

## **Method**

### ***Data***

The total sample consisted of 365 TOEFL dialogue items with 1 item per dialogue. Of this total, 240 items came from eight disclosed post-1995 paper-and-pencil TOEFL forms with 30 items per form. The remaining 125 items were selected from 28 disclosed pre-1995 paper-and-pencil TOEFL forms. As there has been an increased emphasis on limiting the content of the dialogues to campus-related matters, these 125 additional items were selected because they included campus-related content.

For the dialogue items employed in this study, the test taker hears a short conversation between two people, each having one turn to speak, which lasts between 5 and 20 seconds. Then a narrator asks a question about what was said. 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. The sections below and the coding manual in Appendix A include several examples of these dialogue items.

In this section and in the sections that follow, the correct option will be referred to as the *key*, and the incorrect options will be referred to as the *distracters*.

### ***Variables Assessing Item Characteristics***

Below is a summary of the variables assessing item characteristics that were included in this study. Detailed descriptions of how these variables were coded are found in the coding manual in Appendix A. The variables include the five significant variables reported by Nissan et al. (1996) as well as other variables identified in the literature review above or by examination by the author of a sample of hard and easy dialogue items.

Several of these variables were coded separately for the first speaker and for the second speaker, as well as for the total dialogue. The reason for the separate coding of the first and the second speaker is that, in 93% of the TOEFL dialogues in this study, the narrator's question only

refers to what the second speaker has said. Because of this, it is hypothesized that test takers will focus more on what the second speaker has said than on what the first speaker has said; as a consequence, characteristics of the second speaker's utterance may be more closely related to item difficulty than are characteristics of the first speaker's utterance. It should be emphasized that although the narrator's question usually focuses on what the second speaker has said, in most cases the test taker must also comprehend what the first speaker has said in order to respond correctly to the item.

*Word-level variables.* Several measures of vocabulary knowledge were employed. First, the measure of vocabulary knowledge included in Nissan et al. (1996), discussed above, was coded. Their measure of difficult vocabulary was the presence of an infrequent vocabulary word in the dialogue; that is, a dialogue was coded as having an infrequent word if it contained a word that was not on a list of 100,000 common words compiled by Berger (1977).

Examination of the items coded for infrequent vocabulary, using the method in Nissan et al. (1996), revealed two types of items:

1. For one type of item, knowledge of the meaning of the infrequent word was relevant to responding correctly to the item: In the example below, knowledge that the infrequent word *almanac* refers to a kind of book is relevant to identifying the key.

(man) Shall I return this almanac to the reference desk?

(woman) I want to check a few dates first.

(narrator) What does the woman mean?

(A) She needs to check her calendar.

(B) She hasn't finished with the book.\*

(C) The reference material is out-of-date.

(D) She has already returned the almanac.

2. For a second type of item, knowledge of the meaning of the infrequent word does not appear to be relevant to responding correctly to the item, as in the example below where knowledge of the meaning of the infrequent word *antique* does not appear to be needed to respond correctly:



(woman) There's a great antique show at the Grant Auditorium. Let's go see it this evening.

(man) I've worked really hard all day long. Won't it be there for a while?

(narrator) What does the man imply?

(A) He has to work late tonight.

(B) He'd rather go at another time.\*

(C) He's already seen the show.

(D) It'll be hard to get to the auditorium on time.

Based on the above distinction, a variant of the Nissan et al. (1996) measure of vocabulary knowledge was also included in the study; for this variant, only those items were coded where knowledge of the meaning of the infrequent vocabulary word was relevant to responding correctly to the item.

The average word length of the words in the dialogue was also used as a measure of vocabulary knowledge; there is evidence that longer words are generally more difficult than shorter words (e.g., Carver, 1976). Average word length was obtained separately for the first speaker's utterance and for the second speaker's utterance, as well as for the total dialogue.

Items were also coded as to whether or not comprehension of an idiom in the dialogue was relevant to responding correctly to the item. *The American Heritage Dictionary* (2000) defines the word *idiom* as "an expression consisting of two or more words having a meaning that cannot be deduced from the meanings of its constituent parts" (p. xxxvi). Comprehending idioms can be difficult because even high-frequency words in the context of an idiom can mean something quite different from what they commonly mean and thus have a meaning that nonnative test takers are unfamiliar with. Simply coding for infrequent words will not pick up this kind of difficulty. An example of a dialogue coded for this variable is given below; in this example, the idiomatic expression *she's got it made*, which is relevant to responding correctly to the item, includes no infrequent words, but the meaning cannot be inferred from the meaning of the individual words.

(man) If you could, would you trade places with your sister?

(woman) Yeah, she's got it made.

- (narrator) What does the woman mean?
- (A) The sisters share a lot of things.
  - (B) She and her sister will switch seats.
  - (C) Things are going well for her sister.\*
  - (D) Her sister finished her cooking.

Another word-level code concerned whether there were instructions to include sandhi-variation in the dialogue. An example of an item that includes such instructions is given below:

- (woman) You know [Y'know], some TV channels have been rerunning a lot of [lotta] comedies from the sixties. What do you think of [thinka] those old shows?
- (man) Not much. But then, the new ones aren't so great either.
- (narrator) What does the man mean?
- (A) He no longer watches much television.
  - (B) He prefers the comedies from the sixties.
  - (C) Television comedies haven't improved since the sixties.\*
  - (D) He hasn't seen many of the old shows.

A reviewer of this report, who is familiar with the creation of TOEFL dialogue items, made the point that “often the speakers [in the dialogue] elide in the delivery, and this would not necessarily be indicated in the script” (Susan Nissan, personal communication, June 5, 2003). However, one would have to listen to the recording of the dialogue in order to code for sandhi-variation that was not indicated in the script. Although coding for sandhi-variation based on the recording of the dialogue is clearly the superior method for assessing this variable, this was not possible here, as will be explained below.

In addition to sandhi-variation, several other phonological variables unique to listening might also contribute to the difficulty of TOEFL dialogue items, such as speech rate, false start, and repetition rate (see Buck & Kostin, 1999b, for a discussion of phonological variables). However, measurement of variables such as these was not possible in the current study for the following reasons: (a) The recording of each item is embedded in a longer recording of the test in

which the item occurs, (b) to collect the recordings of each dialogue and create a master tape would require accessing excerpts from a great number of original recordings, and (c) analyzing such a tape would require expertise, processes, and equipment that were not available for the current study.

A further word-level variable included in the study was whether or not the key contained an infrequent word. Since the key is presented to test takers in printed form, this variable also taps reading comprehension skill; insofar as the construct being assessed by the dialogue items is the ability to comprehend spoken rather than written text, this variable could be considered, in part, to be a measure of one kind of construct-irrelevant variance.

*Sentence-level variables.* Based on Nissan et al.'s (1996) finding, dialogues were coded with regard to whether or not they contained more than one negative; utterances of the first and second speaker were also separately coded for this variable. Other measures of grammatical complexity that were coded separately for the first and second speaker as well as for the total dialogue were: (a) the number of dependent clauses and (b) the number of words in the longest T-unit, the T-unit being defined as an independent clause with any attached dependent clauses (Hatch & Lazaraton, 1994). The dialogues were also coded for the number of each of four different types of referentials.

Another sentence-level variable coded was whether the key was in the form of a suggestion or a directive. Since most of the test takers probably learned English in a classroom setting, where the instructor probably included frequent suggestions and/or directives in the course of lecturing, it is likely that test takers are very familiar with these grammatical forms, which might tend to make items using such forms easier.

*Discourse-level variables.* The dialogues were coded for the four different utterance patterns identified by Nissan et al. (1996): question-question, statement-question, statement-statement, and question-statement. Also, based on Nissan et al., dialogues were coded as to whether or not the language of one of the speakers in the dialogue was linked to a specific role the speaker played and the role was not one of a casual acquaintance or classmate.

Several additional codes concerned the kind of content in the dialogue. For example, a dialogue was coded as having content dealing with the academic part of campus life if it dealt with the following type of topics: registering for classes; students' attitudes toward their course work; references to materials used for class, such as textbooks and calculators; studying;

interactions with professors involving course work; class attendance; academic requirements; exams; course assignments; classroom experience; and similar content.

*Task-processing variables.* Following Nissan et al. (1996), items were coded with regard to whether or not the item required the test taker to make an inference beyond what was explicitly stated in the dialogue.

Several variables assessing lexical overlap between words in the options and words in the dialogue were included. Some assessed the amount of lexical overlap between the words in the key and the words in the dialogue. Other variables in this category compared the amount of lexical overlap in the distracters with the amount of lexical overlap in the key; the expectation is that distracters that have a greater degree of lexical overlap than the key has would be very attractive and would tend to make an item more difficult.

Additional task-processing variables assessed the location of the lexical overlap, such as, for example, whether or not the lexical overlap involved words in the last clause of the dialogue. As noted above, research has shown that information from the most recent clause in a sentence is more accessible than information from an earlier clause. The expectation is that the relationship between lexical overlap and item difficulty would be stronger if the overlap involved words in the last clause of the dialogue than if it involved words coming earlier in the dialogue.

A further task-processing variable concerned whether there were two pieces of information in the dialogue that functioned as substitutes for each other such that each of these components, in isolation, could yield the correct response. This can be thought of as a form of redundant information in the dialogue. For example, in the following item, the second speaker's utterance contains the following two components: "Oh, it's not a problem anymore" and "I've found an ointment that works just fine." Each of these two components, in isolation, could yield the correct response.

(woman) Have you seen the doctor about your skin condition yet?

(man) Oh, it's not a problem anymore. I've found an ointment that works just fine.

(narrator) What does the man imply?

(A) The doctor was too busy to see him.

(B) He doesn't need to see the doctor.\*

(C) The woman should use the ointment.

(D) His skin condition has gotten worse.

