Does the text matter in a multiple-choice test of comprehension? the case for the construct validity of TOEFL's minitalks

Roy Freedle and **Irene Kostin** *Educational Testing Service, Princeton, NJ*

The current study addresses a specific construct validity issue regarding multiple-choice language-comprehension tests by focusing on TOEFL's minitalk passages: Is there evidence that examinees attend to the text passages in answering the test items? To address this problem, we analysed a large sample ($n = 337$) of minitalk items. The content and structure of the items and their associated text passages were represented by a set of predictor variables that included a wide variety of text and item characteristics identified from the experimental language-comprehension literature. Stepwise and hierarchical regression techniques showed that at least 33% of the item difficulty variance could be accounted for primarily by variables that reflected the content and structure of the whole passage and/or selected portions of the passage; item characteristics, however, accounted for very little of the variance. The pattern of these results was interpreted, with qualifications, as favouring the construct validity of TOEFL's minitalks. Our methodology also allowed a detailed comparison between TOEFL reading and listening (minitalk) items. Several criticisms concerning multiple-choice language-comprehension tests were addressed. Future work is suggested.

I Introduction

1 Purpose of current study

The purpose of this study is to examine a very specific method of evaluating the construct validity of TOEFL's minitalks. The problem to be addressed is whether there is evidence that examinees attend to the text passages when answering the test items.

Address for correspondence: Roy Freedle, Educational Testing Service, Mail Drop 11-R, Princeton, NJ 08541, USA; e-mail: rfreedle@ets.org

Language Testing 1999 16 (1) 2–32 0265-5322(99)LT159OA

2 Background

Several extreme criticisms regarding the construct validity of multiple-choice tests of language comprehension, especially for reading, were summarized by Farr *et al.* (1990). The most serious criticism maintains that examinees do not or need not comprehend the texts accompanying the test items in order to answer the items. In particular, for listening tests, one would have to claim that examinees do not or need not have to listen to the minitalk passage in order to answer the test items. At the very least one should be able to counteract such criticisms by showing that some variables that reflect the structure and content of the text passage are significantly correlated with item difficulty. Finding such significant correlations would indicate that examinees are probably paying attention to text information and are using this information to guide their selection of answers to the items. Such a demonstration could be argued to constitute a modest but still viable claim for the construct validity of a test. We do not claim, however, that a multiple-choice language-comprehension test that shows such effects is necessarily measuring what it 'should' measure; we only claim that the criticisms raised by some critics of standardized testing have been too extreme and that, in part, correlational evidence can be gathered to counteract these extreme criticisms. (At the end of this paper we do, however, raise the issue of evaluating a broader notion of construct validity.)

Related criticisms of multiple-choice language-comprehension testing, especially for reading, can be found (Drum *et al*., 1981). Drum *et al*. studied the predictability of multiple-choice item difficulty of children's reading-comprehension tests by using several types of predictor variables: text variables (e.g., average sentence length of text) and item variables (e.g., counting the number of content words such as verbs, nouns, adjectives in, say, the correct option). Interestingly, using a regression procedure, they reported that their best predictors of item difficulty involved primarily their so-called item variables (see, however, Freedle and Kostin, 1994). Such a result assesses evidence for or against construct validity in a different way. Their result suggests that, while a few of the text predictors were in fact significantly correlated with item difficulty (thereby providing some evidence of validity), nevertheless the multiple-choice language-comprehension test may not be construct valid in a stronger sense because the item predictors appeared to be more prominent than the text predictors. Ideally, a language-comprehension test 'should' assess primarily the difficulty of the text itself – the item structure itself should only be an incidental device for assessing text difficulty. Therefore, since the comprehension test items used by Drum *et al*. did not appear

to be primarily measuring significant aspects of the passage content, but rather seemed to reflect primarily the difficulty of how the items themselves were constructed, they were compelled to question the robustness of their tests' construct validity.

We interpret the Drum *et al*. result as showing that there are different 'levels' of construct validity that can be discerned. In the correlational sense being examined herein, the lowest 'level' of validity requires finding any significant support for the effect of text variables on test item difficulty. A higher 'level' of validity requires that if both text and item variables are significant, then the text variables should be more important than the item variables in accounting for test-item difficulty variance. Freedle and Kostin (1996) recommended considering yet a third 'level' of construct validity which they associated with what they called a text-by-item overlap predictor in combination with a hierarchical regression procedure. This idea of a third 'level' will now be explained.

Freedle and Kostin (1996) presented evidence for three logically different types of predictor variables: text variables item variables, and text-by-item overlap variables. For example, regarding the latter type, suppose some content words that occur in the passage are also used in one or more of an item's options. One might hypothesize that people will tend to choose an option if it contains more text words than some other options. Such a variable for predicting item difficulty would represent a text-by-item overlap variable because measuring it requires examination of both the item content and the passage content in order to enter the appropriate lexical overlap count.

The three distinct 'levels' of construct validity proposed by Freedle and Kostin (1996) are as follows: (1) the lowest 'level' of validity should demonstrate that one or more of the text and/or text-by-item overlap variables are significantly correlated with item difficulty, (2) the intermediate 'level' of validity should show, in a stepwise regression predicting item difficulty, that one or more text and/or text-by-item variables play a more significant role in accounting for item difficulty than do the item variables, (3) the strongest 'level' of validity should show, in a hierarchical regression analysis, that even after partialling out the effect of all item predictor variables, the text and text-by-item overlap variables contribute significantly to item difficulty; ideally, for this 'highest' level, the amount of variance accounted for by the text and text-by-item overlap variables should be greater than the variance accounted for by just the item variables. In our study below we follow the three-level construct validity approach recommended by Freedle and Kostin (1996).

3 The use of listening and reading tests as measures of general language comprehension

While we certainly do not claim that reading and listening are synonymous, we do claim that reading and listening are strongly related due to the fact that they both require the exercise of a general faculty of language comprehension. Indeed, several studies have demonstrated strong intercorrelations between listening and reading comprehension (Sticht and James, 1984; Hale *et al*. 1989). For example, Hale *et al*. report a median correlation of .82 between TOEFL's listening and nonlistening (including reading) sections. Other studies suggest that both listening and reading employ similar cognitive processes (see Kintsch and Kozminsky, 1977). Given this evidence that reading and listening both appear to be largely measures of a more general language comprehension ability, we assume below that it is reasonable that variables that significantly predict reading comprehension difficulty are likely also to significantly affect listening comprehension difficulty. In agreement with our central thesis Carroll (1993: 146) has characterized the language domain as follows: [A]ll language abilities tend to be rather highly correlated; their general degree of correlation can be attributed to the influence of a general higher-order factor of language ability or general language development. … But there is also some specialization of abilities …'

4 Selecting predictor variables

a Literature review of several variables involved in language comprehension. We will frame this section around a brief review of variables that have been demonstrated to affect language comprehension, with a focus on reading; a more in-depth discussion of many of these variables can be found in Freedle and Kostin (1991; 1992; 1993; 1996). The main assumption here that justifies the review of language variables for use in a listening study is that reading and listening, as argued above, are both substantially measures of general language comprehension. We will first develop a set of variables that have already been shown to be predictive of TOEFL reading-item difficulty (Freedle and Kostin, 1993); as stated above it is our expectation that many of these variables will also prove useful in accounting for listening item difficulty. (To conserve space, more detailed examples of many of the categories described immediately below will be given in the Method section.)

