TOEFL®

Monograph Series

MS - 19
SEPTEMBER 2000

TOEFL 2000 Listening Framework: A Working Paper

Isaac Bejar
Dan Douglas
Joan Jamieson
Susan Nissan
Jean Turner

ETS™ Educational Testing Service

TOEFL 2000 Listening Framework: A Working Paper

Isaac Bejar
Dan Douglas
Joan Jamieson
Susan Nissan
Jean Turner

Educational Testing Service
Princeton, New Jersey
RM-00-7

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

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

E-mail: toefl@ets.org

Web site: http://www.toefl.org

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other TOEFL® test development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at Educational Testing Service (ETS®) will evolve into the 21st century. As a first step the TOEFL program recently revised the Test of Spoken English (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, takes advantage of the new forms of assessments and improved services made possible by computer-based testing while also moving the program toward its longer-range goals, which include

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 are the working papers that lay out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, describes the process by which the project will move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extend the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). As such, the current frameworks do not yet represent a final test model. The final test design will be refined through an iterative process of prototyping and research as the TOEFL 2000 project proceeds.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program Office
Educational Testing Service

Abstract

This monograph is an initial attempt to define listening as it will be measured in the TOEFL 2000 test, within the framework delineated in *TOEFL 2000 Framework: A Working Paper* (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000).

This monograph is comprised of six sections. After a brief introduction, an overview of academic listening is presented, which outlines theory and research in the areas of listening in general and academic listening in particular. The third section describes theory and research on the variables that characterize academic listening and that might drive the difficulty of a listening task. The fourth section addresses technological issues for the listening section of the TOEFL 2000 test. A research agenda is presented in the fifth section of the monograph, and the final section describes the features that distinguish the TOEFL 2000 test from its predecessors.

Key words: Listening proficiency, communicative competence, linguistic knowledge, academic listening tasks

Acknowledgments

The authors gratefully acknowledge the support and suggestions of the TOEFL Committee of Examiners, the members of the TOEFL 2000 Committees working on the speaking, reading, and writing frameworks, and the ETS staff for their helpful comments and suggestions.

Table of Contents

List of Tables

Figure

1. Introduction

This report is an initial attempt to define listening as it will be measured in the TOEFL 2000 test. After reviewing the literature on listening comprehension, we identified the variables that are most relevant to the skill of listening comprehension. We propose to research these variables to identify those which have the greatest impact on the difficulty of listening tasks. This information will provide a framework for the development of trial items, test specifications, and research into the validity of the listening section of the TOEFL 2000 test.

This report is divided into six major sections. The next section, an overview of academic listening, outlines theory and research in the areas of listening in general and academic listening in particular. The third section describes theory and research on the variables that characterize academic listening and that might drive the difficulty of a listening task or item. The discussion of each of the variables concludes with a proposal for how to operationalize these variables in a testing context. (The variables are summarized in Appendix A.)

The fourth section addresses technological issues for the listening section of the TOEFL 2000 test. A research agenda is presented in the fifth section of the report, which includes a framework for validation across the different skills. The final section analyzes the features that make the TOEFL 2000 test a better instrument than its predecessors. Some sample tasks are presented in Appendix C.

2. Conceptualizing Listening Proficiency

In this section, theory and research in the nature of listening in general are briefly outlined. This provides a foundation for the next section, which discusses factors that reportedly characterize and affect academic listening and considers how these factors might be operationalized.

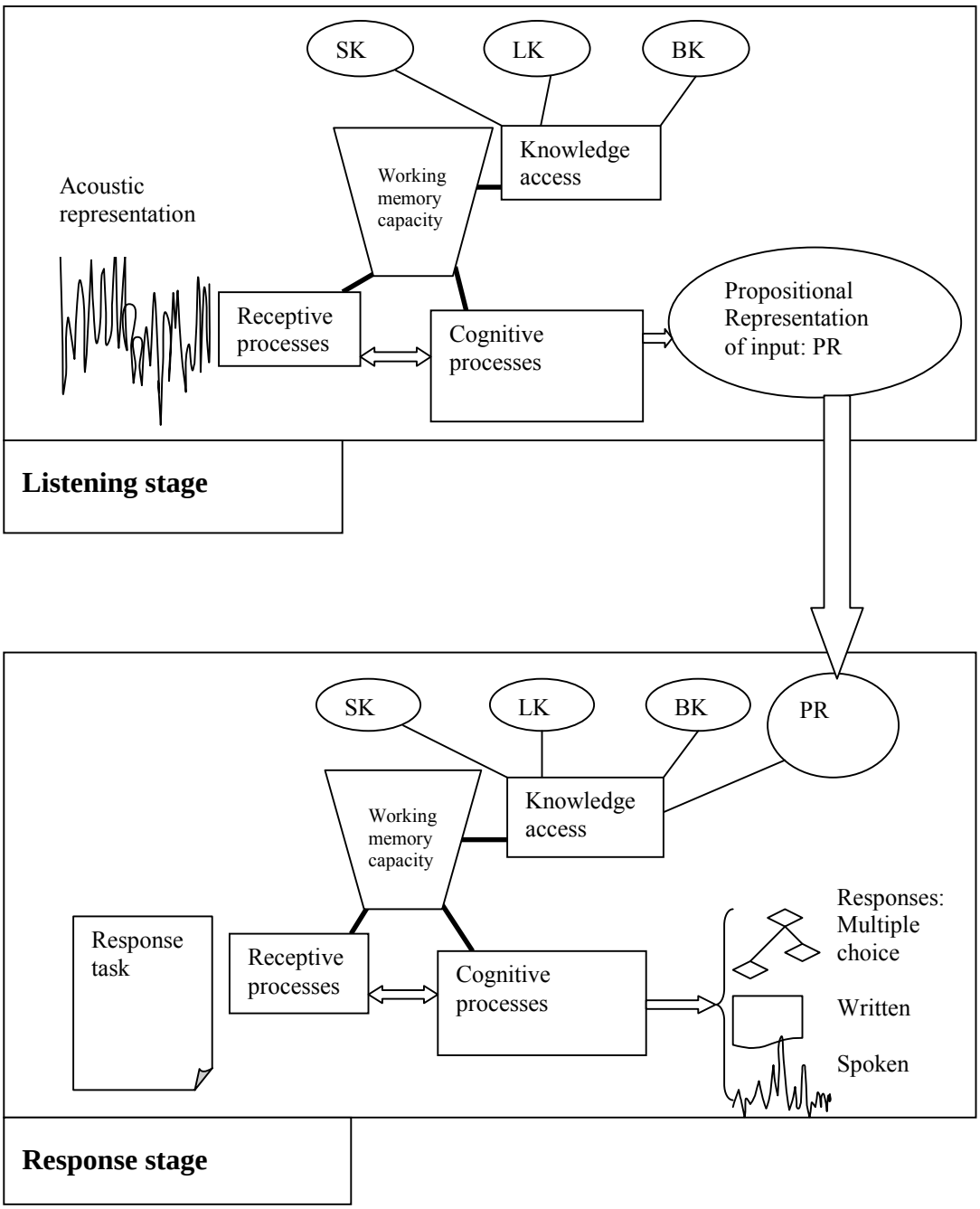
There is a general consensus that no uniformly agreed upon definition exists for listening in either native language studies (Devine, 1978; Dunkel, 1991; Henning, 1991) or second language studies (Brindley, 1998; Buck, 1997; Lynch, 1998). Let us start simply, then, by describing listening as the process of receiving an acoustic signal, which is then structured. While reception of the acoustic signal demands real-time processing, structuring of the acoustic signal requires linguistic, situational, and cognitive interaction and integration. Both reception and structuring are affected by memory load.

First let us examine the reception of the acoustic signal. Weaver (1972) wrote that listening occurs when people receive data aurally. Sound waves strike the tympanic membrane (the eardrum) and cause it to vibrate; the energy in these waves is transformed and carried to the central nervous system where the complicated process that we call *listening comprehension* occurs. As Brindley (1998) writes, there is a "move away from the notion of listening as auditory discrimination and decoding of decontextualized utterances to a much more complex and interactive model which reflects the ability to understand authentic discourse in context."

Figure 1 suggests that listening comprehension consists of a *listening* stage and a *response* stage. During the listening stage the acoustic signal is processed. At least three types of knowledge are accessed in real time during this stage: situational knowledge (SK), linguistic knowledge (LK), and background knowledge (BK). The different types of knowledge are accessed in real time as the incoming signal is processed by specialized receptive and general cognitive processes (depicted by rectangles). The result of this stage is a transformation of the incoming acoustic signal into a set of propositions (PR). Different listeners may arrive at different sets of propositions because of individual differences in knowledge, working memory capacity, and cognitive processing. The incoming stimulus, now mentally represented as a set of propositions, is operated on to produce a response that can be a selection from a set of choices, a written response, or a spoken response. The adequacy of the response (i.e., how well it demonstrates comprehension of the stimulus) is mediated by the representation of the stimulus, which in turn is mediated by knowledge and cognitive factors, and possibly by other proficiencies in the cases where a written or spoken response is required.

Situational knowledge refers to the important role of context. The listener frequently can see the person he or she is listening to and is provided with many visual clues as to the content and the implications of what is said. This important characteristic of listening is captured in this report and the TOEFL 2000 framework document (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000) in the "situational features" category.

Linguistic knowledge includes grammar (phonology, vocabulary, morphology, and syntax), discourse, and pragmatics, which consists of illocutionary and sociolinguistic competence (Bachman, 1990). Linguistic variables are designated as "grammatical features" in this report and the framework document (Jamieson et al., 2000). Cognitive demands include background knowledge, inferencing ability, and memory.



SK=situational knowledge
LK=linguistic knowledge
BK=background knowledge
PR=propositions

Figure 1. Listening and response stage of listening process

The listener has a purpose for listening and is ready to interpret what is said in terms of both expectations and background knowledge. The hearer's process of receiving input and creating meaning, or comprehending spoken language (Pica, 1994; Van Patten & Cadierno, 1993), is affected by memory capacity as well as by purpose and background knowledge. Rost (1994) wrote that 30 to 60 seconds are required by working (short-term) memory to sort out the essential elements of the aural stimuli. Entry into long-term memory requires that the new information be integrated with old, or pre-existing, information. Working memory has processing constraints, however. Thus, if the aural stimulus contains predominantly new information, the majority of the processing time/space is needed to sort it out, and access to old information is necessarily limited. This results, in terms of Figure 1, in a propositional representation of the stimulus that may be incomplete.

Real-time constraints increase the challenge of listening comprehension. Speakers often use pauses and changes of speed to provide clues for the chunking of information (Rubin, 1994). As Buck (1997) wrote, spoken texts which a listener must process consist of short, clause-like units about seven words long and about two seconds in duration, and include a variety of features (e.g., accent and dialect, slang, and up-to-date colloquialisms) and disfluencies (e.g., fillers and hesitations, false starts, and self corrections). These features of authentic texts must be considered in a definition of academic listening.

Apart from the theoretical considerations of processing information, two issues, task and pedagogy, concern language teachers (Ur, 1984). The first is related to the types of listening activities that actually occur in real life. The second concerns the particular difficulties that are likely to be encountered by the learner when coping with these activities. Informed by theory and research, we must address these issues as we pursue our main task: developing a listening test that reflects and measures both real-life academic listening and the difficulties that second language learners encounter. The next part of this report begins with a detailed examination of possible factors that characterize academic listening and which could affect the difficulty of academic listening tasks. A brief discussion of factors outside the construct definition of academic listening that could also affect difficulty appears in Appendix D.

3. Listening Framework for the TOEFL 2000 Test

Identifying the Test Domain

Not all listening tasks possible in the real world are relevant for the TOEFL 2000 test. As the framework document explicitly states, the test "will measure examinees' English language proficiency in situations and tasks reflective of university life in North America" (Jamieson et al., 2000, p. 10). Accordingly, the TOEFL 2000 listening test will focus on listening as it occurs in a university-level academic setting. Familiarity with this context will not be necessary to answer the test questions, however.