Items were also coded as to whether or not test takers could respond correctly to an item solely on the basis of the second speaker's utterance. Items associated with most TOEFL dialogues require the test taker to integrate information from the utterances of the two speakers in order to respond correctly to the item. In contrast, items coded for this variable do not require such integration; comprehension of only the second speaker's utterance suffices to respond correctly. Insofar as most TOEFL dialogue items assess, in part, the ability to integrate information from the utterances of the two speakers, items coded for this variable can be seen as falling short in this regard. The following is an example of an item coded for this variable, where it appears possible to respond correctly to the item if one only comprehends the utterance of the second speaker.

(man) What have you heard about Professor Smith? I'm thinking of taking an advanced engineering course with him.

(woman) You really should. One of his articles just won some sort of award and I heard he's always publishing something in the journals.

(narrator) What does the woman say about the professor?

(A) His classes are very difficult.

(B) His work is well respected.\*

(C) He will publish a book soon.

(D) He is no longer teaching.

An additional code concerned whether there was an apparent inconsistency between an utterance in the dialogue and the item's key. In the dialogue below, for example, there is an apparent inconsistency between the woman's utterance "Then you did get my message" and the key, "Her message did not reach the man." In items such as the following example, comprehension of the narrator's question appears to be essential for responding correctly to the item.

(man) Thanks for letting us know you'd be late for the appointment.

(woman) Oh, good. Then you did get my message.

- (narrator) What had the woman assumed?
- (A) The man had given her the message.
  - (B) The man was late as well.
  - (C) She had plenty of time to make the appointment.
  - (D) Her message did not reach the man.\*

In addition, this code applies to dialogues using sarcasm where there is also an apparent inconsistency between an utterance in the dialogue and the item's key, as in the example below, where there is an apparent inconsistency between the utterance "... another one of Mike's brilliant ideas" and the key, "He [Mike] often makes foolish suggestions."

(man) Can you believe it? Now we're supposed to bring a note from our instructor every single time we want to use the computer!

(woman) [sarcastically] I'll bet that was another one of Mike's brilliant ideas!

(narrator) What does the woman imply about Mike?

- (A) He often makes foolish suggestions.\*
- (B) His instructor won't give him a note.
- (C) He should try using the computer himself.
- (D) He is a very good instructor.

### ***The Coding***

The data analysis is based on the coding of one researcher. A second coder, an ETS staff member who writes and reviews TOEFL dialogues and dialogue items, was recruited to establish intercoder reliability for (a) those variables requiring subjective judgment and (b) the significant variables reported in the Nissan et al. (1996) study of TOEFL dialogue items. Sixty dialogue items from two TOEFL forms were used for this purpose.

For variables that simply code for the presence or absence of a characteristic, the statistic used here to assess intercoder reliability is percent agreement, with an agreement of 90% or more as the desired outcome. Table 1 lists those variables that are simply coded for the presence or absence of a characteristic and the associated percent agreement between the two coders.

**Table 1*****Intercoder Reliability Based on 60 TOEFL Dialogue Items From Two TOEFL Forms***

Variable name	Percent agreement
V01: Infrequent word in dialogue	95%
V02: Knowledge of infrequent word in dialogue is relevant to responding correctly.	92%
V07: Comprehension of idiom in dialogue is relevant to responding correctly.	85%
V11: Two or more negatives in total dialogue	97%
V23: Utterance pattern: question-question	100%
V24: Utterance pattern: statement-question	95%
V25: Utterance pattern: statement-statement	98%
V26: Utterance pattern: question-statement	98%
V27: Speaker has specific role.	100%
V28: Content of dialogue deals with academic campus life.	93%
V29: Content of dialogue deals with nonacademic campus life.	88%
V30: Content of dialogue is related to both campus and a few other domains.	93%
V31: Campus-related terms are present in dialogue but are incidental to main focus.	87%
V32: Content of dialogue is related to noncampus domain.	90%
V45: An inference is required to respond correctly.	92%
V46: More than one element in utterance of second speaker yields key.	90%
V47: Only comprehension of utterance of second speaker is needed to respond correctly.	92%
V49: Key seems inconsistent with content of dialogue.	98%

Using the criterion of percent agreement, the intercoder reliability reaches or exceeds 90% agreement for 15 of the 18 variables in Table 1, and the percent agreement for the remaining variables is close to 90%. Intercoder reliability was also obtained for one of the variables in the study that assessed lexical overlap, namely, for variable V34 (number of words

in key that overlap with words in dialogue); unlike the variables included in Table 1, which were all coded dichotomously (i.e., either 1 or 0), this variable was coded on a continuum, allowing intercoder reliability to be assessed by the Pearson correlation coefficient. (The criteria for judging whether there is lexical overlap between words in the options and words in the dialogue is the same for all variables assessing lexical overlap.) Coding items in the same two forms that were used for coding the variables in Table 1, the correlation between the coding of the first coder and the coding of the second coder for V34 was  $r = .80, p = .000$ , indicating an acceptable level of intercoder reliability for this variable.

*Dependent variable.* The dependent variable in this study is equated delta, a measure of item difficulty (Petersen, Marco, & Stewart, 1982). Higher values are associated with more difficult items and lower values are associated with easier items.

## **Results and Discussion**

Table 2 reports the Pearson correlation coefficients between equated delta and the 49 variables in this study for the data set of 365 TOEFL dialogue items. (Note that all the statistical analyses in this report were carried out using SPSS [Statistical Package for the Social Sciences] software.) In an effort to control for Type I error, the Bonferonni procedure was used to determine the critical probability. Dividing .05 by the number of tests of significance, the critical probability becomes .001. The 11 variables with correlations at this latter level of significance will be discussed below.

The first variable in Table 2 whose  $p$  value is equal to or is less than the critical probability is V02 (knowledge of infrequent word in dialogue is relevant to responding correctly); the correlation indicates that items coded for V02 tend to be more difficult. This variable is a variant of the vocabulary measure used in Nissan et al. (1996), the latter simply coding for the presence of an infrequent word in the dialogue. In contrast to Nissan et al., who reported a significant relationship between this latter vocabulary measure and item difficulty, the corresponding correlation in the current study, where this vocabulary measure is referred to as V01: Infrequent word in dialogue, is not significant. The findings of the current study suggest that it is not the mere presence of a low-frequency word in the dialogue that is associated with item difficulty; rather, the critical factor seems to be whether or not knowledge of the meaning of the infrequent word is relevant to responding correctly to the item. One possible explanation for



the discrepancy between the result in Nissan et al. and the current result is that the Nissan et al. study included more items that required understanding infrequent words than were included in the current study.

**Table 2**

***Correlation of Variables With Item Difficulty (Equated Delta)***

Variable name	Correlation with equated delta	p <sup>a</sup>
<i>Word-level variables</i>		
V01: Infrequent word in dialogue ( <i>N</i> = 132) <sup>b</sup>	.059	.130 <sup>n</sup>
V02: Knowledge of infrequent word in dialogue is relevant to responding correctly. ( <i>N</i> = 52)	.200	.000
V03: Average word length in utterance of first speaker	.084	.109
V04: Average word length in utterance of second speaker	.006	.904
V05: Average word length in total dialogue	.077	.141
V06: Instructions to include sandhi-variation in dialogue ( <i>N</i> = 4)	.124	.017
V07: Comprehension of idiom in dialogue is relevant to responding correctly. ( <i>N</i> = 47)	.245	.000
V08: Infrequent word in key ( <i>N</i> = 9)	.139	.008
<i>Sentence-level variables</i>		
V09: Two or more negatives in utterance of first speaker ( <i>N</i> = 3)	.035	.251 <sup>n</sup>
V10: Two or more negatives in utterance of second speaker ( <i>N</i> = 7)	.125	.008 <sup>n</sup>
V11: Two or more negatives in total dialogue ( <i>N</i> = 31)	.114	.014 <sup>n</sup>
V12: Number of dependent clauses in utterance of first speaker	.064	.225

*(Table continues)*

Table 2 (continued)

Variable name	Correlation with equated delta	p <sup>a</sup>
V13: Number of dependent clauses in utterance of second speaker	.129	.014
V14: Number of dependent clauses in total dialogue	.124	.018
V15: Number of words in longest T-unit of first speaker	.012	.818
V16: Number of words in longest T-unit of second speaker	.085	.104
V17: Number of words in longest T-unit of total dialogue	.049	.347
V18: Number of within clause referentials in dialogue	.122	.020
V19: Number of between clause referentials within a turn in dialogue	.021	.693
V20: Number of referentials in utterance of one speaker that refer to word in utterance of other speaker	.096	.066
V21: Number of special referentials in dialogue	-.055	.292
V22: Number of words in key	.038	.468
<i>Discourse-level variables</i>		
V23: Utterance pattern: question-question ( <i>N</i> = 11)	-.147	.002 <sup>n</sup>
V24: Utterance pattern: statement-question ( <i>N</i> = 41)	-.080	.064 <sup>n</sup>
V25: Utterance pattern: statement-statement ( <i>N</i> = 172)	.104	.024 <sup>n</sup>
V26: Utterance pattern: question-statement ( <i>N</i> = 140)	.003	.483 <sup>n</sup>
V27: Speaker has specific role. ( <i>N</i> = 20)	-.101	n/a <sup>nc</sup>

(Table continues)

Table 2 (continued)

Variable name	Correlation with equated delta	p <sup>a</sup>
V28: Content of dialogue deals with academic campus life. ( <i>N</i> = 125)	.181	.001
V29: Content of dialogue deals with nonacademic campus life. ( <i>N</i> = 30)	.026	.618
V30: Content of dialogue is related to both campus and a few other domains. ( <i>N</i> = 45)	-.069	
V31: Campus-related terms are present but are incidental to main focus of dialogue. ( <i>N</i> = 24)	-.114	.030
V32: Content of dialogue is related to noncampus domain. ( <i>N</i> = 141)	-.087	.098
V33: Total number of words in dialogue	-.018	.732
<i>Task-processing variables</i>		
<i>Lexical overlap variables</i>		
V34: Number of words in key that overlap with words in dialogue	-.149	.004
V35: Percentage of words in key that overlap with words in dialogue	-.180	.001
V36: Key has more words that overlap with dialogue than do three distracters. ( <i>n</i> = 40)	-.135	.010
V37: No distracter has more words than key overlapping with dialogue. ( <i>N</i> = 96)	-.216	.000
V38: The key has no helpful lexical overlap with the dialogue. ( <i>N</i> = 102)	.128	.014
V39: All three distracters have more words than key overlapping with dialogue. ( <i>N</i> = 53)	.107	.040
V40: The key has the last overlapping word with the dialogue. ( <i>N</i> = 73)	-.326	.000