- *Negations*. Carpenter and Just (1975) found that sentence negations typically increased comprehension decision time in comparison with sentences without negations. This increase suggests that the number of negations (e.g., 'no, none, never' and

including prefixed and suffixed words such as '*un*certain', 'point*less*') contained in TOEFL listening passages may also influence multiple-choice item difficulty. Furthermore, one can inquire whether additional negations that are used in the item structure itself (either in the item stem and/or among the response options) may also separately contribute to comprehension difficulty over and above the contribution of text negations. This last observation means that we can consider several predictor variables: a tally of the negations that occur in the minitalk proper, another tally of the negations that occur in the item stem (this auditorily presented segment frames the question to which the examinee must respond), another tally of the negations that occur in the correct option (this option is printed and must be read by the examinee), and finally another tally of the negations that occur among all the incorrect options which are also printed. Thus, four separate predictor variables are to be generated that deal with negations: the minitalk proper, the item stem, the correct option and the incorrect options. For each such predictor we expect that the more negations there are, the harder the minitalk item will be.

The reader should note that the four separate types of predictor variables just described for negations was used, where appropriate, in developing all remaining variables for the categories discussed immediately below.

- *Referentials*. Abrahamsen and Shelton (1989) demonstrated improved comprehension of texts that were modified, in part, so that full noun phrases were substituted in place of referential expressions. This improvement suggests that texts with few referential expressions may be easier than ones with many referential expressions.
- *Rhetorical organizers*. Meyer and Freedle (1984) and Hare *et al*. (1989) studied, in part, the effect of four of Grimes' (1975) rhetorical organizers on passage comprehension difficulty: list/describe, cause/effect, compare and problem/solution. Both studies found significant effects of these four top-level rhetorical organizers on comprehension but differed as to which types were facilitative. A directional prediction will, therefore not be made.
- *Fronted structures*. Freedle *et al*. (1981) reported differences in the use of 'fronted' structures at sentence beginnings (and paragraph beginnings) as a function of the judged quality of student essays. Fronted structures included the following: (1) cleft structures, (2) marked topics and (3) combinations of coordinators and marked topic or cleft structures. Each of these fronted structures initiated an independent clause. Freedle and Kostin (1995)

reported complex effects of fronted structures for a reading test. Therefore a directional prediction will not be made.

- *Serial position effects*. The work of Kieras (1985) and Hare *et al*. (1989) showed that it is easier to locate information (i.e., main idea information) that is early or late in a paragraph; information in the middle of paragraphs is harder to identify. A related study by Bhasin (1990) suggested that bilingual Spanish-English students tended to focus on initial text information in helping them to select a response option in a multiple-choice comprehension test.
- *Lexical overlap*. Freedle and Fellbaum (1987) found that lexical overlap helps to account for multiple-choice item difficulty in the TOEFL listening subsection involving single-sentence comprehension such that an item option that contained the greatest lexical overlap (i.e., key content words) with the auditorily presented single-stimulus sentence tended to be the option that was selected by the test takers.
- *Vocabulary, sentence length, passage length and abstractness of text*. Longer sentence structures and longer and less frequently occurring words tended to be associated with greater difficulty of text comprehension, as can be inferred from their use in traditional readability formulas (see Graves, 1986); in addition, longer passages and abstractness of texts were associated with increased difficulty in comprehending passages (see Newsome and Gaite, 1971 and Paivio, l986, respectively).

b New predictor variables tailored for the TOEFL's minitalk listening environment. Above, we have hypothesized that reading and listening tests are measurements of a general language comprehension ability. In part a study by Hale *et al*. (1989) was used to justify this assertion. However, in spite of the substantial intercorrelations between reading and listening (median correlation of .82), Hale *et al*. did present evidence for the existence of two separate but highly intercorrelated factors: one of listening and the other of a nonlistening factor. Immediately below we list five variables that may clarify the source of some of this factorial difference.

1) *The frequency of emphatic text words*. Texts that contain a fairly large number of emphatic words introduce a variety into the stimulus that is absent from minitalk texts with no emphatic words. Texts that use many emphatic words may thus be more memorable. Hence, items may be easier when they are associated with texts that have many emphatic words. (Anticipating now the method section, we used the emphatic words that were highlighted in the script used to make the minitalk recordings – the tapes themselves were not listened to due to time constraints in

completing the study. For further clarification, we refer the reader to variable v48 below.)

2) *The frequency of filled or unfilled pauses*. A filled pause consists of such sounds as 'um', or 'er'. An unfilled pause is a pause of about 1 second or longer without any audible speech sound in that interval. The professional speakers who recorded the minitalks inserted these filled and unfilled pauses intentionally to render greater appearance of validity to the minitalks.

   It is difficult to anticipate the likely effects of having variable numbers of pauses on items associated with minitalks. Blau (1990) found that pauses at constituent boundaries improved second-language (L2) listening comprehension. But Chaudron and Richards (1986) found that pause fillers did not aid lecture retention. Perhaps the effect of pauses on comprehension will prove to be weakly facilitative. Filled pauses ('um', 'er') were indicated on the typescript for some of the minitalks. Long unfilled pauses were indicated by '…' on the same typescript. (Anticipating again the method section, although these two types of pauses are logically distinct, they were combined for this study because of the low frequency of occurrence of each of the two types; see variable v49 below for further clarification.)

3) *Estimates of the redundancy of information*. Chiang and Dunkel (1992) found that redundancy benefited high listening-proficient Chinese students, whereas no improvement was noted for the low listening-proficient students. While it may vary with ability level, redundancy does, therefore, play a significant role in comprehension. Redundancy in the Chiang and Dunkel study involved a measure consisting of elaboration and repetition of information. As an example of elaboration, they give: 'The food … is very hearty and delicious. Hearty and delicious food is nourishing and tasty.' Repetition of information would consist of repeated segments or paraphrases of information. (The above example can, in fact, be seen to be a combination of both repeated as well as elaborated information over successive sentences.)

   Parker and Chaudron (1987) found that elaborative modifications involving repetition of the information plus clear segmenting of the thematic structure enhanced orally comprehended L2 information. Buck (1990) has reported that more natural listening materials appear to have a high redundancy aspect to them, among other characteristics. Indeed, when he compared examinee reading and listening scores he found significant evidence that listening is a skill distinct from reading providing that the listening materials contain, among other things, more redundant materials.

Given the above evidence, it would appear to be desirable to develop some predictor variables that reflect how test items tap various redundant properties of the minitalks, because such measures of redundancy are likely to be correlated with listening item difficulty. Several scores were developed. Suppose an item's correct option consists of a phrase or clause that essentially restates information that is repeated in more than one sentence of the minitalk. Such an item should be easier than one in which the correct option is a paraphrase of information present in just one sentence of the minitalk, because the first item would be more redundantly represented in the text than the second item. Two types of redundancy scores such as this were developed designated below as ii10 and ii12 (as applied to inference item types) and ss12 and ss14 (as applied to the supporting item types).

Another score involved both redundancy and what we call complementary information. That is, sometimes a correct option implicates one sentence in the minitalk that is an incomplete paraphrase in the sense that one must search additional parts of the minitalk to locate additional (complementary) information needed to complete the paraphrase; such items probably will prove to be harder than items that simply paraphrase information in just one minitalk sentence since the examinee will have to do more cognitive work in locating in memory the complementary information and linking it to the partial paraphrase in order to arrive at the correct answer. This other type of 'redundancy' score was developed for both inference items (see ii11 below) and supporting items (ss13 below).

4) *A new lexical overlap measure*. Freedle and Fellbaum (1987) found that the greater the amount of lexical overlap between correct options and single stimulus sentences (from the TOEFL listening section) the easier the items. This finding encourages one to expect a similar finding to occur for more extended listening materials such as is found for the minitalks.

There is typically a crucial text sentence that one must process to get an item correct (see Freedle and Kostin, 1993). But it is conceivable that the rest of the extended text can have a lexical impact on item correctness as well. We hypothesize that when words in the correct option occur repeatedly throughout the text (i.e., not just for the crucially important sentence) such repetitive effects can also affect item difficulty – incidentally, this repetition is a new kind of redundancy, if you will, at the lexical rather than sentential level. Even the incorrect options may also influence the item difficulty due to lexical repetition. That is, a

particular incorrect option may be made especially attractive if it employs words that are frequently used throughout the minitalk text. These various considerations suggest a new score called a lexical attractiveness score. Below, this new score is designated ii13 (for the inference items) and ss15 (for the supporting idea items).