In focusing on academic listening, the assessment tasks will involve the transmittal of information and will be relevant across disciplines and levels of education. The test will measure examinees' ability to understand both explicit and implicit information, and both concrete and abstract information. Listening skills will be assessed both separately and in combination with other language skills (i.e., reading, speaking, and writing). Because examinees will be required to use other language abilities to respond to most task types, the test will not be a "pure" measure of listening.

The TOEFL 2000 listening test will focus on language use in context, not on decontextualized skills such as phoneme recognition and comprehension of single sentence utterances. Accordingly, examinees will have an opportunity to demonstrate their ability to comprehend coherent, appropriate, and meaningful messages.

The listening tasks will vary in terms of situational features, text materials, and test rubrics, as described in the following section. The text materials will vary with respect to features which characterize authentic texts, including sandhi variation, speed, frequency of vocabulary, use of fillers, hedges, emphatics, and technical terms. The tasks will be designed to discriminate most at the middle to upper levels of English as a second language/English as a foreign language (ESL/EFL) listening proficiency.

Organizing the Test Domain

The framework document and this listening report say that the TOEFL 2000 test "will measure examinees' English language proficiency in situations and tasks reflective of university life in North America" (Jamieson et al., 2000, p. 10). But how should this be operationalized? In other words, what is going to "drive" the development of listening tasks for the test? On what basis do we decide what tasks to include in the test and which to exclude?

One way to design the listening assessment is to decide what authentic tasks we want to include on the test, based on an analysis of the target language use situation, then decide what aspects of performance on those tasks we wish to assess and what interpretations about language ability we wish to make. Another option is to decide what communicative abilities we want to measure, then design tasks to measure these.

Although the listening team has not yet decided which of these approaches to follow, we have identified a set of listening abilities (adapted from Chapelle, Grabe, & Berns, 1997 and Richards, 1983) that we may want to assess in the TOEFL 2000 listening test. These abilities are:

1. comprehension of details and facts,
2. comprehension of vocabulary,
3. comprehension of main ideas and supporting ideas,
4. inferences about content and relationships (e.g., generalizations, cause vs. effect), and
5. comprehension of communicative function of utterances.

Once we have decided specifically which aspects of performance to measure, we will have a clearer notion of the types of questions and response formats that will need to be developed.

Identifying Task Characteristics and Variables

Now that we have identified the domain for the TOEFL 2000 listening test, we will identify and begin to operationalize the task characteristics to be included in the assessment. Based on the TOEFL 2000 framework document (Jamieson et al., 2000), we will concentrate on three sets of task characteristics: situation, text material, and test rubric. The following sections discuss each of these areas in turn.

Situation. Situation can be defined as the "extralinguistic setting in which an utterance takes place" (Crystal, 1991, p. 318). Most applied linguists today agree that a description of communicative language ability must include language use in context (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980; Chapelle, Grabe, & Berns, 1997; Hymes, 1972a, 1972b, 1974, 1976). In the TOEFL 2000 test, listening will be restricted to the types of aural input that students hear in academic situations on North American university campuses. The TOEFL 2000 framework (Jamieson et al., 2000) identified five components of the listening context, or s*ituation features*: participants, content, setting, purpose, and register. We have added another component, *situation visuals,* to characterize the information the visual provides, and have moved register to the pragmatic features of the text. In Jamieson et al., 2000, register is defined as "the degree of formality that is used in language" (p. 15), which is more descriptive of the nature of the text itself than of the situation.

Participants

Differences in participants have been claimed to be associated with differing performances in language use. However, much sociolinguistic work on the connections between interlocutors and language variation has focused on production. Mindful of this limitation, it is still worthwhile to examine reports on the relationship between participants.

Cathcart (1983) found that the age and authority relationship of interlocutors influenced language behavior; Ervin-Tripp (1968), Beebe (1977, 1983), and Young (1987) all found that ethnicity or social convergence were significant variables in language variation. Wolfson (1989) proposed what is known as the "bulge theory": when speech events are considered in relation to the social relationships of the speakers, one finds similarities in many of the interactions between intimates and the ones between strangers. The idea is that in these kinds of interactions, the predictability of responses is known. When the interlocutors are friends, co-workers, or acquaintances, however, there is much more variation in the verbal negotiation necessary for the speech event to be carried out.

Three variables have been hypothesized to influence differences in language use: solidarity (i.e., a participant's intention to emphasize cohesiveness with interlocutors), power (i.e., a participant's intention to emphasize differences in status among interlocutors), and networks (i.e., interlocking social groupings with whom participants interact regularly). Brown and Gilman (1960) is the classic study of the influence of solidarity and power on pronoun use. Studies by Friedrich (1972) and Milroy (1987) also provide evidence for the effect of solidarity on production, and this variable has been proposed as a major factor contributing to the emergence of national varieties, for example, "World Englishes" (Kachru, 1982). The influence of social networks on language variation has been hypothesized since the 1930s (Bloomfield, 1933), and was the basis for Labov's seminal study of the language of an African-American street gang (Labov, 1972).

We will employ aspects of Munby's (1978) taxonomy for communicative syllabus design as a way of organizing our analysis of participants. Although Munby's work has been the target of much criticism over the years as a basis for syllabus and test design, we feel it provides a valuable checklist of variables that might be taken into account in considering which variables are relevant for particular test-taker groups. We begin by positioning our target test-taker group within a social network. In our case, the international student in a North American university would interact with a variety of interlocutors, including other students, both North American and international; professors and other instructors (including teaching assistants); and non-teaching staff such as librarians; administrators, secretaries, international student advisors, lab technicians, and health center personnel. Powers (1986) surveyed faculty members at 34 institutions across 6 disciplines (engineering, psychology, English, chemistry, computer science, and business) to find out their views on the appropriateness of various participants for measuring listening comprehension among international students. The interlocutors were as follows:

1. a professor giving a classroom lecture,
2. a librarian giving information (e.g., on how to use a new computerized card catalog),
3. a registrar explaining procedures (e.g., for adding/dropping courses),
4. a foreign student advisor giving information/advice (e.g., on course selection),
5. a laboratory or teaching assistant explaining (e.g., an experiment or homework assignment),
6. a student health service nurse or doctor providing information (e.g., on medical care), and
7. a fellow student offering advice (e.g., on studying for an impending test).

Powers found that there were no possible interlocutors that faculty generally regarded as inappropriate. Thus, we can establish the "role set" for academic listening as including the above types of interlocutors.

Munby's next step (pp. 69-70) is to identify particulars for each member or group of the role set in terms of number (e.g., individual, small group, large group), age (e.g., adolescent, adult, elderly, mixed), gender (e.g., male, female, mixed), ethnicity (e.g., White, Hispanic, Asian), and nationality (e.g., U.S., international, mixed). Finally, Munby determines all the possible social relationships among the interlocutors implied by the role relationships (pp. 70-73). For example, the role relationship between an international student and a teaching assistant may involve a number of social relationships: instructor/learner, adult/adult, native/non-native speaker, and male/female. The role relationship between the international student and a health service doctor may involve the following social relationships: health provider/patient, professional/non-professional, native/non-native speaker, older/younger, insider/outsider, female/male. A list of plausible social relationships (based on Munby, 1978, p. 72) relevant to an international student and academic listening is given below.

1. superior/subordinate
2. "primes"/"pares"
3. chair/member
4. evaluator/applicant
5. authority/offender
6. employer/employee
7. manager/worker
8. investigator/subject
9. mentor/mentee
10. instructor/learner
11. advisor/advisee
12. health provider/patient
13. professional/non-professional
14. insider/outsider
15. native/non-native speaker

16. older/younger generation

17. male/female

18. minority/non-minority

Note that many of these relationships may be either asymmetrical, such as employer/employee, or symmetrical, for example, female/female or learner/learner. The common characteristic of the asymmetrical relationships is often that of power, while that of the symmetrical relationships is usually solidarity or cohesiveness. Whether a relationship is based on cohesiveness or power is often determined by the status of the interlocutors with regard to the differences between them, which in the academic context may be based on age, gender, occupation, and educational standing. Thus, a social relationship between two students might be asymmetrical if one of them were of higher educational standing than the other (e.g., a senior and a freshman), or if their roles for a particular activity were that of mentor/mentee.

In addition to the symmetry of role-set relationships, two other aspects of participants have been associated with listening difficulty. First, knowledge of the role of the speaker was found to significantly contribute to item difficulty in dialogues in the TOEFL test. In some items, like one involving buying a jacket from a salesperson, the language of the speakers was tied to a specialized role, whereas in other items, the exchange was general and could be spoken by anyone (Nissan, DeVincenzi, & Tang, 1996). It is anticipated that when items have visuals that give clues about speaker roles, this difference may decrease, as the test taker will no longer need to infer the role based only on the aural text. Second, Brown (1995) hypothesized that it is easier to understand texts in which there are fewer rather than more speakers.

In summary, participant variables for the TOEFL 2000 listening test include characteristics of the speakers (e.g., age, gender, ethnicity, and occupation). The number of speakers and their relationship (e.g., whether it is symmetrical or not) will be investigated in research studies as these variables are thought to contribute to difficulty.

Content

The framework document defines content as the "subject matter included in language tasks" (Jamieson et al., 2000, p. 14). Several articles report that familiarity with content, or background knowledge, enhances listening comprehension (Buck, 1997; Chaudron, 1995; Dunkel, 1991; Lynch, 1998; Mendelsohn, 1998; Rubin, 1994), although there is evidence that for other language tasks, such as speaking, familiarity with the topic does not reduce difficulty (Douglas, 1994). In any case, variation in content/topic is a characteristic of authentic language which must be considered in the construct of academic listening.

Three areas of content have been defined as relevant for the TOEFL 2000 listening measure: academic, class related, and campus related.

Setting

Only one study examined to date tried to relate awareness of the place where listening took place to item difficulty (Nissan et al., 1996). In this study, 17 variables were examined to determine the extent to which they could account for difficulty on TOEFL's multiple-choice dialogue items. One of the variables was location of the speakers, in which items were classified dichotomously according to whether the listener had to understand where the conversation was taking place to make sense of the utterance. This variable did not significantly affect item difficulty. However, as Lynch (1998) reports in his review of listening theories, setting provides important contextual support.

Two aspects of setting have been defined as relevant for the TOEFL 2000 listening test: its relevance to content and the actual location. Location is defined as instructional location, study location, or service location.

Purpose

Variation in speaker purpose is a characteristic of authentic language and academic listening tasks. To date, we have found no research on the effect that differing purposes might have on the difficulty of a listening task; however, the TOEFL 2000 reading team has proposed looking at four reading purposes as an organizing principle for the new TOEFL test: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, & Schedl, 2000). Moreover, they suggest that these four purposes provide a hierarchical means for organizing tasks and score reporting. These can be applied as follows to the purposes that people listen to a text:

1. listening for specific information,
2. listening for basic comprehension,
3. listening to learn, and
4. listening to integrate information.

The listening team is considering using these purposes for score reporting. A brief description of each purpose, as it applies to the listening modality, follows.

Listening for specific information requires comprehension of details of the text. Examples of this kind of listening in an academic context are listening to a professor's announcement of a due date of a term paper, and listening to understand the four stages of an insect's life in a biology lecture.

Listening for basic comprehension requires general comprehension of a text, such as comprehending its main idea(s) and important details related to the main idea. Comprehension of all details may not be necessary for this listening purpose.

Listening to learn includes both the ability to listen for specific information and the ability to listen for basic comprehension in a single text. The listener needs to integrate information from different parts of a text and understand the relationships between them.

Listening to integrate information includes the first three purposes, but this purpose requires the listener to integrate information from more than one text. This purpose would apply mostly to tasks that tap more than one skill, for example, where the test taker might read a text on a topic, listen to part of a lecture on the same topic, and then write a paragraph summarizing the differences.

Situation Visuals

Situation visuals will provide preliminary information to the examinees about the features of the "situation" of the stimulus—that is, who the participants are, where they are (the setting), and sometimes, the type of stimulus (text type).

A visual can provide information about the setting and give a sense of where the language is taking place. In some cases, the setting is related to the stimulus itself. For example, a lecture often takes place in a lecture hall, and checking out a book from the library usually occurs at the front desk of the library. These specific context visuals may facilitate comprehension. In other cases, the setting in the visual is not as relevant to the stimulus. For example, a conversation about a history assignment could take place in the library, outside a classroom, or elsewhere. It is not clear what impact this type of visual has on comprehension and on difficulty.