(Table continues)

Table 2 (continued)

Variable name	Correlation with equated delta	p <sup>a</sup>
V41: There is overlap between words in the key and words spoken by second speaker. ( <i>N</i> = 132)	-.206	.000
V42: There is overlap between words in the key and words in last clause of dialogue. ( <i>N</i> = 88)	-.207	.000
V43: Key has synonym of (but no overlapping word with) a word in last clause of dialogue. ( <i>N</i> = 22)	-.084	.111
V44: Overlapping words of all three distracters come later in dialogue. ( <i>N</i> = 55)	.153	.003
<i>Additional task-processing variables</i>		
V45: An inference is required to respond correctly. ( <i>N</i> = 178)	.158	.001 <sup>n</sup>
V46: More than one element in utterance of second speaker yields key. ( <i>N</i> = 27)	-.291	.000
V47: Only comprehension of utterance of second speaker is needed to respond correctly. ( <i>N</i> = 70)	-.163	.002
V48: Key is a suggestion or directive. ( <i>N</i> = 42)	-.161	.002
V49: Key seems inconsistent with content of dialogue. ( <i>N</i> = 7)	.238	.000

<sup>a</sup> The *p* values marked with the superscript <sup>n</sup> are associated with variables that were significant in the Nissan et al. (1996) study. Because there was a clear prediction regarding the direction of the correlation for these variables, the *p* values for them are based on a one-tail test of significance. All other *p* values in the table are based on two-tailed tests of significance. <sup>b</sup> For variables with dichotomous coding (i.e., coded either 1 or 0), the number of items coded for the presence of the variable is given in parentheses after the variable name. <sup>c</sup> The correlation is not in the predicted direction, in which case a one-tailed test is not appropriate.

A second variable meeting the critical probability criterion is V07: Comprehension of idiom in dialogue is relevant to responding correctly; V07 correlates positively with item difficulty. As noted earlier, comprehending idioms can be difficult because even high-frequency

words in the context of an idiom can mean something quite different from what they commonly mean and thus have a meaning that nonnative test takers are unfamiliar with.

The correlation for variable V28 indicates that dialogues dealing with the academic features of campus life are more difficult than dialogues dealing with other subject matter. Some of the more difficult dialogues coded for V28 deal with academic procedures typical of American universities, such as obtaining the required number of credits to graduate, registering for classes, the need for taking basic courses in a subject before taking more advanced courses, and getting a professor's signature to obtain special permission to take a course. It is possible that dialogues with such content are more difficult because nonnative test takers lack background knowledge about these topics.

The correlations of several variables dealing with lexical overlap meet the critical probability criterion. Variable V35 (the percentage of words in the key that overlap with words in the dialogue) was negatively related to item difficulty, indicating that items with a high percentage of lexical overlap in the key tend to be easier items. Similar findings in regard to percentage of lexical overlap in the key have been reported for TOEFL mini-talks (Freedle & Kostin, 1999) and for TOEFL reading (Freedle & Kostin, 1993). One might be concerned that a test taker having little or no comprehension of a dialogue could nevertheless perform well on TOEFL dialogue items by simply choosing the option that had the most lexical overlap with the dialogue. Some information relevant to this concern is provided by results regarding V36 (key has more words overlapping with the dialogue than do any of the three distracters); only 40 of the 365 dialogue items in this study, about 11% of the items, were coded for this variable. Thus, using a strategy of selecting the option with the most lexical overlap would certainly fail to yield a good score on this item type. (Further examination of the TOEFL dialogue items indicates that there is no simple strategy involving lexical overlap that would yield successful performance on these items.)

A further finding suggests that item difficulty is also related to lexical overlap between words in the distracters and words in the dialogue. The correlation for variable V37 indicates that items tend to be easier when no distracter has more words that overlap with the dialogue than does the key. This suggests that if distracters had more lexical overlap with the dialogue as compared to the key, the item would be harder. Supporting this conjecture is the correlation for variable V39,

significant at the less stringent value of  $p = .040$ , which indicates that items tend to be harder when all three distracters have more words overlapping with the dialogue than does the key.

The correlations of some additional variables suggest that item difficulty is also related to the location of the words in the dialogue that overlap with words in the key. In general, the results suggest that the relationship between item difficulty and lexical overlap is strengthened if the lexical overlap involves words coming later in the dialogue. For example, one can consider all instances of lexical overlap between words in the dialogue and words in the options and then identify which of these overlapping words occurs last in the dialogue. The correlation for variable V40 shows that the presence of this “last” overlapping word in the key is negatively related to item difficulty; that is, it is associated with easier items. In a related finding, variable V41, which codes for the presence of lexical overlap between words spoken by the second speaker in the dialogue and words in the key, is also associated with easier items. Likewise, variable V42, which codes for lexical overlap between words in the last clause of the dialogue and words in the key, is also associated with easier items.

The correlation of item difficulty with V45 (an inference is required to respond correctly) also meets the critical probability criterion. As expected, the correlation indicates that items that require the test takers to make an inference beyond what is explicitly stated in the dialogue tend to be more difficult than items that do not require this.

Also meeting the critical probability criterion is the correlation between item difficulty and variable V46, which coded items with respect to whether or not there were two components, (i.e., clauses, phrases, exclamations, or a combination of these) uttered by the second speaker in the dialogue such that each of these components, independent of the other, could yield the key. The presence of this variable was negatively associated with item difficulty (i.e., associated with easier items). The presence of two such components in the dialogue is a kind of redundancy; other kinds of redundancy have been found to facilitate listening comprehension in past research (see Chaudron, 1995; Chiang & Dunkel, 1992).

The last correlation meeting the critical probability criterion is between item difficulty and variable V49, which coded for whether or not there was an apparent inconsistency between the text of the dialogue and the key. The correlation for variable V49 indicates that items coded for this variable tend to be more difficult.

In Table 2, the variables in this study are grouped into four broad categories: word-level variables, sentence-level variables, discourse-level variables, and task-processing variables. The 11 variables discussed above, whose correlation with item difficulty met the critical probability criterion, include representatives from three of these four broad categories, with 2 belonging in the category of word-level variables, 1 in the category of discourse-level variables, and 8 in the category of task-processing variables. Also, some of these 11 variables were discussed in the literature review above. For those variables, the direction of their correlation with item difficulty was consistent with the findings covered in the literature review.

*Regarding the magnitude of the correlations.* Although statistically significant, the correlations between the 11 variables described above and item difficulty are generally small in magnitude: Only 1 exceeds a magnitude of .30, an additional 7 fall between .20 and .30, with the remaining 3 falling below .20. These results are similar to results obtained in an earlier study exploring the relationship between item characteristics and the difficulty of TOEFL mini-talk items (see Freedle & Kostin, 1999). Freedle and Kostin's (1999) comments below regarding the small magnitudes of the significant correlations in the TOEFL mini-talk study can be seen as applying to the present results as well:

Regarding these small magnitudes, it is interesting that a parallel-processing model of language comprehension such as that proposed by Just and Carpenter (1987, pp. 279-281) is consistent with such an observation. That is, if many processes influence comprehension, and if they do operate in parallel, then no single variable is likely to dominate the comprehension process. This fact implies that the correlation of any single variable with a measure of comprehension should be small in magnitude. (The reader should note that if future studies should find large correlations between item difficulty and other variables, this may only mean that the idea of massive parallel processing might be called into question.)  
(p. 19)

The fact that a similar pattern of correlations has been observed for TOEFL dialogues as well as for TOEFL mini- talks can be seen as lending support to the interpretation of both sets of results in terms of a parallel-processing model of language comprehension.

*Results regarding the significant variables in Nissan et al. (1996).* The first variable reported as significant in Nissan et al. was *infrequent vocabulary*, which was measured by the presence of an infrequent word in the dialogue. In the current study, as noted above, this variable, V01, did not have a significant correlation with item difficulty (i.e.,  $r = .059, p = .130$ ). However, a variant of this variable, V02 (knowledge of infrequent word is relevant to responding correctly), did correlate significantly with item difficulty (i.e.,  $r = .211, p = .000$ ). As noted earlier, one possible reason that might account for Nissan et al.'s significant finding and the corresponding nonsignificant one in this study is that the dialogues in Nissan et al.'s study had a much higher percentage of infrequent words that were relevant to responding correctly than was the case in this study.

The second significant variable discussed in Nissan et al. (1996) was *utterance pattern*; items with a statement in the second utterance (i.e., statement-statement and question-statement patterns) were found to be significantly more difficult than those with a question in the second utterance (i.e., question-question and statement-question patterns). There were not enough items in the Nissan et al. study to examine separately the two patterns that had a statement for the second utterance or the two patterns that had a question for the second utterance. These separate patterns were included in the current study. Of the two patterns with a question in the second utterance, the results here suggest that the question-question pattern, V23, is more closely (and negatively) related to item difficulty than the statement-question pattern, V24 ( $r = -.147, p = .002$  and  $r = -.080, p = .064$ , respectively). Of the two patterns with a statement in the second utterance, the results here suggest that the statement-statement pattern, V25, is more closely (and positively) related to item difficulty than the question-statement pattern, V26 ( $r = .104, p = .024$  and  $r = .003, p = .479$ , respectively). In general, the results here replicate the results in Nissan et al. regarding utterance pattern and provide additional information regarding the contribution of the components making up the patterns.