5) *Topical differences in minitalks*. There are differences in the types of topics that are covered by the minitalks in comparison with the reading passages. For example, while the academically oriented minitalks use the same topical areas covered by the reading passages (e.g., physical science, humanities, etc.), there is a range of nonacademic type topics (e.g., sightseeing, extracurricular activities) that are covered in the minitalks but are *absent* from the reading passages. It seems likely that items that inquire about the nonacademic topics may, because of their greater general familiarity, be easier than items about academic topics. In order to examine such topical differences on item difficulty, we have developed several variables that reflect academic vs nonacademic topics. Below these are described by variables v17 through v21 (for academic topics) and v22 through v27 (for nonacademic topics).

II Materials and method

The total item sample consists of 337 listening comprehension items associated with 69 minitalk passages. The passages and associated items were taken from 47 TOEFL forms administered between 1981 and 1992.

1 Sequence of events for each minitalk

The sequence of events associated with each minitalk is as follows. (1) The lead-in (sometimes containing brief contextual information about where a lecture has been given or what its topical content is; e.g., 'Listen to the report on biology', (2) the minitalk itself, (3) the item stem that introduces the problem to be solved and (4) the typed options from which the examinee makes a selection. (The first three events are presented auditorily.) Each of these components of a minitalk was assigned a number of variables that was intended to predict item difficulty (see Appendix A).

2 Types of item studied

Four types of item were studied: detail explicit, detail implicit, gist explicit and gist implicit.

- *Detail explicit* (n = 183); also called Supporting Ideas. Example: 'What year did the teacher give for the critical experiment?'
- *Detail implicit* (n = 117); this item type consists of two subtypes: plain inference items and inference-application items. Eighty-two of these were inference items where the inferences can be made based on information available in the text. Example: 'One can infer that the speaker intends to do which of the following …'. These simple inference items were quite similar to those found in the TOEFL reading section.

  For the other subtype, inference-application items, listeners must use their background knowledge to make the inferences. There were 35 of these items. Example: 'Who is the person giving the lecture?' Usually no direct information is given in the minitalk regarding the lecturer's identity so the information must be inferred strictly from background knowledge that the examinee brings to the task. (Incidentally, inference-application items are not found in the reading section.)
- *Gist explicit and implicit* (n = 37); also called main-idea items. Because the total sample of gist items was so small we decided not to divide this category into its two respective item types.

The measure of item difficulty for each item (equated delta) was based on the performance of approximately 2000 examinees. These examinees were randomly selected from a much larger pool of examinees who responded to each TOEFL test form. The equated delta value slightly adjusts the difficulty of each item across forms so that items can be meaningfully compared across groups of people taking different test forms. The adjustment stems from the fact that the examinees who respond to a particular test form differ slightly in overall ability level from those responding to other test forms.

3 Independent and dependent variables used in this study

Appendix A presents a list of all the coded variables along with a brief definition of each variable. Not all variables were used in the analyses. In Appendix A those variables without a superscript are the ones used in the analyses in this report while those few variables having a superscript were deleted due to low frequency of occurrence or because of collinearity with other predictor variables (collinearity is defined as a correlation of .8 or higher with any other variable; see Nie *et al.*, 1975). For pairs of variables correlating .8 or more, that variable that correlated more strongly with the dependent variable was retained. The following collinear variables were deleted: v6, v13, v15, v35, v37, v46, mm9, mm11, mm13, ii8, ss10. Because of low frequencies of occurrence, defined as two or fewer occurrences in the *n*

= 337 sample, the following variables were also deleted: mm3 and mm6. Additional variables that were deleted are as follows: v22, v23, v24, v25, v26. All of these latter variables represented nonacademic topics that had been individually listed. These five variables were condensed into a single variable (v27); all the component variables were observed to be negatively correlated with the dependent variable and each had a low frequency of occurrence. Additional information concerning all variables is available in Freedle and Kostin's (1996) report.

There are four groups of independent variables used to predict listening item difficulty; the dependent variable constitutes one additional variable.

a Four groups of independent variables

1) *Item variables*. These variables constitute our so-called 'pure' item variables referred to below. That is, these variables can be coded without reference to the contents of the minitalk passage; only the contents of the item itself are used to quantify these particular variables.
2) *Text variables*. These variables characterize the content and structure of the minitalk passage itself; the contents of the items themselves do not figure in the coding of these variables.
3) *Text/item overlap variables*. We define the concept of text-by-item or alternatively text/item overlap variables as variables that necessarily reflect the contents of both the test items as well as the text to which those items apply.
4) *Item types*. We define item type as a special type of text/item overlap. We do this even though item type per se appears to be a pure item variable. On reflection, however, it seems that an item type cannot be correctly determined without checking how it functions for the passage to which it refers; for this reason item types are another kind of text/item overlap code because both the item and the text must be scanned to arrive at a proper coding.

A total of 18 variables were deleted (see rationale for deletion given above) leaving 11 item variables, 31 text variables and 39 text/item overlap variables (the 39 text/item overlap variables include the 4 item type variables).

b Dependent variable. There was one dependent variable – equated delta. The dependent variable (v54) is an item's equated delta. Difficult items have large equated deltas; easy items have small equated deltas.

4 Reliability of variables requiring subjective judgement

While many of our predictor variables are arrived at objectively (e.g., counting the number of words in a passage), the variables listed below required some degree of subjective judgement. The following percentage agreement was obtained for two raters using a sample size of 35 cases: coherence = 74% agreement; referentials = 92% agreement; negations = 96% agreement; frontings = 93% agreement; rhetorical organizers = 89% agreement; location of relevant text = 84% agreement; and abstractness/concreteness = 87% agreement. In general, it is clear that these subjective measures yielded high reliabilities. Because of the high reliability only one coder was used to code the remaining items and passages. Half the variables were coded by author RF while the remaining half were coded by author IK.

III Results and discussion

1 MANOVAs to determine possible interactions with predictor variables

We conducted a series of MANOVAs to help us determine whether there were significant interactions between the predictor variables and the four item types. Only one of the 42 MANOVA analyses suggested a significant interaction; this, however, could easily have been due to chance. Because of these results, all further analyses used the combined item-type samples. (For further information on these analyses, see Freedle and Kostin, 1996.)

2 Two item samples

Two samples of items were defined. The first sample consisted of the following: 337 items (37 main idea items, 82 inferences, 183 supporting ideas and 35 inference application items). The second sample consisted of 302 items (i.e., the 35 inference application items were deleted from the 337 item sample). The reason for deleting the inference application items from the second sample is as follows: the sample of 302 items is more directly comparable to the reading items studied by Freedle and Kostin (1993). Due to space limitations we shall report our regression results only for the 302 item sample. This restriction on which sample we focus on for the regression analyses will not affect our conclusions regarding construct validity (see Freedle and Kostin, 1996).

3 Correlational results

Appendix B presents data that help to identify those variables that are correlationally significant in predicting minitalk item difficulty. In Appendix B we see that 43 different variables – in either or both of the two samples presented in the table – yielded a significant correlation ($p < .05$) with item difficulty (equated delta). Thus of the 81 variables examined (i.e., the non-superscripted variables listed in Appendix A) slightly more than half (43) were significant. A perusal of the correlations between the two item samples ($n = 302$ and $n = 337$) reveals the great similarity of the two samples.

First we will use portions of Appendix B to assess the apparent adequacy of those variables that our literature review suggested would be pertinent to predicting item difficulty. We point out along the way similarities and differences between the reading study results (Freedle and Kostin, 1993) and our current set of listening (minitalk) variables.