The situation visual will also provide information about the participants, and usually the role of the speakers will be apparent in the visual. For example, the visual might be a photograph of a professor in a lecture hall filled with students, a student speaking to a librarian at the front desk of a library, or two students walking into a classroom.

Sometimes the text type (type of stimulus) can be anticipated from the situation visual. For example, if the visual is of a professor in front of a large classroom, the examinees can expect to hear a lecture stimulus. A situation visual may also provide information about the topic of the text.

To sum up, situation visuals can be classified according to the type of situational information they provide:

1. Setting: relevant (e.g., visual of a lecture hall) or not relevant (e.g., visual of two people who could have been anywhere);

2. Participants: specific role (e.g., visual of a librarian) or general role (e.g., visual of two non-specific people talking); and

3. Text type: cued by visual (e.g., visual of a student giving a presentation) or not cued by visual (e.g., visual of two people talking, where it is not clear from the visual what type of discourse will occur)

Lynch (1998) reports that "there seems to be as yet no firm evidence that visual support on screen assists *long term* retention of spoken information." However, the popular belief is that visuals aid in listening. More information about the role of visuals is included below under the area of "content visuals."

Text Material. The framework document (Jamieson et al., 2000) specifies three components of text material: *grammatical features, discourse features,* and *pragmatic features.* Spoken texts also vary in terms of format, varying in terms of channel, content visuals, form, gestures, and length. The discussion that follows reports on research into characteristics of text material that have been associated with difficulty in academic listening.

Format

The format of the text material can vary in terms of the channel and form of its delivery, including the content visuals and gestures that may be a part of the stimulus, as well as the length of text. At this time all of these format features are being incorporated into our research plan to determine their effects on difficulty. Each of these features is discussed below.

Channel. The channel of the text in listening will be aural language. The aural text will always be accompanied by situation visuals, and may or may not be accompanied by content visuals.

Content Visuals. In addition to the *situation visuals* variable, introduced previously as a feature of situation, *content visual* features are included as a characteristic of the text material. Several studies suggest that, in many cases, addition of visuals reduces difficulty. A study described by Duker (1964) found that the addition of visual cues to oral speech was related to improved performance on immediate and delayed recall. Comparing student performance on audiotape vs. videotape in a beginning college-level Spanish course, Parry and Meredith (1984) report not only significantly better performance by the video group, but also more interest and greater motivation to pay attention. Rubin (1994) reported that visuals can enhance listening comprehension, but it depends on whether the visuals are consistent with the speech. Thompson (1995) echoes Rubin's sentiment that while videos can be helpful to listening comprehension, their effects are mediated by the relationship to the speech and the proficiency of the learner. Anderson and Lynch (1988) also suggest that listeners are helped by visuals that are intended to support spoken language. Ginther (2000) investigated the effect of visuals on performance on listening items in the computer-based TOEFL test.

Content visuals can be classified in four ways, based on their relationship to the stimulus (adapted from Levin, 1989, p. 85).

1. Replicate the oral stimulus (e.g., a key word or phrase on the blackboard that is explained in the stimulus). Examinees see exactly (a part of) what they heard.

2. Illustrate the oral stimulus (e.g., a picture of a house with gables in a talk on architecture). Examinees see a picture of what is described orally in the stimulus.

3. Organize information in the stimulus (e.g., an outline of the main points of a lecture, or a diagram of a process that was orally described). Examinees see some of the information in the stimulus presented in a different way.

4. Supplement the oral stimulus (i.e., where the visual provides information that is not in the oral stimulus). Examinees see new information.

Each of the four types of content visuals can be either graphic or text or both. We predict that types 1 through 3 make items easier, and type 4 makes items harder. Texts that are accompanied by supplementary content visuals will probably involve tasks integrated with other skills, e.g., reading.

Form. The form of the text will be language, which may or may not be supplemented by gestures.

Gestures. Gestures may be defined as movements of the body for the purpose of communication, either in conjunction with or independent of verbal communication.

Gestures are of four types (Morris, 1979):

1. Demonstrate: Draw attention to a particular object, person, or place such as by pointing.

2. Symbolize: Communicate shared common meanings that have direct verbal translations, such as the "OK" sign in the U.S., which means "zero" in France and is an obscenity in Brazil.

3. Illustrate: Accompany verbal statements, such as gesticulating excitedly to accompany an exciting story. These are often unconscious.

4. Mimic: Imitate certain objects or actions, such as folding the three middle fingers and extending the thumb and pinkie, while holding the hand near the ear to indicate talking on the phone.

One might hypothesize that demonstrative gestures would enhance communication, that emblematic and mimic gestures would potentially detract from it since they tend to be more culture-bound, and that illustrator gestures would have little effect either way, playing more of a "filler" or "accompaniment" role in communication.

Length. Length of text has been investigated to see whether it impacts item difficulty. In the Nissan et al. (1996) study, the number of words did not contribute to difficulty; however, there was little variation in the length of the texts. Henning (1991) reported on TOEFL listening comprehension items in terms of memory load. Two of his variables were repetition of the passage and passage length. He reported that when a passage was repeated, items tended to be easier, but there was no evidence that repetition had a positive effect on discrimination, item response validity, or format construct validity. A similar pattern of findings was reported when passages were one, two, or three sentences long; the shorter were easier, but test reliability increased with longer passages. He concluded that there is no evidence that any additional burden on memory associated with passage length will negatively affect item performance.

Henning's findings must be cautiously interpreted due to his operationalization of passage length. Three sentences may not have been long enough to show differences in memory. A study on TOEFL listening items (Yepes-Baraya, Yepes, & Gorham, in press) investigated the relationship between text length and memory, using texts of up to 150 seconds, which should shed more light on this area.

Rubin (1994) reports on research which indicates a weak relationship between short-term memory and listening. Henning (1991) also administered a memory test in his study; correlations at .38 were reported with listening items. However, no evidence for the reliability or validity of the memory test was given.

For the TOEFL 2000 listening assessment, the number of words and the length of the text in time will both be examined.

Grammatical Features

The following sections discuss three types of grammatical features of texts that will be considered in developing the TOEFL 2000 test: vocabulary, phonology, and syntax.

Vocabulary. From a lexical perspective, texts vary in such features as frequency of vocabulary, use of technical terms, and modifiers such as hedges and emphatics. A number of articles make the broad claim that vocabulary load affects difficulty in listening comprehension (Dunkel, 1991; Flowerdew, 1994; Powers, 1985), yet it is unclear how one gauges the relative weight of vocabulary from one text to another. Five ideas will be described below.

Frequency of vocabulary in both the input and the rubric has been cited in the literature as one source of difficulty. Supported by Rost (1990), this hypothesis was empirically tested by Nissan et al. (1996), who reported that the presence of infrequent words in the stimulus of listening comprehension TOEFL test items resulted in significantly more difficult items. Nissan et al. used a word list of 100,000 common words based entirely on conversations in the United States, primarily between adults and some between university students. Thompson (1995) also proposed that texts containing predominantly high-frequency vocabulary will be easier to understand than those which include jargon and technical terms.

Not everyone agrees, however, that difficulty in listening for second language (L2) learners is best determined by relative frequency of words in the text of the target language. One issue involves not the frequency of a given word, but whether or not the intended meaning of the word is its most common. Anderson and Lynch (1988) presented another claim: Vocabulary that is high frequency for students will not be so for the general public. To address this concern, vocabulary could be coded as non-technical, pan-technical (e.g., words that span disciplines, such as hypothesis, significant, robust), and technical (i.e., discipline specific) and then subsequently investigated to examine relationships with item difficulty. Anderson and Lynch also claim that difficulty will be affected by the context surrounding the word, knowledge of the topic, and similarity to a word in the examinee's first language.

A second way to describe features of vocabulary is in terms of "lexical bundles." Biber (1997) described lexical bundles as the most frequent lexical sequences to occur in a text. Examples of lexical bundles are "I don't know why..." and "so I said, well... ." Using a corpus containing 5 million words from conversations, Biber reported that about 40% of the words in conversations occurred in lexical bundles. Such frequency might lead one to hypothesize that these lexical bundles are easy to comprehend; however, Ur (1984) cited just such colloquial collocations as examples of difficult words to understand in context due to phonemic assimilation and reduction—in other words, the fast, slurred, pronunciation of these phrases by native speakers make them hard to understand. Anderson and Lynch (1988) also stated that colloquial words and phrases may be the most difficult to comprehend. Biber comments that these bundles are usually followed by an embedded complement clause which increases the syntactic complexity of the utterance. Thus, although lexical bundles might affect the difficulty of listening tasks, it is not clear from the literature whether they make a text easier or more difficult.

Thirdly, understanding key terms is perceived as important to success and difficulty for international students (Powers, 1985). An examination of real-life lectures revealed that key terms were often restated or paraphrased two to three times (Hansen & Jensen, 1994). Whether a word is repeated or expanded seems to affect learners of different abilities differently. Restatement of nouns has been shown to help learners of low ability, whereas elaborations such as paraphrases, use of synonyms, or appositives, help learners of high ability (Chaudron, 1995; Lynch, 1994; Rubin, 1994; Thompson, 1995).

Fourth, textual difficulty is thought to increase with the inclusion of cultural referents (Anderson & Lynch, 1988; Lynch, 1994). However, Nissan et al. (1996) did not find the presence of culture-specific words in the stimulus to be a significant predictor of difficulty. Perhaps it is not the use of words that captures this notion. It may be that in academic lectures, instructors make use of asides to clarify a main point, and in these asides they shift registers from formal to informal and also include broader cultural references that can be captured by word-level analysis (Chaudron, 1995; Rubin, 1994). This, then, is something to consider at the discourse level.

Finally, vocabulary types that might be associated with task difficulty include abstract words as well as modifiers of propositions. Powers (1985) suggested that abstract words and also vague words such as "some" and "very" will increase difficulty. Biber (1988) discusses four classes of words that modify propositions: downtoners, hedges, amplifiers, and emphatics. Downtoners (e.g., hardly, nearly, only, practically) serve to lower the force of the verb. Hedges (e.g., something like, more or less, maybe) are informal markers of uncertainty. Amplifiers increase the force of the verb (e.g., absolutely, extremely, fully, strongly). Emphatics (e.g., for sure, a lot, really) signal certainty. These forms might exemplify what Anderson and Lynch refer to as the "rough-and-ready use of generalized everyday words and phrases" which they associate with L2 listeners' problems.

In sum, four variables which could affect the difficulty of listening comprehension items have been identified: the number of infrequent words; lexical bundles; hedges, emphatics, and vague words; and technical terms. It is hypothesized that a greater number of them will increase the difficulty of items.

The presence and type of restatement of key terms will be captured in the analysis of the discourse variables (degree of redundancy and propositional complexity). Another area beyond the realm of vocabulary is the use of cultural references in asides; this area will also be captured within the features of the discourse (propositional structure).

Phonology. "Phonological competence" as applied to "connected speech" was discussed by Rubin (1994). Phonological competence should be a major aspect of the construct of academic listening because it refers to an aspect that is uniquely relevant to processing oral text. Phonological competence is not comprehension as such, but the preliminary step to comprehension. Unless the listener is able to aurally parse the text, comprehension is not possible.

From a phonological perspective, authentic texts include variation in speed or rate and accent as well as elision, insertion, disassimilation, and assimilation, a collection of phenomena captured in the term *sandhi variation.*

Pause Phenomena. Pauses and fillers such as "um" are thought to facilitate listening comprehension (Bygate, 1987; Hansen & Jensen, 1994), perhaps by allowing more processing time or by providing boundaries for listening "chunks." The results summarized by Rubin (1994) are somewhat conflicting, perhaps because of the lack of comparability in the subject populations. Potential interaction with level of competence was noted as well. For example, pauses may facilitate comprehension, hence reducing task difficulty, for more advanced students but may not help less proficient students for whom the task may be hopelessly difficult.