The third significant variable in Nissan et al. (1996) was *negative in stimulus*; items associated with dialogues that had two or more negatives were found to be significantly more difficult than those that had fewer negatives. Consistent with this result, in the current study the correlation between item difficulty and variable V11 (two or more negatives in the dialogue) is in the expected direction ( $r = .114$ ) and is significant at the level of  $p = .014$ . The results also suggest that the presence of negatives in the utterance of the second speaker may play a greater



role in accounting for this result than the presence of negatives in the utterance of the first speaker: The correlation between item difficulty and V09 (two or more negatives in utterance of first speaker) is  $r = .035$ ,  $p = .251$ , while the correlation between item difficulty and V10 (two or more negatives in utterance of second speaker) is  $r = .125$ ,  $p = .008$ .

The fourth significant variable reported in Nissan et al. (1996) is *implicit versus explicit information tested*. For this variable, items are coded with regard to whether an inference is needed to respond correctly to the item. As noted above, the correlation in the current study for this variable, V45, met the critical probability criterion ( $r = .158$ ,  $p = .001$ ).

The last variable reported as significant in Nissan et al. (1996) was *role of speaker(s)*; items where the language of one of the speakers in the dialogue was linked to a specific role the speaker played and the role was not one of a casual acquaintance or classmate were found to be more difficult than items not having this characteristic. In the current study, the correlation between item difficulty and this variable, V27, was not significant and also was in a direction opposite to prediction. One possible explanation for the discrepancy between the two studies is that the specific roles in the current study may have been more familiar to the test takers than were the roles in Nissan et al. Examples of some specific roles in the current study associated with easier dialogue items are: server at a restaurant, manager at a supermarket or grocery store, and sales person at a store selling luggage. It seems likely that nonnative test takers have some background knowledge concerning roles such as these and can use this knowledge to aid in comprehending the dialogues that include these roles.

*Regression analyses.* Multiple regression was used to estimate how much variance in item difficulty is accounted for by the 49 variables employed in this study. In the regression analysis, equated delta was the dependent variable and the 49 variables in Table 2 were entered as a set. The overall  $F(47, 317) = 6.369$ ,  $p = .000$ ; the multiple  $r = .697$  with an adjusted  $R^2$  of .409, suggesting that about 41% of the variance is accounted for by the variables in the study.

Stepwise regression was used to identify a more parsimonious subset of variables to predict item difficulty. As noted above, the statistical analyses in this report were carried out using SPSS software. The stepwise regression procedure used by this software, as described in the SPSS manual (SPSS, 1999), employs the forward selection procedure to start the process; that is, variables are entered into the model one by one. The variable with the strongest positive (or negative) simple correlation with the dependent variable is entered first. At subsequent steps,

the variable with the strongest partial correlation is entered and tested for significance. However, the stepwise selection procedure tests variables already in the model for removal at each step. (For additional information concerning these procedures, see SPSS, 1999, p.216.)

All 49 variables listed in Table 2 were available for possible selection. Each new variable that was admitted into the solution had to yield a significance level of  $p \leq .05$ . In the final regression equation, 14 variables were left. Results are given in Table 3. In carrying out the stepwise regression, no “already entered variables” needed to be removed from the model because their significance level no longer met the established criterion. We see that the 14 variables accounted for about 40% of the variance with an  $F(14, 350) = 18.15, p = .000$ .

The correlations of item difficulty with all but one of these 14 variables were significant at  $p < .05$  (see Table 2), the one exception being V43. Some of these 14 significant variables were discussed in the literature review above. For such variables, the direction of their beta weights is consistent with the findings covered in the literature review.

It is important to note here that the above estimate of variance accounted for by the 14 variables capitalizes to a considerable degree on chance. A jackknife procedure was used to estimate how much the variance accounted for would vary when using data sets that differ from the original 365-item data set. The jackknife procedure was carried out as follows: First, 10 samples of approximately equal size and approximately equal difficulty were created from the original 365 item data set. Next, a regression procedure was run 10 times; for each run, the 14 variables were used to predict the item difficulty of a data set comprising 9 of the 10 samples, with a different set of 9 samples used for each run. The resulting equation was then used to predict the item difficulty values in the 10th sample. The predicted difficulty values were then correlated with the observed difficulty values in this 10th sample, with the resulting  $R^2$  forming a basis for estimating variance accounted for.

The results of the jackknife procedure are as follows: The correlations between predicted and observed item difficulty in the 10 runs range from  $.517, p < .001$  to  $.742, p < .000$ , with a mean correlation of  $.610, p = .000$ ; thus, the variance accounted for ranges from 26.7% to 55.1%, with a mean of 37.2%. These latter figures can be seen as estimates of variance accounted for when the 14 variables that emerged in the original stepwise regression are used to predict the difficulty of a set of TOEFL dialogue items that differs from the original set of 365 items.

**Table 3**

*Results of Stepwise Multiple Regression, With Only Significant Variables Remaining in the Equation*

	B	Std. Error	Beta	t-test	Prob.
<i>Constant</i>	10.461	.119		87.661	.000
V40: Key has last overlapping word with the dialogue.	-.750	.158	-.214	-4.757	.000
V46: More than one element in utterance of second speaker yields key.	-1.167	.225	-.218	-5.182	.000
V49: Key seems inconsistent with content of dialogue.	1.895	.422	.186	4.493	.000
V07: Comprehension of idiom in dialogue is relevant to responding correctly.	.927	.174	.222	5.332	.000
V02: Comprehension of infrequent word in dialogue is relevant to responding correctly.	.667	.166	.167	4.011	.000
V11: Two or more negatives in total dialogue	.632	.208	.126	3.045	.003
V14: Total number of dependent clauses in dialogue	.157	.046	.141	3.402	.001
V43: Key has synonym of a word in last clause of dialogue.	-.749	.243	-.127	-3.078	.002
V08: Infrequent word is in key.	1.017	.374	.113	2.721	.007
V48: Key is a suggestion or directive.	-.562	.183	-.128	-3.069	.002
V47: Only comprehension of utterance of second speaker is needed to respond correctly.	-.477	.148	-.134	-3.224	.001
V28: Content of dialogue deals with academic campus life.	.329	.122	.111	2.687	.008
V37: No distracter has more lexical overlap with dialogue than key.	-.345	.143	-.109	-2.421	.016
V18: Number of within-clause referentials in dialogue	.618	.273	.093	2.261	.024

*Note.* Multiple  $R = .649$ ;  $R^2 = .421$ ; Adjusted  $R^2 = .398$ ; standard error of estimate = 1.088.

## Conclusions and Implications

First of all, this study has replicated some of the significant findings in Nissan et al. (1996). The following variables that were significant in Nissan et al. were also significantly related to item difficulty in the current study: (a) the presence of two or more negatives in the dialogue, (b) the need to draw an inference beyond what is explicitly stated in the dialogue, and (c) the pattern of utterances in the dialogue. One can have confidence in these results not only because they have been replicated but also because the intercoder reliabilities for them are acceptable. However, these results are based on existing items; it still needs to be determined whether they can provide the basis for creating and/or for modifying items to desired levels of difficulty.

In regard to modifying items, one could follow the approach of Adams, Carson, and Cureton (1993), who revised middle-difficulty GRE<sup>®</sup> discrete items in order to produce items of higher or lower difficulty; in the case of TOEFL dialogue items, for example, one could insert two or more negatives into existing dialogues of middle difficulty that have no negatives and see whether this modification increased the difficulty of the item. However, Adams et al. only needed to change some words in a printed test form to modify these GRE items, which led them to conclude that “producing harder analogies and antonyms by revising items in this manner would be a cost-effective procedure” (see Abstract). In contrast, adding negatives to an existing TOEFL dialogue would require re-recording the dialogue, which might mean that such a procedure would not be cost-effective. Consequently, these results might best be used only as a basis for creating new items of varying levels of difficulty. However, assuming that one has a well-replicated set of variables that predict TOEFL dialogue item difficulty, a reviewer of this report has suggested that “the process of recording dialogues for this item type could be planned in such a way as to prerecord all the variations that would be relevant for later construction [of] sets of appropriate difficulty” (I. Bejar, personal communication, December 30, 2002 ). Also, if the significant findings regarding lexical overlap variables are replicated, these findings could be used as a basis for modifying existing items without the need for re-recording the dialogues. In the case of lexical overlap variables, it would be possible to modify the degree of lexical overlap between the options and the dialogue by simply changing some of the words in the options, which are in printed form.

The correlations between item difficulty and a number of variables other than those from Nissan et al. (1996) met the critical probability criterion. At present, these findings are suitable primarily for hypotheses generation, since they still need to be replicated. However, it is appropriate to note that several of these variables did not come simply from an examination of the items themselves, but also from a survey of the research literature. The direction with which these variables correlated with item difficulty is, in all cases, consistent with the findings in the research literature. This provides evidence to suggest that the results regarding some of these variables will be successfully replicated.

### ***Future Studies***

The primary purpose of the current study was a practical one, that is, to provide test development staff with information that has the potential to help them create harder and/or easier TOEFL dialogue items. However, ideally, future studies that investigate the relationship between item characteristics and item difficulty will be more theoretically guided than the present one; the empirical results of these studies will, hopefully, also yield information about the predictive power of different theoretical orientations. Also, future studies, ideally, will attempt to confirm these predictions using methods other than the regression methods used here.

It has been noted above that the correlational results in the present study are consistent with the findings in the research literature. One can hope that it would be possible in the near future to integrate these separate findings into a more comprehensive theoretical approach to language processing.