Overall, the correlational results suggest that many of those variables found to influence comprehension in the experimental language comprehension literature also influence our multiple-choice listening data. We note that seven of the significant variables are pure item variables. However, the very fact that many of the text and text/item overlap variables are significant suggests that this pattern alone can be taken as support for one view of construct validity of the TOEFL minitalk passages and their associated items – the view that contradicts the extreme claim that examinees do not have to pay attention to the passage in order to get the items correct (see review by Farr *et al*., 1990). Thus the correlations appear to provide positive support for at least this lowest 'level' of construct validity (see introduction for several levels). Evidence regarding other 'levels' of construct validity will be taken up later in this report.

Now we will consider the several classes of variables that our literature review suggested would be effective predictors of language comprehension item difficulty.

For example, in Appendix B we see that the presence of *negatives* was associated with more difficult listening items, as expected. This increased difficulty was found when negatives were present in the correct options (v7), the incorrect options (v11) and the text (v47). The first two scores had a similar effect for TOEFL reading items (see Freedle and Kostin, 1993: 157 for relevant information dealing with reading).

We also see in Appendix B that the presence of more *referentials* had a significant effect on listening-item comprehension, but the effects were not uniform across item components. Thus, the presence of referentials in the item stem (v4) was associated with easier items,

whereas the presence of referentials in the correct option (v9) was associated with harder items. The presence of many extratextual or 'special' text referentials (v45) was associated with easier items. (Terms such as 'you' and 'we' used in a minitalk were classified as extratextual or 'special' referentials). Because it was clear who the 'you' or 'we' referred to in the minitalks (i.e., the listener, and both speaker and listener, respectively), the occurrence of these referentials actually seemed to personalize the listening material and probably for that reason, tended to make items associated with such texts easier. Another explanation is that 'you' and 'we' may have occurred primarily for nonacademic subjects, and such subjects tend to contain easier or more familiar concepts than the academic subjects. In fact, there was some evidence for this latter explanation: there was a high positive correlation of .54 ($p < .01$) between the nonacademic subjects and the occurrence of these extratextual referentials.

Three *rhetorical organizers* had a significant effect on listening item difficulty (v29, v31, v32). The list structure was associated with easier items. The problem/solution and compare structures were associated with harder items for the listening section. Similar results were found for reading regarding the problem/solution and the list organizers (see Freedle and Kostin, 1993: 157).

Fronted structures had a significant effect on listening item difficulty; if the minitalk had a long string of fronted structures in successive sentences (see text variable v41) this pattern was associated with harder minitalk items. A related fronted variable (a text variable) was also found in the Freedle and Kostin (1993) reading study to be associated with more difficult reading items.

Concrete texts (i.e., texts that did not deal primarily with abstract concepts – v16) were associated with easier listening items, as expected. This pattern was also found for reading items (see Freedle and Kostin, 1993).

Average sentence length and *passage length* effects were not significant for listening items. However, a few other variables that assessed additional aspects of length (v1, v10, ss8) were significant. In all three cases the longer the stem, or the incorrect options, or the length of text that was encountered prior to relevant item information in the text, the more difficult the items tended to be – a result that was consistent with prediction. We observe that one of these 'length' type effects (v10 – number of words in incorrect options) was significant for both reading and minitalks tests. Several additional length effects were significant for reading items (see Freedle and Kostin, 1993).

The *vocabulary* effect (v14) was significant when applied to the minitalk texts. However, the result was in a direction opposite from

expected. The more multisyllabic words (involving three or more syllables) in the text, the easier were the items associated with that text. This result is counterintuitive and different from the result reported in Freedle and Kostin's (1993) reading study. We do not have a clear explanation for this unusual result. The development of an alternative vocabulary score (v15) failed to clarify the problem. However, vocabulary did not emerge in any of our regression analyses and for this reason has not affected our conclusion regarding construct validity.

Serial position effects were strongly represented among the predictor variables (mm2, mm4, mm12, ii2, ii3, ii5, ss4, ss5). In general, items dealing with information presented early in the minitalk (mm2, mm4, ii2, ss4) tended to be easier items – this finding applied to main-idea items, inference items and supporting-idea items. Also, if the relevant overlap information occurred in the last sentence of the minitalk, this pattern was associated with easier items (ii5). Finally, if the relevant information for correctly answering an item was found in the middle of a minitalk, this pattern was associated with harder items (ii3, ss5). Similar positional effects were found in the Freedle and Kostin (1993) TOEFL reading study.

Lexical overlap was also prominently represented in our data (ii13, ss11, ss15). In fact ss15 was our best single predictor of minitalk item difficulty. In general, the more lexically 'attractive' a correct option was in comparison with the most chosen incorrect option, the easier the item (ii13, ss15). (Because these variables were specifically designed for the listening study, there were no comparable scores available in the Freedle and Kostin (1993) reading study.) Other variables common to both reading and listening, however, can be compared. We note that while ss11 (percent lexical overlap between words in the correct option and words in the relevant text sentence) was significant for minitalks ($r = -.142$, $n = 302$), the result was significantly stronger for reading ($r = -.350$, $n = 213$), $p < .01$; see McNemar, 1955: 148. We attribute the significant decline in strength of ss11 for minitalks to a memory problem. That is, it is much easier to be influenced by the matching words in the relevant reading text (which is fully visible to the reader) than to remember the exact words in the relevant section after one has listened to a minitalk passage. We conclude that there are substantial lexical overlap effects operating in both listening and reading but the cognitive underpinnings for these effects are probably different.

Redundancy effects were represented among the predictor variables (mm14, ii12, ss13). A close examination of the correlational effects suggests that a somewhat complex redundancy result has occurred for the minitalks. The creation of redundancy (ii10, ss12) by simply

repeating information (typically by paraphrasing) within the text did not appear to significantly influence minitalk item difficulty, at least when the zero-order correlational results were examined. One of the 'complementary' redundancy variables (ss13) was significant – as predicted it was associated with more difficult items, not easier, because it involves more cognitive work to appreciate how two related text sentences complement each other.

There is a possible explanation for why simple repetition of information (e.g., ii10, ss12) failed to yield significant correlations. Inquiry about the test development process showed that minitalk text redundancy is intentionally created in several circumstances when, for example, supporting detail items are tested, or when the content is expected to be particularly difficult for examinees (e.g., when a particular word might be relatively unfamiliar or when special background knowledge may be called upon – knowledge judged to be relatively less familiar to non-native English listeners). The intentional creation of text redundancy by test development staff, especially for selected portions of text having subjectively hard content, may have negated the possibility of obtaining a significant correlation for variables such as ii10.

We should mention that two significant variables (ii13, ss15) already described above under the category of lexical overlap have components involving redundancy in the following sense: when a word is repeated several times in the minitalk this is a type of redundancy effect. When such a repeated word appears in the correct option, this appears to make the option attractive which, in turn, tends to facilitate getting the item correct.

We see that the particular ways in which different kinds of redundancy contribute to listening item difficulty represent a complex result. In general, though, the redundancy results noted above are interpreted as broadly supportive of related findings among ESL researchers in the study of listening comprehension (Chiang and Dunkel, 1992; Parker and Chaudron, 1987).

Subject matter (topical) effects. Most of the academic subject matters (except the social sciences, v19) were associated with harder listening items (see v17, v18, v20, v21 that represent physical science, biological science, humanities, and arts, respectively). All of the non-academic subject matters were associated with easier listening items (see the composite score v27). For reading (see Freedle and Kostin, 1993), the social sciences were represented by items that were significantly harder than the other subject matters, while the humanities were associated with items that were significantly easier. Thus, reading and listening yielded exactly opposite findings with respect to two subject matters: social sciences and humanities.