Speech Rate. Rubin (1994) noted that the evidence from rate studies is not easily comparable. A normal speech rate for native speakers of English is 165 to 180 words per minute (wpm), but this varies depending on the medium—conversations average 210 wpm, and lectures, 140 wpm for British speakers. Since the rate of speech varies among individuals and is negotiable between speakers, up to a point, it might seem reasonable to control difficulty through rate. However, for certain kinds of oral interaction—for example, a lecture or newscast— rate is not negotiable. Should the rate of speech for those items be manipulated to artificially slow them down? Flowerdew (1994) reported that studies show that slowing down speech (to 100 to 150 wpm) does not aid in comprehension, whereas speeding up speech (220 wpm) results in reduced comprehension. In an L1 study using compressed speech, Foulke (1962, cited in Duker, 1964) increased the rate of 175 wpm speech to 225, 275, 325, and 375 without distortion. The subjects, 291 blind children, understood with no fall in comprehension up to 275 wpm; thereafter, there was a general decline in comprehension.

Sandhi Variation. Sandhi variation refers to the influence that sounds have on each other in connected speech, and includes such phenomena as assimilation, insertion, deletion, and reduction. How do we deal with sandhi variation with regard to its effect on difficulty in listening? Ur (1984) cited colloquial collocations as examples of words difficult to understand in context due to phonemic assimilation and reduction—in other words, the fast, slurred pronunciation of these phrases by native speakers makes them hard to understand.

Stress and Intonation. Lynch (1998) writes that these prosodic features influence recognition of speech, but again, there are very few conclusive results regarding this variable's potential effect on item difficulty.

Accent. Flowerdew (1994) summarizes several studies which support the common sense view that unfamiliar accents cause listeners to have difficulty. Accent can be categorized as standard North American, regional North American, standard but not North American (e.g., Australian), or non-native. The theoretical, practical, and psychometric impact of the use of accents other than standard North American accents will be investigated.

In sum, aspects of phonology that will be operationalized and investigated further are pauses, speech rate, sandhi variation, stress and intonation, and accent.

Syntax. It is generally assumed that a text's syntactic complexity is related to difficulty in listening comprehension; the more complex the syntax, the more difficult the text is to understand (Anderson & Lynch, 1988; Bygate, 1987; Chaudron, 1995; Hansen & Jensen, 1994; Rost, 1990). Only one study was cited in Chaudron (1995) that noted no effect of syntactic complexity on listening comprehension; Chaudron states, though, that the operationalization of complexity as simplification vs. elaboration might not have been appropriate to measure its effect on listening comprehension.

How, then, does one measure syntactic complexity? Many articles surveyed proposed isolated examples. Anderson and Lynch (1988) and Bygate (1987) suggested that coordinating structures will be simpler than subordinating structures. The number of S-nodes (i.e., subject-verb structures) per T-units (within a sentence, independent clauses, and all attached clauses, phrases, and words) was the measure of complexity used by Cervantes and Grainer (1992). Dunkel (1991) reported that Carroll (1977) thought uniform sentence patterns within a text would be easier to process than a mix of sentence patterns. Bygate (1987) had a similar proposition: he thought that complex noun groups with many adjectives could be simplified by repetition of the same sentence structure, each with a single adjective. In another study, Henrichson (1984) reported that non-native speakers had much more difficulty than native speakers in comprehending contractions and reductions. Freedle and Kostin (1993) reported that across-clause referents and average sentence length of the first paragraph were significant predictors of difficulty in reading comprehension. Finally, Nissan et al. (1996) reported that the sentence structure of conversational turns was a significant predictor of difficulty.

The study by Hansen and Jensen (1994) was unique among those included insofar as they reported several syntactic structures that would increase syntactic complexity: nominalizations, attributive adjectives, indirect questions, complement and restrictive relative clauses, adverbial phrases, and prepositional phrases.

Obviously, there are many ways to describe syntactic complexity. Flowerdew (1994) states that there is no single parameter that can be used to adequately characterize spoken or written texts. He suggests that investigations into syntactic complexity would be better served by thinking of clusters of

features that work together—that is, by thinking of dimensions rather than single structures, citing work done by Biber.

Biber (1988, 1992, 1995) proposed five dimensions which included over 50 surface linguistic markers to account for discourse complexities of spoken and written registers. His analysis was based on two corpora containing 481 texts across 23 registers; they included approximately 960,000 words. The features and their respective dimensions appear in Appendix B. Each of the five dimensions is given a score.

In sum, a categorization of syntax will be achieved by giving the texts a value for each of Biber's five dimensions: Involved vs. Informational, Narrative vs. Non-narrative Discourse, Situation-dependent vs. Elaborated Reference, Overt Expression of Argumentation, and Non-abstract vs. Abstract Style.

Discourse Features

Discourse features are defined in the framework document (Jamieson et. al., 2000) as those that relate to "the nature and structure of text as a whole, including rhetorical type and textual organization" (p. 16). The listening team proposes that variables categorized as discourse features include propositional structure (including coherence and cohesion), propositional density, and propositional complexity.

Because many of the features of these variables are included elsewhere in the TOEFL 2000 framework, we will restrict our view of discourse to propositional analyses.

Propositional Analyses. The degree of redundancy in a text influences the ease with which readers or listeners can comprehend the information being communicated by the text. In a highly redundant text, it is necessary to process relatively little new information. Conversely, a text with little redundancy is usually more difficult to comprehend, especially when it addresses an unfamiliar topic. Of course, this difficulty varies from listener to listener; a text that would be redundant for one person might be nearly incomprehensible for another. To estimate the degree of redundancy of a text, we expect to use analyses of the propositions in the text.

Propositions are representations of information content (Brown & Yule, 1983). Brown and Yule (1983) note the difficulty of conducting a proposition-based analysis of text, mainly because the process of identifying propositions is subjective (Kintsch, 1974); however, several authors have since attempted to define approaches to identifying the propositional structure of lectures, which includes topics and subtopics as well as coherence and cohesion.

One approach is described by Hansen (1994). In her research, she identifies topics in lectures through topic-shift markers, a structural basis for dividing a text into smaller units (p. 133). According to Schriffin (1987), Hansen (1994), and Brown and Yule (1983), topic-shift markers have these characteristics:

1. They can be syntactically detached from a sentence.

2. They are commonly used at the beginning of an utterance.

3. They operate at the local and global level of discourse and on different planes of discourse.

4. They have no or only a vague meaning or are reflexive (Hansen, 1994, p. 136).

Hansen identifies topic-shift markers and uses a combination of Givón's (1979) sentential topic continuity measurements and a topic framework (Brown & Yule, 1983), which consists of "people, places, entities, events, facts, etc. already activated for both participants [the speaker and the audience] because they have been mentioned in the preceding conversation" (p. 79). The analysis yields major topics (main ideas), subtopics (supporting information), and minor points (modifiers of subtopics).

Young (1994) describes another approach which may be more useful than Hansen's for the identification of propositions because the text is analyzed on the basis of six features rather than a potentially infinite set of topic-related categories. Young identifies the macro-structure and micro-structure of lectures using a technique she calls phasal analysis. In phasal analysis, the strands of discourse that recur in and structure a spoken text are identified. To identify phases, Young used six primary configuration categories:

1. discoursal structuring—segments that identify or announce the direction the speaker will take,

2. conclusion—segments that summarize points,

3. evaluation—segments that evaluate information that has been given or is about to be given by indicating a personal endorsement or disagreement,

4. interaction—segments through which the speaker maintains contact with and reduces distance from the audience to help ensure that understanding takes place,

5. theory/content—segments that reflect the speaker's purpose and transmit information about theories, models, or definitions, and

6. examples—segments that illustrate concepts with concrete examples that are familiar to the audience.

Three steps are taken when applying the configuration categories in phasal analysis. First, each line of text is analyzed for semantic and syntactic choices; second, the configuration category (from the six listed above) is identified; and third, phases are labeled and all lines that constitute a phase are listed.

Both schemes for the identification of propositions require investigation to determine whether propositions can be efficiently and reliably identified to allow meaningful operationalization of propositional structure.

Coherence and Cohesion. Based on the assumption that category of text (function and text type) does not determine difficulty, but that other features of the text might, classifying texts by the type and number of coherence and cohesion features is a possible approach. Halliday and Hassan (1976, reported in Hatch, 1992) identify five types of cohesive ties: reference, substitution, ellipsis, conjunction, and lexical. These five types are broken down as follows:

1.	Reference	pronouns demonstratives comparatives
2.	Substitution	verbs nouns clauses
3.	Ellipsis	verbs nouns clauses
4.	Conjunction	additive adversative causal temporal
5.	Lexical	repetition synonym superordinate related words general related words

The cohesion and coherence of a text may be characterized by a value for each of the five types of ties.

Once cohesive ties, coherence, topics, subtopics, and subtopic modifiers have been identified, the *propositional structure* of a text can be characterized. *Propositional density*, the ratio of new propositions to the total number of propositions, and the ratio of propositions to speaking time, can be calculated. *Propositional complexity*, the degree to which a text deviates from the linear presentation of ideas, can also be determined.

Pragmatic Features

Features related to pragmatics are text type, rhetorical function, and register. Chapelle, Grabe, and Berns (1997) enumerate *text types* for listening as informal conversations, formal discussions, interviews, lectures, debates, newscasts, and orally administered instructions. The listening team has adopted this term and categorized this characteristic of text as a pragmatic feature rather than as a discourse feature. The framework document also states that *pragmatic features* of texts include

function which here refers to the rhetorical function of the speaker. Pragmatics also includes register, which we will characterize by how planned the text is.

Text Type. *Text type* is the term used by the listening team to designate whether the text is a lecture, conversation, or discussion. Shohamy and Inbar (1991) report that different types of spoken texts are associated with different degrees of difficulty; however, it seems likely that it is not text type itself which drives difficulty, but grammatical, discourse, or pragmatic features such as register. Text type is important, however, for test construction purposes and will be a part of the organizing matrix for the TOEFL 2000 listening team. The five text types proposed at this time are:

1. lecture: a monologue delivered in a formal academic setting, usually by a professor,

2. interactive lecture: a lecture delivered in a formal academic setting, where students may ask questions for clarification or the lecturer may ask students questions,

3. consultation: a group of speakers discussing academic or class-related material (e.g., a student asking a professor questions during office hours),

4. group discussion: a group of speakers discussing academic material in an imposed structure (e.g., students in a study group or students performing a classroom activity assigned by the professor), and

5. conversation: a group of speakers talking about class-related or campus-related material (e.g., a student looking for a book on reserve in the library).

Function. The function of the speaker is the speaker's reason for the speech event. Although this variable is not expected to affect difficulty, a list of the functions appears in Appendix A.

Degree of Planning. Sociolinguistic competence, a feature of pragmatic competence (Bachman, 1990), is represented by *register*. The listening team has decided to focus on one aspect of register, the degree of planning, which is related to the delivery of oral texts. Some lectures are carefully scripted before they are given, whereas others are given based on brief notes. Although we have no citations from the literature on how degree of planning itself might affect item difficulty, some of the grammatical features of language that are associated with more vs. less planning, such as the type of vocabulary and sandhi variation, have been discussed earlier in this report. The listening team proposes that the degree of planning of a text be one of the organizing features of test construction. We anticipate that the different degrees of planning will be manifested in the other text variables, such as discourse and grammatical features.

The degree of planning variable will be defined as planned, somewhat planned, and unplanned.

Other sociolinguistic variables which might be included are sensitivity to dialect or variety, sensitivity to naturalness, and understanding of cultural references (Bachman, 1990). We propose that

another sociolinguistic variable described by Bachman, understanding of figures of speech, be included under vocabulary, perhaps as part of the jargon/technical terms category.

Test Rubric. In the framework document (Jamieson et al., 2000), *test rubric* includes three elements: questions/directives (type of information, type of match, and plausibility of distractors), response format, and rules for scoring. We have expanded *test rubric* to five elements: instructions, question format, item-text interaction, response format, and rules for scoring. We added instructions and question format to separate them from questions/directives and renamed questions/directives to focus on the interaction between the item and the text.

Instructions

Instructions provide the test taker with explicit information about the nature of the test task (e.g., the ability being measured by the task, how to respond, and how the response will be scored) (Bachman & Palmer, 1996; Douglas, in press). Instructions may inform the test taker about the measurement objective of the task, its structure in terms of the number of tasks or items, their sequence, and their relative salience and importance. The instructions may also indicate how much time will be allotted for the task and provide information about scoring criteria and procedures. The channel for presenting the instructions may be aural, visual, or both.