## References

- Abrahamsen, E., & Shelton, K. (1989). Reading comprehension in adolescents with learning disabilities: Semantic and syntactic effects. *Journal of Learning Disabilities*, 22, 569-572.
- Adams, R., Carson, J., & Cureton, K. (1993). *Item difficulty adjustment study: GRE verbal discretets*. (ETS RR-92-79). Princeton, NJ: ETS.
- American Heritage Dictionary of the English Language* (4<sup>th</sup> ed.). (2000). Boston: Houghton Mifflin Co.
- Anderson, A., & Lynch, T. (1988). *Listening*. New York: Oxford University Press.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL listening framework: A working paper*. (ETS RM-00-07). Princeton, NJ: ETS.
- Berger, K. W. (1977). *The most common 100,000 words used in conversations*. Kent, OH: Herald Publishing House.
- Breland, H., & Jenkins, L. (1997). *English word frequency statistics: Analysis of a selected corpus of 14 million tokens*. New York: College Entrance Examination Board.
- Buck, G., & Kostin, I. (1999a). *Exploring the cause of item difficulty on TOEFL CBT dialogue items*. Manuscript in preparation.
- Buck, G., & Kostin, I. (1999b). *Developing a scheme to analyze the phonological characteristics of listening-item stimuli*. Manuscript in preparation.
- Carver, R. (1976). Word length, prose difficulty, and reading rate. *Journal of Reading Behavior*, 8, 193-203.
- Chaudron, C. (1995). Academic listening. In D. Mendelsohn and J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 74-96). San Diego, CA: Dominic Press, Inc.
- Chiang, C., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26, 345-374.
- Crystal, D. (1980). *A first dictionary of linguistics and phonetics*. Boulder, CO: Westview Press.
- Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp.162-192). Norwood, NJ: Ablex.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10, 133-170.

- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*, 2-32.
- Gernsbacher, M. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.
- Hale, G. (1988). *The interaction of student major-field group and text content in TOEFL reading comprehension*. (TOEFL Research Rep. No. 25). Princeton, NJ: ETS.
- Hatch, E., & Lazaraton, A. (1994). *The research manual—Design and statistics for applied linguistics*. Boston: Heinle & Heinle.
- Henrichson, L. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning, 34*, 103-126.
- Just, M., & Carpenter, P. (1987). *The psychology of reading and language comprehension*. Boston, MA: Allyn & Bacon.
- Kelly, P. (1991). Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching, 29*, 135-150.
- Kintsch, W., & Kozminsky, E. (1977). Summarizing stories after reading and listening. *Journal of Educational Psychology, 69*, 491-499.
- Kintsch, W., Kozminsky, E., Streby, W., McKoon, G., & Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14*, 196-214.
- McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.
- Miller, G.A. (1999). On knowing a word. *Annual Review of Psychology, 50*, 1-19.
- Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. (TOEFL Research Rep. No. 51). Princeton, NJ: ETS.
- Petersen, N., Marco, G., & Stewart, E. (1982). A test of the adequacy of linear score equating models. In Holland, P. & Rubin, D. *Test Equating* (pp. 71-136). New York: Academic Press.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London, England: Longman.

- Recht, D., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, *80*, 16-20.
- Rost, M. (1990). *Listening in language learning*. New York: Longman.
- Smiley, S., Oakley, D., Campione, J., & Brown, A. (1977). Recall of thematically relevant material by adolescent good and poor readers as a function of written versus oral presentation. *Journal of Educational Psychology*, *69*, 381-387.
- Spilich, G., Vesonder, G., Chiesi, H., & Voss, J. (1979). Text processing of domain related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 275-290.
- SPSS, Inc. (1999). *SPSS base 9.0: Application guide*. Chicago: SPSS, Inc.
- Sticht, T., & James, J.H. (1984). Listening and reading. In P.D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 293-317). NY: Longman.
- Yi'an, W. (1998). What do tests of listening comprehension test?—A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*, 21-44.



## Appendix A

### Coding Instructions

#### *Word-level Codes*

##### ***V01: Infrequent Word in Dialogue***

A word in the dialogue is considered to be an infrequent word if it does not appear in *The Most Common 100,000 Words Used in Conversations*, by Kenneth Berger (1977).

*Coding instructions for V01.* If there is at least one content word in the dialogue that does not appear in Berger's word-frequency list, code 1; else 0.

Additional coding instructions for V01:

1. Words with the same root but with different endings are considered to be the *same word* (e.g., the word *offering* in a dialogue would get coded 0 if the word *offered* appeared on Berger's list but the word *offering* did not, since both words have the same root).
2. A compound word in a dialogue would get coded 0 if (a) its component words appeared in Berger's list and (b) the meaning of the compound word could be inferred from its components (e.g., the word *weekday* would get coded 0 because both *week* and *day* appear on Berger's list.)
3. To help in coding V02 below, coders should look up all the words in the dialogue that they believe might not appear in Berger's word-frequency list and make note of all those words that don't appear on the list.

##### ***V02: Knowledge of An Infrequent Word in the Dialogue Is Relevant to Responding Correctly to the Item.***

*Note.* This variable is only coded for those items assigned a 1 for code V01 (infrequent word in dialogue).

1. Below is an example of an item where knowledge of the infrequent word *almanac* is relevant to responding correctly to the item:

(man) Shall I return this almanac to the reference desk?

(woman) I want to check a few dates first.

(narrator) What does the woman mean?

(A) She needs to check her calendar.

(B) She hasn't finished with the book.\*

(C) The reference material is out-of-date.

(D) She has already returned the almanac.

2. Below is an example of an item where knowledge of the infrequent word *antique* is NOT needed in order to respond correctly to the item:

(woman) There's a great antique show at the Grant Auditorium. Let's go see it this evening.

(man) I've worked really hard all day long. Won't it be there for a while?

(narrator) What does the man imply?

(A) He has to work late tonight.

(B) He'd rather go at another time.\*

(C) He's already seen the show.

(D) It'll be hard to get to the auditorium on time.

*Coding instructions for V02.* If knowledge of an infrequent word in the dialogue is relevant to responding correctly to the item *AND* if the infrequent word does not also appear in the key, code 1; else 0. (It is assumed here that knowledge of the infrequent word in the dialogue may not be needed when the infrequent word is also present in the key because, in the latter case, a simple matching strategy might yield the key.)

### ***V03: Average Word Length in the Utterance of the First Speaker***

*Coding instructions.* Use grammar tool in MS-Word to get the average word length in characters of the utterance for the first speaker.

### ***V04: Average Word Length in the Utterance of the Second Speaker***

*Coding instructions.* Use grammar tool in MS-Word to get the average word length in characters of the utterance for the second speaker.

***V05: Average Word Length in Total Dialogue***

*Coding instructions.* Use grammar tool in MS-Word to get the average word length in characters of the utterance for the total dialogue.

***V6: Instructions to Include Sandhi-variation in the Dialogue***

Below is an example of an item that includes instructions to include sandhi-variation in the dialogue:

(woman) You know [Y'know], some TV channels have been rerunning a lot of [lotta] comedies from the sixties. What do you think of [thinka] those old shows?

(man) Not much. But then, the new ones aren't so great either.

(narrator) What does the man mean? (12 seconds)

(A) He no longer watches much television.

(B) He prefers the comedies from the sixties.

(C) Television comedies haven't improved since the sixties.\*

(D) He hasn't seen many of the old shows.

*Coding instructions.* If the speakers in the dialogue are instructed to alter the pronunciation of the words that they speak, code 1; else 0.

***V07: Comprehension of an Idiom Or an Idiomatic Multiword Verb Is Relevant to Responding Correctly to the Item.***

*The American Heritage Dictionary* (2000) defines the word *idiom* as “an expression consisting of two or more words having a meaning that cannot be deduced from the meanings of its constituent parts” (p.xxxvi). Similarly, according to Quirk, Greenbaum, Leech, and Svartvik (1985), idiomatic multiword verbs are those whose “meaning is not predictable from the meanings of its parts” (p. 1162). Some examples of idiomatic multiword verbs given by Quirk et al. are: *come by* (acquire), *turn up* (make an appearance), *give in* (surrender), *catch on* (understand), and *blow up* (explode).

1. Below is an example of an item where comprehension of the idiom *she's got it made* is relevant to responding correctly to the item:

(man) If you could, would you trade places with your sister?

(woman) Yeah, she's got it made.

(narrator) What does the woman mean?

(A) The sisters share a lot of things.

(B) She and her sister will switch seats.

(C) Things are going well for her sister.\*

(D) Her sister finished her cooking.

2. Below is an example of an item where comprehension of the multiword idiomatic verb *turned down* is relevant to responding correctly to the item:

(woman) But David, you mean you didn't apply for a scholarship?

(man) I did, but I was turned down.

(narrator) What does David mean?

(A) He decided to quit school this term.

(B) He didn't bring his application form.

(C) He made a wrong turn downtown.

(D) He didn't receive financial aid.\*

3. Below is an example of an item where comprehension of the idiom *gets on my nerves* does NOT appear to be needed in order to respond correctly to the item:

(man) Why did you come to the meeting late? I left a message with your roommate about the time change.

(woman) She has a very short memory, and it really gets on my nerves sometimes.

(narrator) What does the woman imply?

(A) The man shouldn't have invited her roommate to the meeting.

(B) Her roommate was unable to attend the meeting.

(C) Her roommate is unreliable about delivering messages.\*

(D) She forgot about the time change.

*Coding instructions for V07.* If comprehension of an idiom or multiword idiomatic verb is relevant to responding correctly, code 1; else 0.

***V08: There Is an Infrequent Word in the Key.***

*Coding instructions for V08.* If a word in the key has an Standard Frequency Index (SFI) of less than 40.0 in the Breland word-frequency count (Breland & Jenkins, 1997) AND if this word does not also appear in the dialogue, code 1; else 0. (It is assumed here that comprehension of the infrequent word in the key may not be needed if the infrequent word is also present in the dialogue because, in the latter case, a simple matching strategy might yield the key.)

***Sentence-level Codes***

***V09: Two or More Negatives in Utterance of First Speaker***

Negative markers (e.g., *no* and *not*) are counted, as well as negative prefixes (e.g., *un-* and *dis-*). Negative tags are also counted, even if their meaning is not negative.

*Coding instructions for V09.* If the number of negatives in the utterance of the first speaker is 2 or greater, code 1; else 0.

***V10: Two or More Negatives in Utterance of Second Speaker***

Negative markers (e.g., *no* and *not*) are counted, as well as negative prefixes (e.g., *un-* and *dis-*). Negative tags are also counted, even if their meaning is not negative.

*Coding instructions for V10.* If the number of negatives in the utterance of the second speaker is 2 or greater, code 1; else 0.