To help explain these differences between reading and listening, inquiries about the TOEFL test assembly process revealed that preparation for the reading and listening stimuli are markedly different. The reading materials are sampled from materials and topics likely to be encountered by native speakers. In contrast, the listening stimuli are fictional, scripted out excerpts intended to represent the type of speech actually encountered, say, in university classrooms or when attending other university functions.

Additional correlational findings. An item variable, v5, that recorded the position of the correct answer showed that if the correct option occurred late among the list of options, this pattern was associated with easier items. There were three significant item type variables (tt1, tt2, tt3). Main idea items (tt2) and inference-application items (tt1) tended to be easy items while inference items (tt3) tended to be hard items.

There are additional findings. Contrary to hypothesis, the more filled ('umm', 'er') or long pauses (…) in the minitalk, the harder were the listening items associated with that passage. Apparently any disruption in the coherent reception of a speaker's ideas made it harder to process the message. If there were content words in the lead-in that overlapped with the minitalk text, these content words were associated with easier items. Such words in the lead-in probably acted as advance organizers for the material that followed and hence facilitated the comprehension of the items associated with the passage. Finally, as the number of focal shifts in the minitalk (v52) increased, the easier were the items associated with such minitalks. We interpret this pattern to mean that since most items focus on the topical content of the minitalk (and less on the introductory or ending episodes – which involve focal shifts), the inclusion of more nontopical episodes shortens the critical content that has to be attended to hence leading to generally easier items.

4 General statement regarding significant correlates of listening comprehension

Broadly conceived, we have shown that in many places our scored variables reveal several ways in which TOEFL's reading and listening sections are similar; a few differences were also noted and we have explored several ideas as to why these differences may exist.

As already suggested, based on the variables that are significant, one might be tempted to conclude, without further analyses, that the correlational results in Appendix B appear to provide some support for the construct validity of the TOEFL minitalk section. That is, in terms of the 'level' of evidence for construct validity, the zero-order

correlational results point towards favourable, though weak, evidence for construct validity. This conclusion would seem to hold whether we examine just the text/item overlap variables or examine just the text variables. For the 302 sample (Appendix B) there were 10 text or text-by-item overlap variables that had correlations with equated delta that were larger in magnitude than the best pure-item predictor. However, this conclusion concerning construct validity should still be considered tentative because some of the variables are significantly intercorrelated and, furthermore, the correlations are rather low in absolute magnitude – only 5 of the 43 significant variables in Appendix B meet or exceed a magnitude of .20. Regarding these small magnitudes, it is interesting that a parallel-processing model of language comprehension such as that proposed by Just and Carpenter (1987: 279–81) 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.)

We seek below to provide some evidence favouring a 'higher' level of construct validity. To accomplish this demonstration, regression methods will be used.

5 Regression analysis

Criteria for admitting variables into a stepwise regression. The following criteria were used for admitting variables into the regression. All non-superscripted variables listed in Appendix A were available for possible selection. Each new variable that was admitted into the solution had to yield a significant individual t value ($p < .05$); also, for the final step, the new t values for all previously admitted variables had to be significant, and finally the variables had to be consistent with the directionality (if any) predicted by our literature review. If the next variable admitted showed a non-significant t (or was contradictory to our literature review), then the previous stepwise solution was considered the definitive one.

We will explore below the implications for construct validity of both stepwise and hierarchical (Cohen and Cohen, 1983) regression analyses for our combined item sample.

6 Stepwise regression analysis of 302 reading items involving three item types (main ideas, inferences and supporting ideas)

This sample was constructed to make it more similar, in its item type structure, to the Freedle and Kostin (1993) study of TOEFL reading items. That is, the 302 item sample has only main ideas, inferences and supporting idea items as in the Freedle and Kostin (1993) reading item study. As we see from Table 1, a regression analysis yielded 12 significant predictors of the difficulty of minitalk items (equated deltas) for three item types. The overall F (12,289) = 11.97, $p < .0001$; the multiple-R = .576, the R-squared = .332, which accounts for 33.2% of the item variance. Table 1 lists the 12 significant variables in the order that they emerged from the regression analysis. Although space limitations prohibits presenting the details, Freedle and Kostin (1996) also demonstrated that this significant regression result cannot be attributed to chance effects associated with a large number of predictor variables.

The best predictors are ss15 (relative attractiveness of correct vs. incorrect options) and v27 (nonacademic content). There are five positional variables that are significant for the 302 sample (ss4, mm12, ii4, ii3, mm2).

Table 1 Multiple regression results for one of the TOEFL minitalk item samples

Variable		Sample (n = 302)	
		beta weight	p value
ss15	Supporting, correct more attractive than most chosen incorrect option	–.235	.001
v27	Nonacademic topic	–.306	.001
tt2	Main idea	–.227	.001
v10	Number of words in incorrects	.120	.020
mm12	Incorrect option lexical overlap among first 20 words plus lexically related words	.141	.001
ss4	Supporting, first three sentences	–.145	.001
v29	List rhetorical structure	–.122	.020
ii3	Inference, critical information only in middle of text	.244	.001
ii4	Inference, critical information in second and third sentence from end	.168	.001
v19	Social science content	–.148	.001
ii12	Inference, critical information in only one text sentence	.099	.040
mm2	Main idea, critical information in second text sentence	–.095	.050

Notes:
- Variables are listed in the order that they emerged from the linear regression.
- n = 302 items analyzed (3 item types excluding inference-application).
- Multiple regression results: F (12,289) = 11.97, $p < .0001$, multiple-R = .576, and R-squared = .332.

The 302 sample has a pure item variable (v10) as a significant contributor to the regression result. But the rest of the predictors are either text variables (v27, v19, v29), text/item overlap variables (ss15, ss4, mm12, ii4, ii3, mm2, ii12) or a special text-by-item overlap variable involving item type (tt2).

7 Hierarchical regression of 302 sample

Before we can reach any further conclusions concerning the construct validity of this sample (with disclaimers as above) we need to report the results of a hierarchical regression analysis. The first step extracted all the eleven pure item variables; this step accounted for a significant 7.8% of the item difficulty variance as shown by the F test: $F\ (11,290) = 2.22$, $p < .01$. The second step in the hierarchical procedure extracted all remaining text and text-associated predictors. This step augmented the variance predictability by an additional 37.5%. This increase was significant (see Cohen and Cohen, 1983: 145–47): $F\ (62,229) = 2.53$, $p < .01$.

8 Discussion of results

The correlational analyses of the 302 and 337 item samples, and the stepwise and hierarchical regression analyses of the 302 item sample allows us to reach a general conclusion concerning construct validity. The most important predictors are the text and text/item overlap variables; the pure item variables play a minor role in determining minitalk item difficulty. The correlational results provided support for the first 'level' of construct validity inasmuch as several text and text-by-item overlap variables were significant. The stepwise regression results provided evidence for a higher second 'level' of validity inasmuch as the text and text-by-item variables were more potent predictors than any of the item variables. Finally, the hierarchical regression result provided evidence for a third 'level' of construct validity inasmuch as the amount of variance accounted for by the text and text-by-item variables significantly exceeded the variance due solely to the item variables even after the variance due to all the item variables had been extracted. Hence, consistent with the arguments advanced in our introduction (also see Freedle and Kostin, 1994), there appears to be some evidence favouring the construct validity of the listening minitalk items.

We caution the reader that a replication of the exact set of predictors that emerged from these several regression analyses cannot be guaranteed if a new data set were to be analysed; this is partly a consequence of the large number of predictor variables that have been

used in our study. However, we would expect robust replication of the broad classes of variables: we expect pure item variables to account for very little of the item difficulty variance while substantial variance is expected to be accounted for by the text and text-associated (text-by-item overlap) variables. That is, we place greater stress on the replicability at the level of classes of predictor variables and less stress at the level of individual predictor variables.