Question Format

Question format refers to how the questions are formulated and is defined by channel, form, and time allotment (Bachman, 1990; Bachman & Palmer, 1996; Douglas, 2000). The channel may be aural, visual, or both; that is, the questions may appear on a computer monitor or in a test booklet, or they may be presented only aurally, or they may be presented in both mediums simultaneously. The questions may be presented as language, the most common form, but they may also be presented non-linguistically, in the form of pictures or diagrams, for example, or in both mediums. The questions may be presented for a limited amount of time, restricted either mechanically (e.g., when presented visually, they disappear from the screen after some number of seconds) or naturally (e.g., when presented aurally). The test takers may also be given unlimited time to process the questions, as, for example, when they are presented visually, and the test taker has control over when to respond and move to the next question.

Another aspect of the question format that could be considered is the directedness of the question, which refers to how specifically the question is asked. For example, "What does the man mean?" is less directed (i.e., less specific) than "What does the man mean to say about computers?" However, Nissan et al. (1996) found that this variable did not impact difficulty significantly.

Question format will be categorized in terms of channel, form, and time allotment.

Item-Text Interaction

The interaction between the item or question and the input text refers to the kind of information that listeners have to understand to successfully answer a question. It includes both the question or

stimulus, and the oral input text. Features reported to affect listening difficulty in this regard are: the presence of negatives and infrequent words in the question and the text, the "inference load" imposed by the question, the abstractness or concreteness of the information requested, the type of match between the question and the information in the text, the memory load imposed on the test taker, and the plausibility of the distractors.

Type of Information Requested. Difficulty of listening items varies as a function of the inference load the item imposes. This has been examined from two perspectives: phonological processing and cognitive load. Weaver (1972, p. 9) described two kinds of data in a verbal message: explicit and implicit. "A good listener hears the explicit data, but he 'hears' the implicit data too." In the case of dialogue items, the explicit case occurs if the correct answer is a paraphrase of something that is actually stated in the text; the implicit case occurs when the correct answer requires an inference beyond what is explicitly stated in the text. This variable impacted difficulty in the Nissan et al. (1996) study: explicit items were easier than implicit items. The basis for the inference may not so much be what is explicitly stated in the text or not, but aspects closer to oral communication, such as intonation and other spoken language variables. Although intonation was studied in the Nissan et al. (1996) study, it was not found to be a significant predictor of difficulty. However, because their analyses were based on written transcripts of the stimuli, Nissan et al. recommend that future studies be based on the aural stimuli themselves. Thus, both the linguistic ability to understand intonation and the cognitive ability of inferencing seem important in listening.

Finally, an item can be more difficult if the information it requires is more abstract than concrete (see Jamieson et al., 2000).

For the TOEFL 2000 listening test, type of information will be categorized in terms of whether the information tested was stated explicitly in the text or whether it was implicit. Also, the concrete nature of the information will be captured using the scale defined in Jamieson et al. (2000, pp. 64-66).

Type of Match. The type of match refers to the way examinees process text when they respond to a question. It includes the processes used to relate information in the question to the necessary information in the text, as well as the processes needed to either identify or construct the correct response (see Jamieson et al., 2000).

Henning (1991) investigated the effect of level of processing on item difficulty. He developed a "comprehension hierarchy" where items requiring information from one sentence were at the lowest end, and items in which information was required from three sentences were at the highest end. The level of comprehension processing had no effect on item difficulty. However, the differences between one and three sentence texts may not be great enough to require different kinds of processing.

Nissan et al. (1996) reported that presence of negatives and infrequent words in the stimulus accounted for item difficulty. This suggests the need to examine the vocabulary and syntax in the question as well as in the specific area in the oral text where the answer is located. Evidence that the linguistic characteristics of the section of the text where the answer is located has been indirectly provided for with regard to vocabulary. Restatement of nouns has been shown to help learners of low

ability, whereas elaborations such as paraphrases, use of synonyms, and appositives helped learners of high ability (Chaudron, 1995; Lynch, 1994; Rubin, 1994; Thompson, 1995).

Jamieson et al. (2000, p. 21) provide a framework for capturing the process that test takers go through when answering a question. For listening, we have changed the first category (locate) to "remember," as in listening, test takers do not go back to a written text to locate an answer. Thus, the four process categories for listening are remember (a detail), cycle (two or more details), integrate (cycle and identify a relationship), and generate (cycle and construct a relationship). In addition, the features of the text where the answer is located will be classified, as it is hypothesized that they will contribute to difficulty. For the initial analyses, we propose to examine the position of the answer in the text, as well as the syntax, vocabulary, rhetorical function (identify, define, describe, etc.), and phonological features of the text where the answer is located.

Plausibility of Distractors. The distractor set plays an important role in many item types. In the multiple-choice format, the attributes of the option set may contribute to the psychometric attributes of the item. Systematic rationales can be used in the distractor construction. For example, the distractor sets for analogy items are very different for GRE® and SAT® items. For GRE analogies, which need to be harder than SAT analogies, the options differ from the answer only slightly, but only one exhibits the exact semantic relationship with the stem. By contrast, for SAT analogies at least some distractors are often from different semantic classes or from no discernible class.

Some research has discussed distractors in terms of their relationship to the text. First, although explicit vs. implicit information in some sense deals with the nature of the stem-key relationship, it has potential implications for the construction of distractor sets. For example, if a key is a paraphrase, by what principle should the distractors be constructed? Should they also be paraphrases of stimuli text?

Second, items are least difficult when none of the distractors refers to information in the text. As more distractors refer to parts of the text, difficulty should increase. As the distractors share more in common with the key, difficulty should increase. Similarly, if the distractors refer to text in the proximity of the text corresponding to the key, difficulty should increase. In fact, a review of research prepared by Freedle and Kostin (1996) noted that "lexical overlap" was strongly implicated in the difficulty of listening items and was also related to difficulty in reading assessment items.

Third, Nissan et al. (1996) recommended that "designing distractors with plausible responses to the first speaker's utterance appears to be a reasonable way to create difficult items" (p. 28).

Other studies investigated only the response set. Studies by Nissan et al. (1996) and Freedle and Kostin (1993) investigated the lexical and syntactic characteristics of the correct answer and the distractors and found a number of variables which predict difficulty. Henning (1991) investigated the length of answers and found that shortened response options were generally easier than the unshortened TOEFL response options. Position of correct answer simply refers to the location of the key within the option set. It is reassuring that this variable did not significantly impact difficulty in the Nissan et al. (1996) study.

There appears to be convergent evidence that the difficulty of listening items can be manipulated through the characteristics of the option set. It remains to be discussed and investigated what the appropriate role of that variable might be in the design of a new listening exam. Specifically, should all items be constructed according to a standard option set, and thereby hold the effect on difficulty constant across items? Or should items vary in their option set? Since this variable is, at least to some extent, an artifact of the response format it would seem that the former is preferable. That is, the construct as currently defined stresses the "nontest purpose" of the exam. Therefore, we might maximize correlation with performance criterion by not relying excessively on a difficulty effect that is achieved through the option set. Of course, if, as was the case with the GRE analogy items, the processing of the option set can be made to tap the construct we are interested in, then the objection would not apply.

We have revised the Jamieson et al. (2000) analysis of distractors to capture the number of plausible distractors in the text, adding a category for distractors that are not in the text but are plausible responses to the speaker. In addition, the location of the distractor(s) will be identified.

Response Format

Characteristics of the format of the response the test taker is asked to provide include the channel, type, and time allotment (Bachman & Palmer, 1996). The test taker may be directed to respond orally, in writing, or to read options. The response type may be selected, as in a multiple-choice task; limited, as when a single word or short phrase is required; or extended, as in an essay. The test taker may be given a limited amount of time to complete the task, or be placed in control over when to respond and go on to the next question or task.

Rules for Scoring

The rules for scoring include the criteria for correctness of responses and procedures for scoring or rating them. The criteria may refer to aspects of language knowledge (such as grammar, vocabulary, or pragmatics), to communicative abilities (such as comprehending main ideas and details, extrapolating from the text to other situations, analyzing the text critically, and inferring attributes of the speakers, the situation, or the topic), or to whether a task was appropriately completed. Scoring procedures may involve merely deciding which responses will be counted as right or wrong, as in a multiple-choice task; assigning partial credit to responses; or developing a scale for rating various qualities of responses (Douglas, in press).

Defining Task Parameters

Tasks in the TOEFL 2000 listening test will be defined in terms of rubric and input. *Rubric* refers to that part of the test which gives the test taker explicit information about the nature of the task as a measuring device; it consists of instructions, question format, item-text interaction, response format, and rules for scoring. *Input* refers to the specific material that the test taker must process in taking the test, and consists of situation prompts and text.

To ensure that the new listening test is comprehensive and well balanced, it will be necessary to design tasks that target a large array of task parameters. Table 1 and 2 present checklists indicating the specific aspects of rubric and input that test developers must consider in constructing the listening tasks for the TOEFL 2000 test. Three sample tasks, each accompanied by a completed set of checklists, are presented in Appendix C.

Table 1
Checklist of Task Parameters: Rubric

Instructions		
	Channel	aural, visual, both
	Structure	number, salience, sequence, importance of tasks
	Time allotment	limited, unlimited
	Scoring	criteria, procedures
Question format		
	Channel	aural, visual, both
	Form	language, language and non-language
	Time allotment	limited, unlimited processing time
Item-text interaction		
	Type of information requested	concrete/abstract explicit/implicit
	Type of match	process: remember, cycle, integrate, generate text characteristics: position (beginning, middle, end), syntax, vocabulary, phonology
	Plausibility of distractors	number: none, plausible response, one, two or more location: end, middle, beginning
Response format		
	Channel	oral, written
	Type	selected, limited production, extended production
	Time allotment	limited, unlimited response time
Rules for scoring		
	Criteria for correctness	areas of language knowledge, communicative abilities, task completion
	Procedures for scoring	right/wrong, partial credit, rating scale

Table 2
Checklist of Task Parameters: Input

Situation prompt			
	Participants		instructor, student, service personnel
	Topic		
		Academic	life sciences, social sciences, humanities/arts, physical sciences
		Class related	assignments, due dates, textbooks
		Campus related	registration, advising, health care, library help
	Setting		
		Instructional location	lecture hall, classroom, seminar room, laboratory
		Study location	dorm study room, library, instructor's office, computer center
		Service location	health center, bookstore, registrar's office, business office, advisor's office
	Purpose		listen for information, listen for comprehension, listen to learn, listen to integrate
	Situation visuals		topic, setting, participants, text type
Text			
	Format		
		Channel	aural, aural and visual
		Content visuals	replicate, illustrate, organize, supplement
		Form	language, language and non-language
		Gestures	demonstrative, symbolic, illustrator, mimic
		Length	words, time
	Grammar		
		Vocabulary	technical, pan-technical, non-technical frequency lexical bundles, hedges, emphatics
		Phonology	pauses, fillers, contractions, interruptions, stress, intonation, accent
		Syntax	complexity
	Discourse		
		Propositional density	ratio of new to total propositions ratio of propositions to speaking time
		Propositional structure	cohesion coherence topics, subtopics, details
		Propositional complexity	degree to which text deviates from linear presentation of ideas
	Pragmatics		
		Function of the speaker	give directions/instructions, recommend/suggest/advise/persuade, complain/apologize/forgive, give opinion/agree/disagree, describe/define/compare/summarize/classify, hypothesize/predict/speculate, request/invite, narrate
		Text type	lecture, interactive lecture, consultation, group discussion, conversation
		Degree of planning	planned, somewhat planned, unplanned

4. Technological Issues

We read with interest the TOEFL monographs on technology for the TOEFL 2000 test, particularly *A Review of Computer-based Speech Technology for TOEFL 2000* (Burstein et al., 1999). Although the paper suggests that speech synthesis technology could be used for certain types of listening comprehension tasks, we do not at the present time see any advantage to using such technology in place of compressed digitally recorded input texts. We will, of course, continue to stay informed about developments in this area and may suggest at a later time that experimental test tasks using speech synthesis be devised and field-tested. For the moment, we offer the following observations with regard to technology in the TOEFL 2000 listening component:

1. We assume that the technology being used to deliver the listening component of the current CBT TOEFL test—that is, compressed digital sound and visuals—will continue and even be enhanced by the time the TOEFL 2000 test is operational.