***V11: Two or More Negatives in Total Dialogue***

Negative markers (e.g., *no* and *not*) are counted, as well as negative prefixes (e.g., *un-* and *dis-*). Negative tags are also counted, even if their meaning is not negative.

*Coding instructions for V11.* If the number of negatives in the total dialogue is 2 or greater, code 1; else 0.

***V12: Number of Dependent Clauses in Utterance of First Speaker***

*Coding instructions for V12.* Code the number of dependent clauses in the utterance of the first speaker.

***V13: Number of Dependent Clauses in Utterance of Second Speaker***

*Coding instructions for V13.* Code the number of dependent clauses in the utterance of the second speaker.

***V14: Number of Dependent Clauses in Total Dialogue***

*Coding instructions for V14.* Code the number of dependent clauses in the total dialogue.

***V15: Number of Words in Longest T-unit in Utterance of First Speaker***

A T-unit is defined as an independent clause with any attached dependent clauses (Hatch & Lazaraton, 1994).

*Coding instructions for V15.* Code the number of words in the longest T-unit in the utterance of the first speaker.

***V16: Number of Words in Longest T-unit in Utterance of Second Speaker***

*Coding instructions for V16.* Code the number of words in the longest T-unit in the utterance of the second speaker.

***V17: Number of Words in Longest T-unit of Total Dialogue***

*Coding instructions for V17.* Code the number of words in the longest T-unit in the total dialogue.

***V18: Number of Within-clause Referentials in the Dialogue***

The line of dialogue below contains the within-clause referential *his*.

(man) Roy wouldn't let me borrow his notes, even though I needed them.

*Coding instructions for V18.* Code the number of within-clause referentials in the dialogue.

***V19: Number of Between-clause Referentials Within a Speaker's Turn in the Dialogue***

The line of dialogue below contains the between-clause referential *he*.

(man) Julia asked me to pick up the guest speaker, Bob Russell, at the airport this afternoon. Do you know what he looks like?

*Coding instructions for V19.* Code the number of between-clause referentials within a speaker's turn in the dialogue.

***V20: Number of Referentials in the Utterance of One Speaker That Refer to a Word in the Utterance of the Other Speaker***

In the dialogue below, the pronoun *they*, spoken by the man, refers to the word *packages*, spoken by the woman.

(woman) Those packages took forever to arrive.

(man) But they did arrive, didn't they?

*Coding instructions for V20.* Code the number of referentials used by one speaker that refer to a word in the utterance of the other speaker.

***V21: Number of Special Referentials in the Dialogue***

Special referentials are those that refer to things outside of the text. In the example below, the pronouns *you* and *I* refer to the speakers themselves rather than to words in the dialogue.

(woman) Do you have change for a fifty-dollar bill?

(man) A fifty-dollar bill! I hardly have fifty cents!

*Coding instructions for V21.* Code the number of special referentials in the dialogue.

***V22: Number of Words in the Key***

*Coding instructions for V22.* Code the number of words in the key.

**Discourse-level Codes**

***Variables V23-V26***

Each item needs to be coded for one of the following four variables having to do with utterance patterns.

***V23: Utterance Pattern: Question-question***

*Coding instructions for V23.* If the utterance pattern takes the form of question-question, code 1; else 0.

***V24: Utterance Pattern: Statement-question***

*Coding instructions for V24.* If the utterance pattern takes the form of statement-question, code 1; else 0.

***V25: Utterance Pattern: Statement-statement***

*Coding instructions for V25.* If the utterance pattern takes the form of statement-statement, code 1; else 0.

***V26: Utterance Pattern: Question-statement***

*Coding instructions for V26:* If the utterance pattern takes the form of question-statement, code 1; else 0.

***Additional Coding Instructions for V23-V26***

If an utterance includes two sentences, one a question and another a statement, the item's key needs to be examined to determine whether the focus is on the question or on the statement. For example, in the dialogue below, the woman both asks a question and makes a statement. The woman's response is coded as a statement because the key focuses on the statement part of her response.

(man) All I can turn in today is my chemistry homework.

(woman) Is everything all right? You usually have everything completed on time.

(narrator) What does the woman imply about the man?

(A) He usually turns in his assignments late.

(B) He didn't have time to complete everything.

(C) He is usually a conscientious student.\*

(D) He usually completes only his chemistry work on time.

***V27: The Speaker Has a Specific Role.***

For variable V27, use the following instructions from Nissan et al. (1996): judge 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 (pp. 9-10).

*Coding instructions for V27.* If the language of one of the speakers is linked to a specific role the speaker plays and the role is not that of a casual acquaintance or classmate, code 1; else 0.

### **Variables V28-V32**

Each item needs to be coded for one of the next five variables; these concern the content of the dialogues with regard to if and/or how the content is related to campus life.

#### ***V28: The Content of the Dialogue Deals With the Academic Part of Campus Life.***

The content of the dialogue is related to university academic activities. This includes content such as registering for classes; students’ attitudes toward their course work; references to materials used for class such as textbooks, calculators, and the like; studying; interactions with professors involving course work; class attendance; academic requirements; exams; homework; course assignments; classroom experience; and similar content. One example is given below:

(man) All I can turn in today is my chemistry homework.

(woman) Is everything all right? You usually have everything completed on time.

(narrator) What does the woman imply about the man?

- (A) He usually turns in his assignments late.
- (B) He didn't have time to complete everything.
- (C) He is usually a conscientious student.\*
- (D) He usually completes only his chemistry work on time.

*Coding instructions for V28.* If the content of the dialogue is related to university academic activities, code 1; else 0.

***V29: The Content of the Dialogue Deals With the Nonacademic Part of Campus Life.***

This includes nonacademic features such as references to life in a dormitory, student government, discounts for students, extracurricular activities, getting transportation to school, finding a place to live while at school, jobs on campus, and similar content. The following is an example:

(woman) You know, the noise in my dorm has really gotten out of control. My roommate and I can rarely get to sleep before midnight.

(man) Why don't you take the problem up with the dorm supervisor?

(narrator) What does the man suggest the woman do?

- (A) Discuss the situation with the person in charge of the dormitory.\*
- (B) Ask her roommate not to make so much noise.
- (C) Go to bed after midnight.
- (D) Send a letter to the residents.

*Coding instructions for V29.* If the content of the dialogue is related to nonacademic features of campus life, code 1; else 0.

***V30: The Content of the Dialogue Is Related to Campus Life But Could Also Be Related to One or Two Additional Domains.***

This includes references to content such as the following, where it is not clear whether the context is campus, recreation, or work related: working on a project, gyms, cafeterias, roommates, books, presentations, health clinic, library, references to equipment such as computers and photocopy machines, and similar content. In the example given below, the *three projects* could be conducted either at a university or in a work-related setting.

(woman) I'm getting really stressed out. I just don't have the time to work on all three projects.

(man) You need to set priorities—just take the time to figure out what has to be done first.

(narrator) What does the man suggest the woman do?

(A) Calculate how much each project will cost.

(B) Take time to relax.

(C) Discuss her stress with the project leader.

(D) Decide which project is most urgent.\*

*Coding instructions for V30.* If the content of the dialogue is related to campus life but could also be related to one or two additional domains because the context is not specified, code 1; else 0.

***V31: Campus-Related Terms Are Present But Are Incidental to the Main Focus of the Dialogue.***

One example is given below:

(man) You know, I've been watering my plants regularly, but they're still not doing well in my new dorm room.

(woman) Maybe instead of keeping them in the corner you should put them directly in front of the window.

(narrator) What does the woman imply?

(A) The plants may need more light.\*

(B) The plants should get less water.

(C) The area in front of the window is too cold for plants.

(D) Plants rarely do well in the dormitory.

*Coding instructions for V31.* If campus-related terminology is present but is incidental to the main focus of the dialogue, code 1; else 0.

***V32: The Content of the Dialogue Is Either Related to a Noncampus Domain Or Is Very General.***

Two examples are given below:

1. The content of the dialogue below is related to the noncampus domain of shopping.

(woman) I thought the department store was open late from Tuesday through Friday night.

(man) No, just Thursdays and Fridays.

(narrator) On what nights is the store open late?

(A) Thursdays and Fridays.\*

(B) Tuesdays and Fridays.

(C) Wednesdays and Thursdays.

(D) Tuesdays, Thursdays, and Fridays.

2. The content of the dialogue below is very general and could occur in a great variety of settings.

(man) You know, every time I talk to Mary I get the feeling she's being critical of me.

(woman) Don't you think you're overreacting a bit?

(narrator) What does the woman mean?

(A) She thinks Mary is too critical.

(B) She doesn't know how to react.

(C) She thinks the man is too sensitive.\*

(D) She wants to know what the man thinks.

*Coding instructions for V32.* If the content of the dialogue is either very general or clearly related to a noncampus domain, code 1; else 0.

***V33: Total Number of Words in the Dialogue***

*Coding instructions for V33.* Code the total number of words in the dialogue.

## Task-processing Codes

### *Codes Involving Lexical Overlap*

#### ***V34: Number Of Words in the Key That Overlap With Words in the Dialogue***

*Coding instructions for V34.* Using the instructions for coding lexical overlap given in Appendix B, code the number of words in the key that overlap with words in the dialogue.

Note that most of the words in the key that overlap with words in the dialogue are content words; however, in certain cases, lexical overlap is also coded for function words as described in Appendix B.

#### ***V35: Percentage of Words in the Key That Overlap With Words in the Dialogue***

*Coding instructions for V35.* Divide the number of words coded for variable V34 by the number of words coded for variable V22.

#### ***V36: The Key Has More Words That Overlap With Words in the Dialogue Than Do Any of the Three Distracters.***

*Coding instructions for V36.* If the key has more words that overlap with words in the dialogue than do any of the three distracters, code 1; else 0.

#### ***V37: No Distracter Has More Words Than the Key That Overlap With Words in the Dialogue.***

*Coding instructions for V37.* If no distracter has more words that overlap with words in the dialogue than does the key, code 1; else 0. Note that all items assigned a 1 for V36 should also be assigned a 1 for V37.