III Conclusion

The construct validity of minitalks

Several analyses have been presented that provide some support for the relevance of variables, isolated from our earlier literature review, in predicting listening item difficulty. The broad array of predictor variables that were isolated from the experimental literature on language comprehension has revealed many underlying similarities in the skills being tapped by TOEFL's reading and listening (minitalk) items.

There are additional implications of our results. In particular, we have interpreted earlier critiques of multiple-choice reading tests as implying that text and text-by-item overlap variables should play a minor role in predicting minitalk item difficulty, with the added implication, according to our reading of Drum *et al.* (1981), that item variables might play a major role. By and large our stepwise and hierarchical regressions provide evidence that suggest just the opposite conclusion: pure item variables appear to play a minor role, while text and text associated (text/item overlap) variables play by far the major role in accounting for minitalk item difficulty. Consistent with the arguments advanced earlier, especially the fact that our hierarchical regression supports the idea that pure item variables play a weak role in influencing item difficulty, we are led to conclude that there is modest evidence to support the claim that the TOEFL minitalks and their associated items appear to be construct valid.

a Similarity of TOEFL's listening (minitalk) and reading sections. We have explored a few of the similarities and differences between TOEFL's reading and listening (minitalk) passages and their associated items.

We have noted that while the psycholinguistic literature has emphasized the empirical similarities between reading and listening measures, there is nevertheless some evidence that TOEFL's reading and listening items actually fall on two distinct factors (see Hale *et al*., 1989); while distinct, the listening and nonlistening (including

reading) section scores were highly intercorrelated, varying from .79 to .84, depending upon language group membership, with a median of .82 (also see Carroll, 1993: 178–81). Our earlier comparisons of the few differences that exist between TOEFL reading and minitalk items for particular variables (e.g., our results above concerning different magnitude of effects of lexical overlap) presumably help to explain why there is evidence for these two separate factors. An additional example, that underlies factor differences, may be the following: we have noted that while minitalks include text passages on nonacademic as well as academic topics, reading materials are restricted to academic topics. Yet the numerous strong similarities among the many variables that were compared for both reading and the minitalks also suggests why the two factors are so strongly intercorrelated. In other words, our focus on analysing the content and structure of individual test items (by defining specific predictor variables) could potentially help to pinpoint and explain broad factor analytic results.

We have noted that many of the variables that successfully accounted for reading-item difficulty also accounted for listening-item difficulty which is consistent with the view that both these receptive skills are measures of a general underlying language comprehension ability. The evidence suggests that there are similar effects for reading and listening (minitalks) associated with respect to negations, some frontings, several of the rhetorical organizers (e.g., list and problem/solution), the effects of concreteness vs. abstractness of the text, and serial position effects. Presumably some or all of these similarities help in part to account for the strong positive correlation between listening and nonlistening factors reported by Hale *et al.* (1989).

b Limitations of the current study. It would be useful to conduct additional empirical work on the TOEFL minitalk passages by developing additional variables that listening researchers (see Rost, 1990) have specifically found to be helpful in explaining the relative difficulty of various listening materials. Researchers who carry out such additional work could explore with greater thoroughness other aspects of construct validity regarding TOEFL's listening materials not covered by the current study.

c Broader issues of validity: evaluating language tests using theories of communicative competence. The main purpose of our study has focused upon a very special idea of construct validity: To be valid, a multiple-choice test of listening (or reading) must demonstrate sensitivity to the information in the text passages. But there are obviously more complex validity issues that could be addressed. Chapelle *et al.*

(1997) have outlined a model for evaluating some broader issues of validity using a communicative-competence framework. Their model attempts to integrate the seminal work of Hymes (1972), Canale and Swain (1980) and Bachman (1990). Using the framework of Chapelle *et al.*, it may be useful for us to augment and reclassify the system of predictor variables used in our study to more directly reflect the four main categories of interest from a communicative-competence perspective: discourse competence, grammatical competence, strategic competence and sociolinguistic competence. Hierarchical regression techniques could then be applied to the enlarged system of variables in order to study the *relative* contributive of grammatical, sociolinguistic, discourse and strategic competences in the prediction of minitalk (or reading) item difficulty. One can hypothesize that such a methodology would potentially allow one to begin to address more complex issues of validity for current formats of the TOEFL test. Of course, the more general intent of the Chapelle *et al*. model is to lay a foundation for constructing new language tests that meet the growing needs of language-teaching specialists with respect to evaluating not only a test's construct validity, but also its relevance, utility, value implications and social consequences.

Acknowledgements

We want to thank John Upshur, Paul Angelis, Thom Hudson, Carol Chapelle, Linda Schinke-Llana and Fred Davidson of the TOEFL Research Committee for their helpful comments on an earlier draft of this paper. Additional comments by Gary Buck, Felicia DeVincenzi, Ann Gordon and Larry Stricker are also gratefully acknowledged.

IV References

Abrahamsen, E. and **Shelton, K.** 1989: Reading comprehension in adolescence with learning disabilities: semantic and syntactic effects. *Journal of Learning Disabilities* 22, 569–72.

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bhasin, J.I. 1990: The demands of main idea tasks in reading comprehension tests and the strategic responses of bilingual poor comprehenders. Unpublished Dissertation, Columbia University Teachers College.

Blau, E.K. 1990: The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly* 24, 746–53.

Buck, G. 1990: The testing of second language listening comprehension. Unpublished doctoral dissertation. University of Lancaster, Lancaster.

Canale, M. and **Swain, M.** 1980: Theoretical bases of communicative

approaches to second language teaching and testing. *Applied Linguistics* 1, 1–47.

Carpenter, P.A. and **Just, M.A.** 1975: Sentence comprehension: a psycholinguistic processing model of verification. *Psychological Review* 82, 45–73.

Carroll, J.B. 1993: *Human cognitive abilities: a survey of factor-analytic sutdies*. New York and Cambridge: Cambridge University Press.

Chapelle, C., Grabe, W. and **Berns, M.** 1997: *Communicative language proficiency: definition and implications for TOEFL-2000*. TOEFL Monograph Series MS-10. Princeton, NJ: Educational Testing Service.

Chaudron, C. and **Richards, J.** 1986: The effect of discourse markers on the comprehension of lectures. *Applied Linguistics* 7, 113–27.

Chiang, C.S. and **Dunkel, P.** 1992: The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly* 26, 345–74.

Cohen, J. and **Cohen, P.** 1983: *Applied multiple regression: correlational analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Drum, P.A., Calfee, R.C. and **Cook, L.K.** 1981: The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly* 16, 486–514.

Farr, R., Pritchard, R. and **Smitten, B.** 1990: A description of what happens when an examinee takes a muiltiple-choice reading-comprehension test. *Journal of Educational Measurement* 27, 209–26.

Freedle, R. and **Fellbaum, C.** 1987: An exploratory study of the relative difficulty of TOEFL's listening cpmprehension items. In Freedle, R. and Duran, R., editors. *Cognitive and linguistic analyses of test performance*. Norwood, NJ: Albex, 162–92.

Freedle, R., Fine, J. and **Fellbaum, C.** 1981: Predictors of good and bad essays. Paper presented at the annual Georgetown University Roundtable on languages and linguistics, Washington DC.

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

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

Freedle, R. and **Kostin, I.** 1993: the prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* 10, 133–70.

Freedle, R. and **Kostin, I.** 1994: Can multiple-choice reading tests be construct-valid? *Psychological Science* 5, 107–10.

Freedle, R. and **Kostin, I.** 1995: The prediction of SAT reading comprehension item difficulty for two item types: inference and explicit statements. Unpublished report. Princeton, NJ: Educational Testing Service.

Freedle, R. and **Kostin, I.** 1996: *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: implications for construct*

validity. TOEFL Research Report No. 56; also available as RR-96-29. Princeton, NJ: Educational Testing Service.