2. We would like to explore the possibilities offered by real-time video. We need to know whether there is a measurement advantage in the use of video over the use of sequenced still photos. There is no doubt that video offers the potential for enhanced face validity and authenticity, although there is a lot of concern about its potential for distraction. There is also a question about its cost effectiveness, owing in part to video production costs and in part to uncertainty about the recyclability of video vs. sequenced still photos.

   As for the measurement advantage, we have in mind the potential offered by interactive compact disk-computer technology, which offers test takers options such as revisiting various parts of the input for confirmation and comprehension purposes, calling up resource materials such as a dictionary or thesaurus, and requesting repetition of questions. The computer can easily track the number and type of such requests and the speed with which responses are made (see Douglas, 1988, for a review). Of course, it would still be unclear whether there were any advantages to using video vs. single or sequenced still photos. However, the listening team would like to explore this issue before deciding against video for purely financial reasons.

   We would like to explore, too, the possibility that the high cost of video could be ameliorated somewhat by using parts of each tape for different tests and tasks. For example, a thirty-minute video of a class in which the instructor gave a mini-talk, engaged the students in an academic discussion, and then assigned them to small groups to complete a task could provide input for several listening and integrated tasks.

3. As for integrated tasks, in which test takers listen to some input and then respond by writing or speaking, we might like to explore with the speaking and writing teams the possibilities offered by computerized template-scoring of spoken and written production.

5. Research Agenda

In this section we discuss the outline of a research agenda in support of listening comprehension. The primary goal of the agenda is to yield a test blueprint within a three-year period. The agenda is structured in such a way as to make it possible to document the improvements that we will want to claim with respect to score meaning and consequences, as well as to guide the research to support those claims.

Because the TOEFL 2000 test will be delivered by computer, a comprehensive research agenda needs to include research on issues motivated by the medium, in addition to proficiency-specific research. In a recent review of several computer-based testing efforts, Bennett and Bejar (1997) argued that almost every aspect of the system is implicated in score meaning and test consequences. In their scheme, construct definition, task design, and test design occupy a central role. Much of the foregoing framework has been devoted to laying a foundation for construct definition, task design, and test design. We view the research agenda as a series of studies that will inform the development of a test blueprint, which is a detailed set of psychometric and content specifications. Item pools are created based on this test blueprint, and from the pools, tests are produced. The tests will yield scores that lead to valid inferences about the test takers.

Task design refers to the enumeration of a set of possible tasks, and evaluation of their psychometric feasibility. Test design refers to a series of constraints, such as the amount of time allocated to the exam, and the required number of items of a particular type.

The test blueprint can be thought of as a "base form" that serves as the basis for constructing a pool of items in such a way that all the items in the pool are used with essentially the same frequency. This precise and highly targeted approach to item pool design is in contrast with the traditional approach, where items accumulate in a pool as a result of the pretesting process, and then test forms are created from the pool according to a selection algorithm (Stocking & Swanson, 1992). By building the pool with a particular base form in mind we are enforcing selection criteria, as in the traditional approach, but with the advantage that the item pool is constructed with those criteria in mind. Therefore, no effort is wasted creating tasks that will seldom be used.

Construct definition, task design, and test design are critical components in developing a new assessment, but not the only ones. Another important component is the nature of the scoring process, especially if it is an automated process. In that case, scoring is strongly linked to interface and tutorial design. For example, to avoid overly complex scoring procedures there might be a temptation to use simpler tasks that are more amenable to automated scoring, but this could be at the expense of construct coverage. To avoid that possibility, while maintaining construct coverage, the interface may need to be somewhat complex, in which case tutorial design becomes a major mechanism to prevent irrelevant constructs, such as facility with computers, from affecting performance. These same considerations affect, and are affected by, the availability of appropriate test development tools, such as on-line corpora and dictionaries. Certain item types may be very expensive to produce without appropriate tools, which could lower their likelihood of appearing in the blueprint. Finally, all of these considerations ultimately impact score reporting.

Organization of Validity Evidence

Clearly, the transition to the TOEFL 2000 test needs to be supported by research that will ultimately support our assertions about test scores. It is convenient and useful to map the reality of CBT design onto validational schemes to ensure that scores satisfy validational criteria. To achieve this objective we resort to broad schemes of test validation (Messick, 1989). Specifically, we distinguish between evidentiary and consequential aspects of validity. The evidentiary basis of validity refers to the empirical linkages one can bring to bear to support score meaning. The consequences of using a measurement instrument are now also considered part of validation (Messick, 1989). Of particular relevance to this agenda are the consequences with respect to language instruction. Both the evidentiary and consequential aspects of validity can be approached empirically through a distinction made by Embretson (1983), namely construct representation and nomothetic span. Construct representation refers to the internal consistency of evidence regarding the definition of the construct, whereas nomothetic span refers to evidence relating the construct to other constructs.

The crossing of these two dichotomies—evidentiary/consequential and construct representation/nomothetic span–yields a 2 x 2 table (Table 3). This table can be used to plan a series of research studies, which, if carried out with positive results, would form the basis for claims about score meaning and consequences. Moreover, by using this scheme at the design stage we are forced to consider what we wish and do not wish to measure as part of the construct of listening comprehension, and what consequences we might hope for or strive to avoid. Thus, we can attempt to build validity into the scores rather than leaving validation as the last step in the research and development process.

Table 3
A 2 x 2 Scheme to Guide Test Development and Validation

	Construct representation	Nomothetic span
Evidentiary	What is the construct as judged by internal evidence regarding construct definition, task design, test design, scoring procedures?	What is the evidence regarding what the construct is or is not by examining construct definition, task design, test design, scoring procedures, *and* examining listening comprehension performance in relation to other linguistic proficiencies?
Consequential	What are the positive and negative implications of the construct as represented through task design, test design, interface, and tutorials, and as perceived by different constituencies? Are there pragmatic considerations, such as security and scheduling, that may inadvertently affect construct representation?	What are the positive and negative implications of the set of interrelationships between listening comprehension and other linguistic proficiencies or variables, such as linguistic background?

Evidentiary/Construct Representation. This cell of Table 3 deals primarily with evidence regarding construct definition, task design, test design, and scoring procedures. These components essentially define the meaning of scores that are ultimately reported; therefore, research to justify the choices we ultimately make is critical. By carefully documenting deliberations regarding these components, we record some of the justifications for score interpretation. Moreover, the cell provides an action plan. Earlier in this document we recast research on listening comprehension into a set of variables to guide the conception of listening comprehension tasks and to study the basis for variation in difficulty. Consequently, the research agenda calls for investigations of these variables as determinants of difficulty and the creation of listening comprehension tasks consistent with the framework. To the extent that the tasks are not objectively scored, research is also needed to identify scoring procedures and the extent to which the scoring can be automated**.** Factors that are not relevant to the construct but that may affect difficulty are discussed briefly in Appendix D.

Evidentiary/Nomothetic Span. This cell is concerned with gathering evidence regarding what the construct is, or is not, through examination of its relationship with external variables. Whereas construct representation can be examined at the item level, nomothetic span is best examined at the score level. Therefore, we may not be able to collect information about the relationship between listening comprehension performance and other variables until it is possible to estimate proficiency. Nevertheless, some aspect of nomothetic span can be examined early on at the item level, for example, by looking at performance on certain kinds of tasks based on listening and reading texts. Some discussion of reading and listening appears in Appendix E.

Consequential/Construct Representation and Nomothetic Span. The bottom two cells of Table 3 deal with real or perceived consequences. Because it may be too late by the time unintended negative consequences are discovered, it is important to be vigilant during the design stage to avoid them if possible. The experience of other CBT efforts is likely to be a valuable source of learning. For example, the scheduling algorithm that is or is not used in test centers (e.g., Lewis, Sheehan, & Swanson, 1994) can affect construct representation in ways that have consequences to students. If scheduling bottlenecks were to occur they would constitute an unintended negative consequence of computer-based testing that could have, or could be perceived to have, potential fairness implications.

Testing Time and Test-Task Design

Discussions about testing time can be facilitated by clearly defining *the range of proficiency* we aim to measure so that the best measurement can be obtained in the shortest period of time. However, because the emphasis in the exam is on the assessment of communicative competence, it may be necessary to use a significant number of open-ended tasks. An important consideration is the relative amount of testing time to be allocated to open-ended tasks versus selected-response tasks. In general, open-ended tasks require more time and yield less reliability per unit of testing time than do selected-response tasks. Therefore, the time requirements of different possible tasks will be kept in mind and ways of scoring performance to maximize information yield will be identified.

Security

Computer-based testing raises important test security issues as illustrated by the first generation of computer-based tests. A critical aspect of the development of computer-administered tests is that design practices and procedures be implemented to preclude the possibility of items, or the entire item pool, becoming compromised through overexposure. Significant progress has been made in the multiple-choice arena to achieve this objective under an adaptive multiple-choice test design (e.g., Stocking, 1993). Applications for an open-ended response format have been less well investigated. One possible approach for an open-ended test design would be to create several test forms and rotate their use. A natural extension of this idea is to construct linear forms "on the fly"[1] for any given test taker, so that each test taker is administered a different set of items. The approach is also well rooted in psychometric theory in that it is in concert with psychometric models that emphasize sampling as a basis for score interpretation and response modeling as opposed to the postulation of a latent trait (Tryon, 1957; Shoemaker, 1975).

Systematic Item Writing

Although the approach described above has favorable security implications, it also has a price. The specifications for constructing items or tasks need to be far more detailed than the usual specifications. Templates, such as the ones described in the section on sample tasks, need to be developed for each kind of listening task. Moreover, the specifications need to address not only content considerations but psychometric ones as well. This framework and the research that follows it can guide the overall

[1] This is currently being done in the reading section of the TOEFL computer-based test.

conceptualization of the item pool as well as the systematic construction of the items. Moreover, by relying on computer-based tools for test development the process can be more efficient.

Alternative Test Designs

In computer-based testing, two aspects of efficiency are important: item exposure rate and testing time. From a purely efficiency perspective, a test design that minimizes the exposure of item pools and is as short as possible is preferable. Higher exposure rates lead to higher maintenance costs because as items reach a certain exposure level they can become compromised. Longer tests lead to higher costs because in computer-based testing, test center costs are computed as a function of an hourly rate. To lower costs, test designs may be possible that capitalize on the computer as a delivery medium. For example, it is possible to design a two-stage adaptive test, where candidates below a desired level of proficiency are terminated in the first stage and only applicants not terminated in the first stage proceed to the second stage. The efficiency gains in this case result from savings in testing time and minimizing the exposure of tasks in the item pool designed to discriminate at a higher level of proficiency.

Logistics

The foregoing suggests the highly interrelated and complex nature of all aspects of the design. Much needs to be investigated and there is relatively little time to carry out the research. Therefore, we plan to start with the most basic research as soon as possible, and to proceed so that the research from one year feeds into the research for the following year.

Year 1. The first year will be devoted to four major efforts: conducting a needs analysis, analyzing variables in current computer-based TOEFL test items, prototyping new tasks, and establishing an oral academic corpus. The needs analysis will be based on responses from a variety of constituencies, such as test takers, university professors, and ESL experts.

The second major effort is to conduct analyses of computer-based TOEFL test items with the goal of identifying which variables fit the current framework and determining which variables contribute to difficulty. These studies will be along the lines of the Nissan et al. (1996) study but based on the variables as defined in the current framework. The following variables from the present framework have been identified as applicable to computer-based TOEFL test items:

1. number and symmetry of participants,
2. purpose of listener,
3. text type, degree of planning, and task type,
4. content visuals (Types 1 through 3),

5. discourse analyses: determine a system and how to apply it efficiently to capture propositional structure, propositional density, propositional complexity, and

6. item-text interaction.

The third major effort is to develop prototypes of potential tasks not in the current computer-based TOEFL test. The prototypes will provide information on:

1. the number of speakers in a stimulus,

2. the use of supplement visuals to accompany the audio stimulus,

3. the impact of gestures in video and still photos,

4. the use of response types other than selected response,

5. tasks integrated with the modalities of speaking, writing, and reading, and

6. the effects of notetaking.