#### ***V38: The Key Has No Helpful Lexical Overlap With the Dialogue.***

*Coding instructions for V38.* If the key has no words that overlap with words in the dialogue *OR* if the key has lexical overlap with the dialogue that is identical to the lexical overlap of all three distracters, code 1; else 0.

#### ***V39: All Three Distracters Have More Words Than Key That Overlap With Words In the Dialogue.***

*Coding instructions for V39.* If all three distracters have more words that overlap with words in the dialogue than does the key, code 1; else 0.

***V40: The Key Has the Last Overlapping Word With the Dialogue.***

*Coding instructions for V40.* A 1 is assigned for this code if (a) only the key has the last overlapping word with the dialogue, *OR* (b) the key and only one distracter have the last overlapping word with the dialogue, but the key's other overlapping words come later than those of this one distracter, *OR* (c) the key and only one distracter have the last overlapping word but are otherwise equal in regard to lexically overlapping words; else 0.

In the example below, only the key has the last overlapping word with the dialogue, that is, the word *tea*. No distracter has an overlapping word with the dialogue that comes later than the word *tea*.

(man) It's really nice of you to visit me when I'm so miserable with the flu. I'm sure I'd feel much better if I just had some of my mom's homemade chicken soup.

(woman) That will be [that'll be] hard to come by, but a cup of hot tea might help.

(narrator) What will the woman probably do next?

(A) Make some tea for the man.\*

(B) Take the man to see a doctor.

(C) Ask the man's mother to come over.

(D) Look up a recipe for chicken soup.

***V41: There Is Overlap Between Words in the Key and Words Spoken by the Second Speaker in the Dialogue.***

*Coding instructions for V41.* If the key has a word or words that overlap with those of the second speaker in the dialogue, code 1; else 0.

***V42: There Is Overlap Between Words in the Key and Words in the Last Clause of the Dialogue.***

*Coding instructions for V42.* If the key has a word or words that overlap with those of the last clause in the dialogue, code 1; else 0.

***V43: The Key Has a Word That Is Synonymous With a Word in the Last Clause of the Dialogue.***

*Coding instructions for V43.* For items coded 0 for V42, if the key has a word that is synonymous with a word in the last clause of the dialogue, code 1; else 0.

***V44: All Three Distracters Have Lexical Overlap With the Dialogue That Comes Later in the Dialogue Than Does Any Lexical Overlap of the Key.***

*Coding instructions for V44.* If all three distracters have lexical overlap with the dialogue that comes later in the dialogue than does any lexical overlap of the key, code 1; else 0.

In the example below, there is overlap between the word *go* in the key and the word *go* in the dialogue. Each of the three distracters have words that overlap with words in the dialogue that come later in the dialogue than does the word *go*.

(man) Dennis would like us to go bowling with him this weekend.

(woman) I'd love to—but not until I get this project out of the way ... and that could take weeks!

(narrator) What does the woman mean?

(A) She doesn't like bowling.

(B) She probably won't be able to go.\*

(C) She'll go bowling with Dennis next week.

(D) She'll help Dennis with his project this weekend.

**Other Text-Processing Codes**

***V45: An Inference Is Required to Respond Correctly to the Item.***

Variable V45 identifies 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 test inference have stems worded “What does the man/woman imply?” or “What does the man/woman imply about x?” One example is given below.

(woman) What did you think of the new doctor at the infirmary?

(man) You mean Dr. Randolph? 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. Randolph.

*Coding instructions for V45.* If responding correctly to the item requires an inference, code 1; else 0.

*Additional coding instructions for V45.* Do NOT assign a 1 for this variable if the only inference involved is inferring the referent of one or more pronouns in the dialogue.

***V46: The Utterance of the Second Speaker in the Dialogue Contains Two Sentences, Clauses, Phrases, Exclamations, or Some Combination of These Such That Each of These Sentences, Clauses, Phrases, or Exclamations, in Isolation, Can Yield the Key.***

In the example below, it is possible to respond correctly to this item if one *only* comprehends the sentence, "Oh, it's not a problem anymore" *or* if one *only* comprehends the sentence, "I've found an ointment that works just fine." It is not necessary to comprehend both sentences to respond correctly to this item.

(woman) Have you seen the doctor about your skin condition yet?

(man) Oh, it's not a problem anymore. I've found an ointment that works just fine.

(narrator) What does the man imply?

(A) The doctor was too busy to see him.

(B) He doesn't need to see the doctor.\*

(C) The woman should use the ointment.

(D) His skin condition has gotten worse.

*Coding instructions for V46.* If there are two sentences, clauses, phrases, exclamations, or some combination of these in the turn of the second speaker in the dialogue such that each of them, in isolation, can yield the key, code 1; else 0.



*Additional coding instructions for V46.* When coding this variable, one should assume that the test taker has correctly inferred the referents of any pronouns used by the second speaker. In the example below, one should assume that the test taker has inferred that the pronoun *it*, spoken by the man, refers to the South Dorm.

(woman) I need a place to live next semester. The ride back and forth to class this year was too much.

(man) Did you check out the South Dorm? The rooms are pretty small, but it's close to everything.

(narrator) What does the man suggest the woman do?

(A) Move out of the South Dorm.

(B) Find a bigger room.

(C) Look for a room in the South Dorm.\*

(D) Stay where she lives now.

***V47: Only Comprehension of the Utterance of the Second Speaker Is Needed to Respond Correctly to the Item.***

In the example below, it is only necessary to comprehend what the second speaker has to say in order to respond correctly to this item.

(man) What have you heard about Professor Smith? I'm thinking of taking an advanced engineering course with him.

(woman) You really should. One of his articles just won some sort of award—and I heard he's always publishing something in the journals.

(narrator) What does the woman say about the professor?

(A) His classes are very difficult.

(B) His work is well respected.\*

(C) He will publish a book soon.

(D) He is no longer teaching.

*Coding instructions for V47.* If it is not necessary to comprehend what the first speaker says in order to respond correctly to this item, code 1; else 0.

*Additional coding instructions for V47.* This code is NOT assigned to an item if the key for the item uses any term used by the first speaker unless the term is also present in the response of the second speaker and/or in the question asked by the narrator.

***V48: The Key Is a Suggestion or Directive.***

*Coding instructions for V48.* If the key is a suggestion or directive such as including the word *should* or using the imperative form of a verb, code 1; else 0. Below are two examples of items coded for this variable.

Example 1:

(woman) How often do the buses run?

(man) Every half hour on weekdays, but I'm not sure about weekends.

There's a schedule on the corner by the bus stop.

(narrator) What does the man imply?

(A) The woman should check the bus schedule.\*

(B) The buses stop running on Fridays.

(C) The bus doesn't stop at the corner.

(D) The schedule on the corner is out-of-date.

Example 2:

(woman) I need to be in the city by 9 a.m. to get to a 9:30 [nine-thirty] doctor's appointment.... Do you think I should take the bus or the train?

(man) Let's see ... the bus doesn't arrive till 9:45 [nine-forty-five].... Oh! But the train gets in at quarter to nine.

(narrator) What does the man suggest the woman do?

(A) Reschedule her appointment.

(B) Travel by bus.

(C) Meet him at the bus station.

(D) Take the train to the city.\*

***V49: The Key Seems to Be Inconsistent With the Content of the Dialogue.***

Examples of items coded for this variable are given below.

1. In a number of items where the narrator asks about what the second speaker assumed, the key seems to be inconsistent with what is said in the dialogue. In the example below, there is an apparent inconsistency between the key (“Someone would drive them (the cousins) home”) and “So they (the cousins) didn’t manage to get a lift after all” in the dialogue.

(man) Your cousins just called. They’re stranded at the beach.

(woman) So they didn’t manage to get a lift after all.

(narrator) What had the woman assumed about her cousins?

(A) Their friends would take them to the beach.

(B) They wouldn’t mind taking the bus.

(C) Someone would drive them home.\*

(D) They wouldn’t be able to find a phone.

2. In a number of dialogues that involve sarcasm, the key seems to be inconsistent with what is said in the dialogue. In some of these cases, there is apparent praise of someone or something in the dialogue, whereas there is criticism in the key.

(man) Can you believe it? Now we’re supposed to bring a note from our instructor every single time we want to use the computer!

(woman) [sarcastically] I’ll bet that was another one of Mike’s brilliant ideas!

(narrator) What does the woman imply about Mike?

(A) He often makes foolish suggestions.\*

(B) His instructor won't give him a note.

(C) He should try using the computer himself.

(D) He is a very good instructor.

3. Another example of where the key seems to be inconsistent with what is said in the dialogue is where a seemingly negative response to a request is actually a positive one.

(woman) Mind if I borrow your economics notes for a while?

(man) Not at all.

(narrator) What does the man mean?

(A) He'll only give her part of his notes.

(B) He doesn't know anything about economics.

(C) He's not taking an economics class.

(D) He's happy to lend her his notes.\*

*Coding instructions for V49.* If the key seems to be inconsistent with what is stated in the dialogue, code 1; else 0.

*Additional coding instructions for V49.* This code is NOT assigned if a statement in the dialogue appears to be inconsistent with a later statement in the dialogue itself, as in the example below:

(woman) A lot of people were excited about the class election.

(man) But they didn't turn out to vote, did they?

(N) What does the man imply about the students?

(A) They weren't really interested in the election.\*

(B) They didn't vote for the best people.

(C) Their votes weren't counted.

(D) They remained enthusiastic about the candidates.

## Appendix B

### Instructions for Coding Lexical Overlap

Only words with helpful lexical overlap are coded, that is, if the key has lexical overlap with the dialogue that is identical to the lexical overlap of all three distracters, it is not coded for lexical overlap. For example, in the item below, the word *Nancy*, which appears in the dialogue, is common to all four options; this word is not coded for lexical overlap.

(man) We got a thank-you note from Nancy today. She said she's already worn the scarf we sent.

(woman) That's great. I wasn't sure if she'd wear red.

(narrator) What had the woman been concerned about?

(A) Nancy wouldn't send a thank-you note.

(B) Nancy hadn't received the scarf.