Graves, M. 1986: Vocabulary learning and instruction. In Rothkopf, E., editor, *Review of research in education*, Vol. 13. Washington DC: Amercian Educational Research Association.

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

Hale, G., Rock, D. and **Jirele, T.** 1989: *Confirmatory factor analysis of the Test of English as a Foreign Language*. TOEFL Research Reports, Report 32. Princeton, NJ: Educational Testing Service.

Hare, V., Rabinowitz, M. and **Schieble, K.** 1989: Text effects on main idea comprehension. *Reading Research Quarterly* 24, 72–88.

Hymes, D. 1972: *Towards communicative competence*. Philadelphia, PA: Pennsylvania University Press.

Just, M. and **Carpenter, P**. 1987: *The psychology of reading and language comprehension*. Boston, MA: Allyn and Bacon.

Kieras, D.E. 1985: Thematic processes in the comprehension of technical prose. In Britton, B. and Black, J., editors, *Understanding expository text*. Hillsdale, NJ: Erlbaum, 89–107.

Kintsch, W. and **Kozminsky, E.** 1977: Summarizing stories after reading and listening. *Journal of Educational Psychology* 69, 491–99.

McNemar, Q. 1955: *Psychological statistics*, New York: Wiley.

Meyer, B. 1985: Prose analysis: purposes, procedures, and problems. In Britton, B. and Black, J., editors, *Understanding expository text*. Hillsdale, NJ: Erlbaum, 11–64.

Meyer, B. and **Freedle, R.** 1984: The effects of different discourse types on recall. *American Educational Research Journal* 21, 121–43.

Newsome, R.S. and **Gaite, J.H.** 1971: Prose learning: Effects of pretesting and reduction of passage length. *Psychology Reports* 28, 128–29.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K. and **Bent, D.** 1975: *Statistical packages for the social sciences* (2nd edn). New York: McGraw-Hill.

Paivio, A. 1986: *Mental representations*. New York and Oxford: Oxford University Press.

Parker, K. and **Chaudron, C.** 1987: The effects of linguistic simplification and elaborative modifications on L2 comprehension. *University of Hawaii Working Papers in EST. 6*, 107–33. Cited in Chiang & Dunkel, 1992.

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

Sticht, T. and **James, J.H.** 1984: Listening and reading. In Pearson, P.D., Barr, R., Kamil, M. and Mosenthal, P., editors, *Handbook of reading research*. New York and London: Longman, 293–317.

Appendix A List of coded variables and their definitions

Item Variables

Variables for item's stem

v1 – *Words in stem* – number of words in stem

v2 – *Negative stem* – number of negations in stem (no, not, etc.)

v3 – *Fronted stem* – number of fronts in stem including use of clefts (*It* was cold); phrases or clauses preceding subject of main clause (*In truth*, he fled. *But, because he slept*, he revived. Etc.)
v4 – *References in stem* – sum intertextual referentials either within clauses (*He*, the general, fought), across clauses (John fell; *that* hurt) and/or extratextual reference (*One* needs to study).

Variables for item's correct option (examples: see v1 to v4)
v5 – *Answer position* – the ordinal position of the correct option
v6 – Words correct – number of words in correct option[a]
v7 – *Negative correct* – number of negatives in correct option
v8 – *Fronting correct* – number of frontings in correct option
v9 – *Reference correct* – number of referentials in correct option

Variables for item's incorrect options (examples: see v1 to v4)
v10 – *Words incorrects* – total number words in all incorrect options
v11 – *Negative incorrects* – total negations used in all incorrects
v12 – *Fronted incorrects* – total frontings used in all incorrects
v13 – Reference incorrects – total referentials used in all incorrects[a]

Text Variables

Vocabulary variable for text
v14 – *Vocabulary* – number words with more than two syllables among first 100 words
v15 – Vocabulary variety – number different words used with more than two syllables among first 100 words[a]

Concreteness/abstractness of text
v16 – *Concreteness* – text focused primarily on concrete imagery (e.g., apple, car) as opposed to abstract imagery (justice, truth) (5=mostly concrete; 1=mostly abstract)

Subject matter variables of text
v17 – *Physical science* (1=yes; 0=no)
v18 – *Life science* (1=yes; 0=no)
v19 – *Social science* (1=yes; 0=no)
v20 – *Humanities* (1=yes; 0=no)
v21 – *Arts* (1=yes; 0=no)
v22 – Miscellaneous topic (1=yes; 0=no)[a]
v23 – Sightseeing topic (1=yes; 0=no)[a]
v24 – Extracurricular topic (1=yes; 0=no)[a]
v25 – University procedure topic (1=yes; 0=no)[a]
v26 – Academic procedure topic (1=yes; 0=no)[a]
v27 – *Sum of Nonacademic topics* (sum of v22 through v26)

Type of rhetorical organization – (see Meyer, 1985)
v28 – *Argument* (author favors one of several viewpoints)
v29 – *List/describe* (rhetorical structure uses list of text elements or a descriptive 'list')
v30 – *Cause* (author uses a cause/effect type text organization)
v31 – *Compare* (author uses a compare/contrast type organization)

v32 – *Problem/solution* (author uses problem/solution organization). (For v28 through v32 a 1=yes; 0=no)

Coherence of text

v33 – *Coherence* (3=max coherence; 0=min coherence). High coherence means elements of opening text sentence densely represented throughout text, etc.

Lengths of various text segments

v34 – *Text words* (number of words in text)
v35 – Text sentences (number of sentences in text)[a]
v36 – *Text's average sentence length*

Occurrence of different text frontings (fronting examples: see v3)

v37 – Percent fronted text clauses[a]
v38 – *Frequency fronted text clauses*
v39 – *Frequency combinations of fronted text structures*
v40 – *Frequency of text clefts* (*It* is a shame. *There* are many.)
v41 – *Longest fronted run* (max number consecutive fronted text clauses)

Text questions

v42 – *Text questions* (number of interrogative text sentences)

Text referentials (frequency of occurrence per variable)

v43 – *Reference within text clauses* (*He*, the boss, left.)
v44 – *Reference across text clauses* (George came. Then *he* left.)
v45 – *Frequency special reference* (*One* can always hope.)
v46 – Reference sums (sum of v43 through v45)[a]

Text negations (examples: no, not, *un*clear, point*less*)

v47 – *Text negatives* (total number of negations in text)

Special Listening Text Variables

v48 – *Frequency words emphasized in text* (number words emphasized in typescript of talk)
v49 – *Frequency pauses* (total number of filled pauses 'um' 'er' and unfilled pauses … in typescript of talk)
v50 – *No. words in lead-in* (words in the spoken direction prior to talk proper. E.g., 'Listen to the following talk.')
v51 – *Lead-in overlaps with text* (overlap or not of any content word in lead-in with the text proper. E.g., 'Listen to the talk on art,' where *the word* 'art' may also occur anywhere in the talk proper.)
v52 – *No. of focal shifts* (Number of parts of the minitalk involving a focal shift. E.g., 'I'm the new teacher. The history of Rome is quite involved …. Class is now dismissed.' This example consists of three episodes or 'focal shifts' – the introductory statement, the lecture proper, and the dismissal statement.
v53 – *No. words prior to 'lecture' content* (using the example in v52, it's the number of words that precede the 'history' lecture, which is 4 in this case).