Each prototype will be evaluated according to whether it taps aspects of the construct that are not included in the computer-based TOEFL test. Practical issues will also be evaluated, such as:

1. Can multiple instances be produced efficiently and stored?

2. What range of proficiency does the task measure?

3. Can responses be scored automatically or objectively? If not, can judges be trained to reliably score them?

The conception of new item types will necessarily proceed without complete knowledge of how the framework variables affect difficulty. To wait until the potential new item types have been prototyped and sufficiently refined to collect data entails the risk of not having sufficient time to incorporate results.

The fourth crucial effort for Year 1 will be to create an appropriate corpus of spoken academic English.

Year 2. At the end of the first year there will be a set of potential listening tasks and information on the role of the framework variables in predicting difficulty. It is likely that the set of such tasks will be larger than can be accommodated. Year 2 will be devoted to refining the most promising subset of tasks, taking into account operational considerations such as interface and tutorial demands. Moreover, larger data collection will take place to evaluate the tentative models of difficulty formulated on the

basis of Year 1 results. Research on construct representation will continue in Year 2 by studying the following variables in the context of some of the task prototypes developed in Year 1:

1. phonology (based on corpus established in Year 1),
2. content visuals: Type 4—Supplement (prototype integrative tasks),
3. length (informed by Yepes-Baraya ongoing study and corpus),
4. vocabulary (using corpus),
5. comparison of stills vs. video (e.g., as applied to gestures),
6. syntax (using corpus),
7. discourse: investigate difficulty drivers (using system of discourse analysis devised in Year 1),
8. integrated tasks,
9. note-taking (integrated tasks, scoring), and
10. response formats.

Year 3. Year 3 will be devoted to completing ongoing research, formulating a test blueprint, and developing scoring rules. Blueprint development is a still evolving methodology (e.g., Elliot & Nelson, 1984; Raymond, 1996; Schmeiser & Estes, 1986). The formulation of a blueprint can be thought of, as indicated earlier, as the construction of a "base form" (Bejar, 1993). The outcome of the research agenda can be thought of as a database for identifying the base form that best satisfies a set of constraints. Schematically, the database consists of information on all the task types that have been investigated and that have survived scrutiny. The identification of a base form is "simply" the identification of one or more subsets of task types, together with their frequency, that maximizes or minimizes the "scores" characterizing the set in terms of how well it represents the construct, the adequacy of the psychometric attributes for the range of proficiency we wish to measure, as well as timing and computer resources requirements.

6. A Better Listening Test

The TOEFL 2000 project represents one of the most ambitious psychometric research and development efforts ETS has ever undertaken. The resulting test should represent a major improvement over the existing instrument.

Construct Representation

A likely advantage of the TOEFL 2000 test over the current computer-based TOEFL test is improved construct representation. For one thing, a wider variety of contexts is likely to be sampled. These contexts will have been identified from existing research and further refined from research conducted as part of the research and development effort. The difficulty of task types, whether they are multiple-choice or open-ended and integrative, will be modeled. To the extent that variation in difficulty across task types can be attributed to the variables we postulated as determinants of difficulty, we will be able to claim an improved understanding of performance on the test. That understanding in turn will allow us to "engineer" the test rather than assemble it from tasks whose difficulty cannot be predicted precisely. Furthermore, a clear definition of the range of proficiency we wish to measure is likely to translate into improved measurement precision. More discrimination at the higher end of the proficiency scale will be possible.

By being aware during the research and development process of threats to validity that affect computer-based testing (CBT), and by being mindful of the potential positive and negative consequences the test could have, we expect to produce an assessment that satisfies the demands of several constituencies without sacrificing construct representation. For example, we expect to have a positive effect on language instruction and applied linguistics research.

Communicative Function of Language

The tasks that we envision for the TOEFL 2000 listening component will be communicative in that they will be situated in contexts that are relevant to both test takers and test users, specified in terms of participants, topic, setting, and purpose. We anticipate that this will encourage language teachers and materials developers to focus more on communicative language use in academic contexts, and that so-called "TOEFL preparation" courses will more closely resemble communicatively oriented academic English courses. The test will include tasks that require integration with other language skills, as do real world tasks.

In addition, we assume that the TOEFL 2000 listening tasks will help spawn applied linguistics research into a previously much neglected area, that of understanding listening strategies and processes among English as a second language (ESL) students in North American universities, and how they are affected by changes in such variables as number of participants, purpose of listening, and communicative event.

Interpretation of Test Results

Another advantage of the TOEFL 2000 listening component over the current computer-based exam will be in its potential for making and reporting richer interpretations of performance. For example, the new test could make it possible to report levels of listening performance according to purpose and/or

task type, defined by the variables affecting difficulty, as we have discussed in this paper. This, in turn, would make it possible to report a profile of listening skills that will be useful to different constituencies: the students themselves, instructors, and administrators. Students will potentially obtain a fuller picture of their listening skills in various contexts and for various purposes; instructors will potentially receive detailed information about their students that will make it possible to tailor academic listening courses to groups of learners with similar profiles; and administrators will be able to use the information for planning and allocation of resources necessary for ESL course and program development.

References

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

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

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* New York: Oxford University Press.

Beatty, M., & Payne, S. (1984). Listening comprehension as a function of cognitive complexity, a research note. *Communication Monographs, 5,* 483-489.

Beebe, L. (1977). The influence of the listener on code-switching. *Language Learning, 27*, 331-339.

Beebe, L. (1983). Risk-taking and the language learner. In H. Seliger & M. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 39-66). Rowley, MA: Newbury House.

Bejar, I. I. (1993). *Optimization approach to the design of tests consisting of complex tasks.* Paper presented at a meeting of the Psychometric Society, Barcelona, Spain.

Bennett, R. E., & Bejar, I. I. (1997). *Validity and automated scoring: It's not only the scoring* (Research Report No. RR-96-13). Princeton, NJ: Educational Testing Service.

Biber, D. (1988). *Variation across speech and writing.* New York: Cambridge University Press.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes, 15,* 133-163.

Biber, D. (1995). *Dimensions of register variation.* New York: Cambridge University Press.

Biber, D. (1997). Lexical bundles: What the grammar books don't tell you. In *Perspectives on spoken and written discourse.* Colloquium conducted at the meeting of Teachers of English to Speakers of Other Languages, Orlando, FL.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart, and Winston.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistic, 18,* 171-191.

Brown, G. (1995). Dimensions of difficulty in listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 59-73). San Diego, CA: Dominie Press, Inc.

Brown, G., & Yule, G. (1983). *Discourse analysis.* Cambridge: Cambridge University Press.

Brown, R., & Gilman, R. (1960). The pronouns of power and solidarity. In T. Sebeok (Ed.), *Style in language* (pp. 253-276). Cambridge, MA: MIT Press.

Buck, G. (1997). Testing language skills. *Encyclopedia of Language and Education,* 7.

Burstein, J., Kaplan, R., Rohen-Wolff, S., Zuckerman, D., & Lu, C. (1999). *A review of computer-based speech technology for TOEFL 2000* (TOEFL Monograph Series Report No. 13). Princeton, NJ: Educational Testing Service.

Bygate, M. (1987). *Speaking.* New York: Oxford University Press.

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

Carroll, J. B. (1977). On learning from being told. In M. C. Wittrock (Ed.), *Learning & instruction* (2nd ed., pp. 496-512). Berkeley, CA: McCutchan.

Cathcart, R. (1983). *Situational variability in the second language production of kindergartners*. Unpublished doctoral dissertation, University of California at Berkeley.

Cervantes, R., & Grainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*, 767-770.

Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series Report No. 10.) Princeton, NJ: Educational Testing Service.

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: Dominie Press, Inc.

Crystal, D. (1991). *A dictionary of linguistics and phonetics.* Cambridge, MA: Basil Blackwell, Inc.

Devine, T. (1978). Listening: What do we know after fifty years of research and theorizing? *Journal of Reading, 21,* 296-304.

Douglas, D. (1988). Testing listening comprehension in the context of the ACTFL Proficiency Guidelines. *Studies in Second Language Acquisition, 10*, 245-261.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 1*, 125-144.

Douglas, D. (2000). *Assessing language for specific purposes.* Cambridge: Cambridge University Press.

Duker, S. (1964). Listening. *Review of Educational Research, 34*, 156-163.

Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly, 25,* 431-457.

Elliot, S. M., & Nelson, J. (1984). *Blueprinting teacher licensing tests: Developing domain specifications from job analysis results*. Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Enright, M., Grabe, B., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper*. (TOEFL Monograph Series Report No. 17). Princeton, NJ: Educational Testing Service.

Ervin-Tripp, S. (1968). An analysis of the interaction of language, topic, and listener. In J. Fishman (Ed.), *Readings in the sociology of language* (pp. 192-211). The Hague: Mouton.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension--An overview. In J. Flowerdew (Ed.), *Academic listening* (pp. 7-29). New York: Cambridge University Press.

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items* (TOEFL Research Report No. 44). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity* (TOEFL Research Report No. 56). Princeton, NJ: Educational Testing Service.

Friedrich, P. (1972). Social context and semantic feature: The Russian pronominal usage. In J. Gumperz & D. Hymes (Eds*.), Directions in sociolinguistics: The ethnography of communication* (pp. 270-300). New York: Holt, Rinehart, and Winston.

Ginther, A. (2000). *Effects of the presence of different types of visuals on subjects' performance and preferences on CBT TOEFL listening comprehension stimuli* (TOEFL Research Report No. 67). Princeton, NJ: Educational Testing Service.

Givón, T. (1979). *On understanding grammar* (Perspectives in Neurolinguistics and Psycholinguistics). New York: Academic Press.

Hale, G., & Courtney, R. (1994). The effect of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing, 11,* 29-47.

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English.* London: Longman.

Hansen, C. (1994). Topic identification in lecture discourse. In J. Flowerdew (Ed.), *Academic listening* (pp. 131-145). New York: Cambridge University Press.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241-268). New York: Cambridge University Press.

Hatch, E. (1992). *Discourse and language education.* Cambridge: Cambridge University Press.

Henning, G. (1991). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance* (TOEFL Research Report No. 33). Princeton, NJ: Educational Testing Service.

Henrichson, L. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning, 34*, 103-126.

Hymes, D. H. (1972a). Models of the interaction of language and social life. In J. J. Gumperz & D. H. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35-71). New York: Holt, Rinehart, & Winston.

Hymes, D. H. (1972b). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Hymes, D. H. (1974). *Foundations in sociolinguistics.* Philadelphia: University of Pennsylvania Press.

Hymes, D. H. (1976). *Ethnography, linguistics, narrative inequality.* Bristol, PA: Taylor and Francis, Inc.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper*. (TOEFL Monograph Series Report No. 16). Princeton, NJ: Educational Testing Service.

Kachru, B. (1982). Models for non-native Englishes. In B. Kachru (Ed.), *The other tongue: English across cultures* (pp. 31-57). Oxford: Pergamon.

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

Labov, W. (1972). *Language in the inner city.* Philadelphia: University of Pennsylvania Press.

Levin, J. R. (1989). A transfer-appropriate-processing perspective of pictures in prose. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 83-100). Amsterdam: North-Holland.

Lewis, C., Sheehan, K., & Swanson, L. (1994). *The examinee scheduling algorithm - A new tool for computer-based testing* (Research Memorandum No. 94-12). Princeton, NJ: Educational Testing Service.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal*, *75*, 196-204.

Lynch, T. (1994). Training lecturers for international audiences. In J. Flowerdew (Ed.), *Academic listening* (pp. 269-289). New York: Cambridge University Press.

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics, 18,* 3-19.

Mendelsohn, D. (1998). Teaching listening. *Annual Review of Applied Linguistics, 18,* 81-101.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: American Council on Education.

Milroy, L. (1987). *Observing and analyzing natural language*. Oxford: Blackwell.

Morris, D. (1979). *Gestures: Their origins and distributions.* London: Jonathan Cape.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* (TOEFL Research Report No. 51). Princeton, NJ: Educational Testing Service.