(C) Nancy wouldn't like the gift.\*

(D) Nancy doesn't wear scarves.

The instructions below typically refer to lexical overlap between words in the dialogue and words in the key. It should be noted that the instructions apply equally well to lexical overlap between words in the dialogue and words in the distracters.

I. For *content words* (i.e., nouns, main verbs, adjectives, and adverbs), use the instructions below to determine whether there is lexical overlap between a word in the key and a word in the dialogue.

1. Lexical overlap between a word in the key and a word in the dialogue is coded if the root of the words is the same; for example, *expecting* and *expected* would be coded as lexically overlapping words because both share the same root (i.e., *expect*). In the example below, lexical overlap is coded between the word *reading* in the dialogue and the word *read* in the key because both have the same root (*read*). There is also lexical overlap in this item between the word *page* in the dialogue and the identical word *page* in the key.

(man) You've certainly been reading that one page for a long time now.

(woman) Well, I'm being tested on it tomorrow.

(narrator) What does the woman imply?

(A) She's reading a very long book.

(B) The man is mistaken.

(C) She needs to read the page carefully.\*

(D) She's working on a long assignment.

2. To code lexical overlap between a word in the key and a word in the dialogue, the words need to have the same or similar meanings; for example, the word *left*, when used to refer to a direction, would NOT be coded as having lexical overlap with the word *left*, when it is the past tense of the word *leave*. In the following item, lexical overlap is NOT coded between the word *go* in the key and the word *going* in the dialogue, since these two forms of the word *go* have quite different meanings.

(woman A) That famous violinist our professor was talking about is going to be the soloist in next week's concert!

(woman B) Great! I don't want to miss it. Where can we get tickets?

(narrator) What will the speakers probably do next week?

(A) Find out where their professor is going to perform.

(B) Go to a concert.\*

(C) Perform in a musical recital.

(D) Interview the violinist.

3. If a word appears twice in a dialogue but refers to two different things, lexical overlap is **only** coded between the word in the key and the word with the same referent in the dialogue. In the example below, the word *salad* refers to two different things in the dialogue. One only codes for lexical overlap between the word *salad* in the key and the word *salad* spoken by the second speaker because these two words have the same referent (i.e., tuna salad), whereas one does NOT code for lexical overlap between the word *salad* in the key and the word *salad* spoken by the first

speaker, since in the key the word *salad* refers to tuna salad whereas the word *salad* spoken by the first speaker refers to a different referent, namely, chicken salad.

(man) Are you sure this is what I ordered? This looks like chicken salad.

(woman) Oh, I'm sorry. You ordered the tuna salad, didn't you? I'll be right back with it.

(narrator) What does the woman mean?

(A) She wants to eat chicken salad.

(B) The chicken salad is gone.

(C) She dropped the man's food.

(D) She'll bring the tuna salad.\*

4. A word in the key is coded as having lexical overlap with a word in the dialogue if the same word appears as part of a compound word in the dialogue or vice-versa. In the example below, lexical overlap is coded between the word *hall* in the key and *hall* in the compound word *hallway* in the dialogue.

(man A) I can hardly read because it's so dark in this classroom.

(woman B) It is in the hallway, too.

(narrator) What does the woman mean?

(A) The hall is also dark.\*

(B) It's difficult to read while class is going on.

(C) The reading assignment was too long.

(D) All the classrooms are the same.

5. Lexical overlap is coded between a word that is commonly used as a substitute for a longer word of which it is a part and the longer word itself. In the example below, lexical overlap is coded between the word *dorm* in the dialogue and the word *dormitory* in the key, since *dorm* is part of the longer word *dormitory* and is frequently used instead of the longer word.

(woman) You know, the noise in my dorm has really gotten out of control. My roommate and I can rarely get to sleep before midnight.

(man) Why don't you take the problem up with the dorm supervisor?

(narrator) What does the man suggest the woman do?

(A) Discuss the situation with the person in charge of the dormitory.\*

(B) Ask her roommate not to make so much noise.

(C) Go to bed after midnight.

(D) Send a letter to the residents.

II. For *function words* (i.e., determiners, auxiliary verbs, conjunctions, prepositions, and pronouns), use the instructions below to determine whether there is lexical overlap between a word in the key and a word in the dialogue.

1. Determiners such as *a* and *the* in the key are coded as having lexical overlap with the same words in the dialogue only when they directly precede the same content word. For example, if *the dog* appears in the key and *the dog* also appears in the dialogue, both words are coded as having lexical overlap.

a) In the example below, lexical overlap is coded between the words *the party* in the key and the same words *the party* in the dialogue.

(man) My math assignment's due tomorrow morning and I haven't even started it yet.

(woman) I'll miss you at the party tonight.

(narrator) What does the woman imply?

(A) The party will be crowded.

(B) The man will do his assignment before the party.

(C) She's not going to the party.

(D) The man won't be able to go to the party.\*

b) In the example below, lexical overlap is only coded between the word *machine* in the key and the word *machine* in the dialogue. Lexical



overlap is NOT coded between the word *the* in the key and the word *the* in the dialogue because the word *the* in the dialogue does not directly precede the word *machine*.

(man) I can't seem to get the copy machine to work.

(woman) Have you checked the switch?

(narrator) What does the woman imply?

(A) The machine works like that other one.

(B) The man should change machines.

(C) The machine might not be turned on. \*

(D) The man might be charged for the copies.

2. Auxiliary verbs in the key are coded as having lexical overlap with the dialogue only when they have the same function in the key as in the dialogue, that is, they precede the same or similar content. In the example below, the auxiliary verb *hasn't* precedes content in the key that is similar to the content it precedes in the dialogue.

(woman) Has Alice decided on a major yet? I know she was thinking about American history.

(man) She has so many interests—as far as I know she hasn't been able to make up her mind.

(narrator) What does the man say about Alice?

(A) She isn't interested in being a historian.

(B) She hasn't chosen a course of study.\*

(C) She's studying American history.

(D) She's a very good student.

*Additional coding instructions for auxiliary verbs.* The above instructions also apply to contracted auxiliary verbs (e.g., *'ll* as in *she'll* or *I'll*).

3. Forms of the verb *to be* in the key are coded as having lexical overlap with the dialogue only when they have the same function in the key as in the dialogue, that is, they precede the same or similar content.

- a) In the example below, lexical overlap is coded between the verb *been* in the key and the verb *been* spoken by the second speaker in the dialogue because *been* is followed by similar content in both cases. (Lexical overlap is also coded for this item between the word *paper* in the key and the word *paper* in the dialogue.)

(woman) I haven't seen you at the student center all week. Have you been sick?

(man) I've been overwhelmed with my history paper.

(narrator) What does the man mean?

(A) He decided to attend extra history classes.

(B) He hopes to meet the woman at the student center.

(C) He was too sick to work on his paper.

(D) He's been busy working on his paper.\*

- b) In the example below, lexical overlap is NOT coded between the verb *is* in the key and the verb *is* spoken by the first speaker in the dialogue because the content following the verb is quite different in the two cases.

(woman) This is the car you bought? I've never seen such an old jalopy!

(man) It may not look like much, but it gets me where I'm going.

(narrator) What does the man mean?

(A) The car is dependable.\*

(B) The car isn't very old.

(C) This car is better than his old one.

(D) He paid too much for the car.

4. Prepositions in the key are coded as having lexical overlap with the same preposition in the dialogue when the preposition has the same function in the key as it has in the dialogue (i.e., when the preposition precedes the same word, or when it precedes a synonym of the word, or when it precedes a word that refers to the same thing in the

key as it does in the dialogue). In the example below, lexical overlap is coded between the preposition *with* in the key and the preposition *with* spoken by the second speaker, since both instances are followed by words that refer to the same thing. (Lexical overlap is also coded for this item between the word *ski* in the key and the word *skiing* in the dialogue.)

(woman) Can you come skiing with me this weekend, or do you have to study for your exams?

(man) I'll come along with you, but I'm so tired from studying that I'm afraid I won't be doing much skiing.

(narrator) What will the man probably do?

(A) Stay home and study all weekend.

(B) Stay home and rest all weekend.

(C) Go with the woman and ski all weekend.

(D) Go with the woman and rest rather than ski.\*

5. Pronouns in the key are coded as having lexical overlap with the same pronouns in the dialogue when the pronoun refers to the same thing in both cases. In the example below, the word *she* in the dialogue and the word *she* in the key both refer to the same person, Laura.

(woman A) What's Laura doing here today? I thought she was supposed to be out of the office on Mondays.

(woman B) She decided she'd rather have Fridays off instead.

(narrator) What can be inferred about Laura?

(A) She has changed her schedule.\*

(B) She was sick on Friday.

(C) She works less than she used to.

(D) Her vacation started on Monday.

6. Conjunctions in the key are coded as having lexical overlap with the same conjunctions in the dialogue when the conjunction has the same function in the key

as it has in the dialogue, that is, when the conjunction precedes the same or similar content. In the example below, lexical overlap is coded between the word *why* in the key and the word *why* in the dialogue because the words precede similar content.

(man) Joe took a taxi home alone ten minutes ago.

(woman) I wonder why he didn't wait for me to go with him.

(narrator) What does the woman mean?

(A) She wanted to visit Joe's home.

(B) She doesn't understand why Joe left without her.\*

(C) Joe should take a taxi to her house.

(D) Joe didn't want to take the taxi to his house.

7. Negative forms of verbs such as *can't*, *doesn't*, and *haven't* are not coded as having lexical overlap with positive forms of these verbs; that is, lexical overlap is not coded between *can't* and *can*.



**Test of English as a Foreign Language  
PO Box 6155  
Princeton, NJ 08541-6155  
USA**

---

To obtain more information about TOEFL programs and services, use one of the following:

**Phone: 1-877-863-3546  
(US, US Territories\*, and Canada)**

**1-609-771-7100  
(all other locations)**

**Email: [toefl@ets.org](mailto:toefl@ets.org)**

**Web site: [www.ets.org/toefl](http://www.ets.org/toefl)**

\* America Samoa, Guam, Puerto Rico, and US Virgin Islands