Text/Item Overlap Variables

Text/item overlap variables applicable only to main idea information

mm1 – *Main idea in 1st text sentence*
mm2 – *Main idea in 2nd text sentence*
mm3 – Main idea in 3rd text sentence[a]
mm4 – *Main idea early in text* (among 1st three sentences)
mm5 – *Main idea in middle of text*
mm6 – Main idea near text end (2nd or 3rd sentence from the end)[a]
mm7 – *Main idea in last text sentence*
mm8 – *Main idea implicit*
mm9 – First 20 text words, lexical match with correct option (Lexical matches between first 20 text words and word(s) in correct option)[a]
mm10 – *First 20 words, lexical match including related words.* Same as mm9 except includes related words ('Related' means all morphological variants – plurals, past tense, etc.)
mm11 – First 20 text words, lexical match with most chosen incorrect[a]
mm12 – *First 20 text words, lexical match with most chosen incorrect including related words* (e.g., work/works)

Special listening main idea variable

mm13 – Frequent word count (suppose 'sleep' and 'home' occur in correct option; further suppose 'sleep' occurs 7 times in text while 'home' occurs 2 times; only the frequency of the greater count, here '7' is recorded)[a]
mm14 – *Frequent word count plus related words* (same as mm13 but includes 'related' words such as sleep/sleepily)

Text/item overlap variables applicable to inferences

ii1 – *Unique word same sentence* (content word in stem corresponds to unique word in text and relevant information for answering item is in same text sentence)
ii2 – *Information in general beginning of text* (relevant information is among first three text sentences)
ii3 – *Information middle of text* (relevant information in middle)
ii4 – *Information near end* (relevant information in 2nd or 3rd sentence from end of talk)
ii5 – *Information in last sentence* (last sentence contains the critical information)
ii6 – *Words before critical information* – number of words that precede the start of the critical text sentence containing the relevant information for answering the item.
ii7 – *Number words in key text sentence* – Words in relevant sentence for answering the inference item. If two sentences are critical, enter the number of words in the longer sentence.
ii8 – Percent lexically matched words – percent words in correct answer that overlap with words in key text sentence(s)[a]
ii9 – *Percent lexically matched words plus related words* (same as in ii8 including lexically related words (e.g., shirt/shirts)
ii10 – *Information Redundant* (see introduction for full definition). If answering an inference item depends upon two text sentences, and both sentences are close paraphrases enter a '1' for this variable.

ii11 – *Information Complementary* (see introduction for full definition). If answering an inference item depends upon two (or more) text sentences and one sentence presents information complementary to the other sentence(s), then enter a '1' for this variable.

ii12 – *Relevant information is in only one text sentence.*

Special listening inference variable

ii13 – *Attractiveness of the correct option versus the most chosen incorrect option* (see introduction for further discussion). This variable consists of two parts. The first part evaluates the content words common to the correct option and the text and compares it to the content words common to the most favored incorrect option and the text. If the correct option has the greater number of shared words, this first part gets scored as +1, if the incorrect has the greater number of shared words the score is −1, if the two are equal the score is 0. The second part examines which *particular* word from the correct option and which word from the most chosen incorrect option overlaps most often with the text. Assign a score of +1 if the correct option has a greater overlap with text than the most favored incorrect option. Score −1 if the incorrect option has a greater overlap score. And assign a 0 if they are equal. Add the two parts and enter this as the final score.

Text/item overlap variables applicable to supporting ideas

ss1 – *Unique word same sentence* (content word in stem corresponds to unique word in text, and, relevant information is in same sentence)

ss2 – *Unique word, earlier sentence* (content word in stem corresponds to a unique word in text, but relevant information is in an earlier text sentence)

ss3 – *Key stem word(s) occur in multiple places in text*

ss4 – *Critical information in general beginning* (first three text sentences)

ss5 – *Critical information in middle of text*

ss6 – *Critical information near end* (2nd or 3rd sentence from end)

ss7 – *Critical information in last text sentence*

ss8 – *Number words before critical information*

ss9 – *Number words in key text sentence*

ss10 – Percent lexically matched words[a]

ss11 – *Percent lexically matched words plus related words*

ss12 – *Information Redundant* (same as ii10 above except here it applies to a supporting item rather than an inference item)

ss13 – *Information Complementary* (same as ii11 above except here it applies to a supporting item rather than an inference item)

ss14 – Relevant *information in just one text sentence*

Special listening supporting idea variable

ss15 – *Attractiveness of correct* (same as ii13 above except here it applies to a supporting item rather than an inference item)

Item Types (special types of text-by-item overlap variables)

tt1 – *Inference application* (e.g., 'Who was the speaker?')
tt2 – *Main idea* (e.g., 'What is the main topic of the talk?'
tt3 – *Inference* (e.g. 'According to the speaker one can infer …')
tt4 – *Supporting idea* (e.g., 'What is the boiling point of lead?')

Dependent Variable

v54 – *Item difficulty* (equated delta, a standardized measure of difficulty that allows combining item information across several TOEFL test forms)

Note: The superscript 'a' designates variables that were excluded from all analyses. See method section for reasons for exclusion. Slightly more detailed descriptions of the variables listed in Appendix A are available in Freedle and Kostin (1996).

Appendix B Significant correlations of variables with item difficulty (equated delta)

Significant variables		Sample size	
		(n = 302)	(n = 337)
Item Variables			
v1	Stem: words in stem	.133***	.175***
v4	Stem: sum of referentials	–.062ns	–.093**
v5	Correct: position of correct answer	–.085*	–.090**
v7	Correct: negatives	.125**	.130***
v9	Correct: referentials	.143***	.175***
v10	Incorrect: no. words	.146***	.182***
v11	Incorrect: no. negatives	.134***	.148***
Text Variables			
v14	Vocabulary	–.157***	–.110**
v16	Concreteness	–.067ns	–.080*
v17	Physical science	.087*	.082*
v18	Biological Science	.127**	.120***
v19	Social science	–.093**	–.096**
v20	Humanities	.091*	.105**
v21	Arts	.116**	.142***
v27	Nonacademic topics	–.219***	–.224***
v28	Argue	.096**	.096**
v29	List rhetorical	–.200***	–.218***
v31	Compare rhetorical	.142***	.139***
v32	Problem/Solution rhetorical	.130***	.145***
v41	Longest run of fronts	.111**	.114**
v45	Outside text referentials	–.133***	–.147***

v47	Negatives in text	.117**	.108**
v49	No. filled or long pauses	.094**	.079*
v51	Lead-in's content words overlap text	–.157***	–.134***
v52	No. focal shifts in minitalk	–.178***	–.168***
	Text/Item Overlap Variables		
mm2	Main idea in 2nd sentence	–.129**	–.118**
m4	Main idea: general beginning	–.084*	–.078*
mm12	Main idea, 1st 20 words, lex. match for incorrects including lex. related wds	.152***	.138***
mm14	Main idea, freq. word count including lexically related words	–.086*	–.079*
ii2	Inference, info. among 1st three text sent.	–.106**	–.086*
ii3	Inferences, info only in middle text	.114**	.106**
ii5	Inference, info. in last sent.	–.079*	–.075*
ii12	Inference, info. in only one sent.	.087*	.080*
ii13	Inference, correct more attractive than incorrect	–.125**	–.115**
ss4	Supporting, general beginning	–.156***	–.132***
ss5	Supporting, info. only in middle text	.120**	.114**
ss8	Supporting, no. wds before info. begins	.165***	.156***
ss11	Supporting, % lex. overlap inc. lex. related	–.142***	–.131***
ss13	Supporting, information in more than one complementary sentence	.076*	.079*
ss15	Supporting, correct more attractive than incorrect	–.258**	–.238**
tt1	inference application (item type)	NA	–.232***
tt2	Main idea (item type)	–.195***	–.151***
tt3	Inference (item type)	.200***	.226***

Notes:
*** significant at $p < .01$, 2-tailed
** significant at $p < .05$, 2-tailed
* significant at $p < .05$, 1-tailed
ns not significant, $p > .05$, 2-tailed
NA not applicable ($n = 302$ does not include inference-application items)
If a variables was not significant for the 2-tailed test but where direction was predicted we applied a 1-tailed test. A positive correlation indicates that the presence (or more) of the variable makes the item harder while a negative correlation would mean that the presence (or more) of the variable makes the item easier.