Parry, T., & Meredith, R. (1984). *Videotape vs. audiotape for listening comprehension tests: An experiment.* (ERIC Document Reproduction Services ED 254 107).

Pica, T. (1994). Questions from the language classroom: Research perspectives. *TESOL Quarterly, 28,* 49-79.

Powers, D. (1985). *A survey of academic demands related to listening skills.* (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service.

Powers, D. (1986). Academic demands related to listening skills. *Language Testing, 3*, 1-38.

Raymond, M. R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education, 9*, 237-256.

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, 219-240.

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

Rost, M. (1994). On-line summaries as representations of lecture understanding, In J. Flowerdew (Ed.), *Academic listening* (pp. 93-127). New York: Cambridge University Press.

Rubin, A. (1980). A theoretical taxonomy of the differences between oral and written language. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 411-438). Hillside, NJ: Erlbaum.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78,* 199-221.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Schmeiser, C. B., & Estes, C. A. (1986). *Translating task analysis results into test specifications: Clarity or confusion?* Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Shoemaker, D. M. (1975). Toward a framework for achievement testing. *Review of Educational Research, 45*, 127-147.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question. *Language Testing, 8,* 23-40.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing program* (Research Report No. RR-93-2). Princeton, NJ: Educational Testing Service.

Stocking, M., & Swanson, L. (1992). *A method for severely constrained item selection in adaptive testing* (Research Report No. RR-92-37). Princeton, NJ: Educational Testing Service.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, *49(2)*, 219-274.

Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 31-58). San Diego, CA: Dominie Press, Inc.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229-249.

Ur, P. (1984). *Teaching listening comprehension*. New York: Cambridge University Press.

Van Patten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition, 15,* 225-243.

Weaver, C. (1972). *Human listening–Processes and behavior.* Indianapolis, IN: Bobbs-Merrill Educational Publishing.

Wolfson, N. (1989). *Perspectives: Sociolinguistics and TESOL*. Cambridge, MA: Newbury House.

Yepes-Baraya, M., Yepes, J., & Gorham, J. (in press). *TOEFL listening comprehension: Effects of task and listener variables on understanding and memory* (TOEFL Research Report). Princeton, NJ: Educational Testing Service.

Young, L. (1994). University lectures–Macro-structures and micro-features. In J. Flowerdew (Ed.), *Academic listening.* Cambridge: Cambridge University Press.

Young, R. (1987). *Variation and the IL hypothesis*. Paper presented at the 11th University of Michigan Conference on Applied Linguistics, Ann Arbor.

Appendix A: Listening Variables

I. Situational Features

A. Participants

1. Characteristics of speakers

1. Number of speakers

2. Age
   1 = young (20's) 2 = old (30's +) 3 = mixed

3. Gender
   1 = male 2 = female 3 = mixed

4. Ethnicity
   1 = white 2 = minority 3 = mixed

5. Occupation
   1 = student 2 = instructor 3 = other 4 = mixed

2. Relationship between speaker(s) and listeners

1. Symmetrical (e.g., two students, two teaching assistants)
2. Asymmetrical (e.g., student and professor, librarian and student)
3. Changes during the stimulus (e.g., starts out symmetrical, becomes asymmetrical)

B. Content (Topic)

1. Academic
   1. Life Sciences
   2. Social Sciences
   3. Humanities and Arts
   4. Physical Sciences

2. Class related (related to a class the speakers are taking)
   (e.g., assignments, due dates, text books)

3. Campus related
   (e.g., registration, faculty advisor, health care, library help)

C. Setting

1. Relevance to content
    1. Relevant
    2. Not relevant

2. Location

    1. Instructional location
        (e.g., lecture hall, class, seminar room, laboratory)
    2. Study location (not classroom)
        (e.g., dorm study room, library, instructor's office, computer center)
    3. Service location
        (e.g., health center, bookstore, registrar's office, dining area, business office, faculty advisor's office)

D. Purpose of listener

1. Listening for specific information
2. Listening for basic comprehension
3. Listening to learn
4. Listening to integrate information

E. Situation visuals (context)

1. Setting
    1. Relevant
    2. Not relevant

2. Participants
    1. Specific
    2. Not specific

3. Text type
    1. Cued
    2. Not cued

II. Text Material

A. Format

1. Channel
    1. Aural
    2. Aural and visual

2. Content Visuals
    1. Replicate
    2. Illustrate
    3. Organize
    4. Supplement

3. Form
    1. Language
    2. Language and non-language

4. Gestures
    1. Demonstrate
    2. Symbolize
    3. Illustrate
    4. Mimic
    5. None

5. Length
    1. Number of words
    2. Time

B. Grammatical Features

1. Vocabulary
    1. Number of infrequent words (definition of infrequent to be established)
    2. Number of lexical bundles (e.g., “what I meant to say was”)
    3. Number of vague words
    4. Number of technical words

2. Phonology
    1. Pauses/fillers
    2. Rate
    3. Sandhi variation (assimilation/reduction/ellipsis)
    4. Stress and intonation
    5. Accent
        1. Standard
        2. Regional, North America
        3. Standard, not North America
        4. Non-native

3. Syntax—complexity score

C. Discourse Features

1. Propositional structure

    1. Cohesion and coherence
        1. Reference
        2. Substitution
        3. Ellipsis
        4. Conjunction
        5. Lexical

    2. Young's (1994) six categories
        1. Discoursal structuring
        2. Conclusion
        3. Evaluation
        4. Interaction
        5. Theory/content
        6. Example

2. Propositional density

    1. Ratio of new propositions to total propositions of the whole text (low percent means extensive redundancy, and is hypothesized to be easier)

    2. The density ratio in the first 10 seconds of the text, the first minute, etc.

3. Propositional complexity
   How much the text deviates from a linear presentation of ideas

D. Pragmatic Features

1. Text type
    1. Lecture
    2. Interactive lecture
    3. Consultation
    4. Group discussion
    5. Conversation

2. Functions
    1. Give directions/instructions
    2. Recommend/suggest/advise/persuade
    3. Complain/apologize/forgive
    4. Give opinion/agree/disagree
    5. Describe/define/compare/summarize/classify
    6. Hypothesize/predict/speculate

7. Request/invite
8. Narrate

3. Register
    1. Formal (i.e., read or carefully planned, perhaps elaborate notes)
    2. Consultative (pre-defined goal, e.g., to explore an unambiguous point of a lecture)
    3. Informal (i.e., spontaneous conversation, no planning)

III. Test Rubric

A. Instructions

1. Channel
2. Structure of task
3. Time
4. Scoring information

B. Question format

1. Channel
    1. Aural
    2. Aural and visual (stills, video)

2. Form
    1. Language
    2. Language and non-language (e.g., pictures)

3. Time allotment
    1. Limited
    2. Unlimited

C. Item-text interaction

1. Type of information requested

    1. Concrete/abstract
        1. Identify a person, place, or other concrete object.
        2. Identify amount, time, type.
        3. Identify manner, goal, purpose, condition.
        4. Identify cause, effect, pattern, similarity, opinion, explanation.
        5. Identify difference, theme, equivalent (e.g., defining an unfamiliar word).

    2. Explicitness
        1. Explicit
        2. Implicit

2. Type of match

    1. Process
        1. Remember
        2. Cycle
        3. Integrate
        4. Generate

    2. Characteristics of text where answer is located

        1. Position
           3 = beginning 2 = middle 1 = end
        2. Syntax
        3. Vocabulary
        4. Function
        5. Phonology

3. Plausibility of distractors

    1. Number (Adapted from Appendix H of Jamieson et al., 2000)
        1. No plausible distractors in text
        2. No plausible distractors in text, but distractor is plausible response to first speaker
        3. Only one distractor contains information in the text
        4. Two or more distractors contain information in text

    2. Location
       end = 3 middle = 2 beginning = 1

D. Response format

1. Channel
    1. Spoken
    2. Written
    3. Read

2. Type of response
    1. Selected (e.g. multiple choice)
    2. Limited (e.g., one word answer)
    3. Extended e.g., short essay)

3. Time
    1. Limited
    2. Unlimited

E. Rules for scoring

Appendix B: Five Dimensions of Structural Complexity in English (from Biber, 1995)

Dimension 1: Involved vs. Informational Production

"Involved production" (positive features)

private verbs
that deletion
contractions
present tense verbs
second person pronouns
do as pro-verb
analytic negation
demonstrative pronouns
general emphatics
pronoun *it*
be as main verb
causative subordination
discourse particles
indefinite pronouns
general hedges
amplifiers
sentence relatives
wh questions
possibility modals
non-phrasal coordination
wh clauses
final prepositions
adverbs

"Informational production" (negative features)

nouns
word length
prepositions
type-token ratio
attributive adjectives
place adverbials
agentless passives
past participial postnominal clauses

Dimension 2: Narrative vs. Non-narrative Discourse

"Narrative discourse" (positive features)
past tense verbs
third-person pronouns
perfect aspect verbs
public verbs
synthetic negation
present participial clauses

"Non-narrative discourse" (negative features)
present tense verbs
attributive adjectives

Dimension 3: Situation-dependent vs. Elaborated Reference

"Situation-dependent reference" (positive features)
time adverbials
place adverbials
adverbs

"Elaborated reference" (negative features)
wh relative clauses on object positions
pied-piping constructions
wh relative clauses on subject positions
phrasal coordination
nominalizations

Dimension 4: Overt Expression of Argumentation

(positive features)
infinitives
prediction modals
suasive verbs
conditional subordination
necessity modals
split auxiliaries
possibility modals

(***no negative features)***

Dimension 5: Non-abstract vs. Abstract Style

(***no positive features)***

(negative features)

conjuncts
agentless passives
past participial adverbial clauses
by-passives
past participial postnominal clauses
other adverbial subordinators

Appendix C: Sample Tasks

Below we present three sample tasks—two "independent" listening tasks, and one integrated listening/writing task.

Sample task 1: Sequoia task (audio input)

In this task, test takers listen to a lecture about the life and contributions of Sequoia, the Native American leader. After listening to the lecture, test takers answer multiple choice questions.

Rubric		
Instructions		
	Channel	aural and visual
	Structure	"Listen to a lecture in a linguistics class." "Now get ready to answer the questions."
	Time allotment	general test instructions tell test takers they do not have a limit on response time for each question
	Scoring	no information given to test takers in instructions
Question format		
	Channel	stems: aural and visual choices: visual
	Form	both language and non-language
	Time allotment	unlimited processing time
Item-text interaction		
	Type of information requested	concrete, explicit and implicit
	Type of match	process: remember, cycle location: beginning, middle, end
	Plausibility of distractors	number: two or more; location: middle
Response format		
	Channel	written
	Type	selected
	Time allotment	unlimited response time
Rules for scoring		
	Criteria for correctness	recognize lexical/semantic relations, variations in meaning; discriminate among forms and structures
	Procedures for scoring	right/wrong

Input			
Situation prompt			
	Participants		instructor, students
	Topic	Academic	social science: linguistics
	Setting	Instructional location	Classroom
	Purpose		listen for comprehension
	Situation visuals		setting: photo of professor addressing class participants text type: photo of professor addressing class
Text			
	Format		
		Channel	both visual and aural
		Content visuals	replicate: words "Cherokee", "Sequoia" illustrate: pictographs, syllabic symbols
		Form	both language and non-language
		Gestures	Illustrator
		Length	336 words, 3.5 minutes
	Grammar		
		Vocabulary	technical, pan-technical some infrequent
		Phonology	normal pauses, few fillers or contractions, no interruptions, normal stress and intonation, standard accent
		Syntax	moderately informational, fairly non-narrative, neutral situation dependency, fairly non-argumentative, moderately abstract
	Discourse		
		Propositional density	fairly high ratios
		Propositional structure	highly organized, well marked cohesively
		Propositional complexity	fairly linear
	Pragmatics		
		Functions	describe/define/compare/summarize/classify
		Text type	lecture
		Degree of planning	planned

Sample task 2: At the bookstore (video input)

A student goes into a bookstore to get books for English 101. She finds that there are several books for this course, some marked "Required" and some "Recommended." She asks a clerk for help: "The book by Johnson and the one by Smith are required, but the one by Clark and the college dictionary are recommended. What's the difference and do I need to buy all the books? I'm a bit short of money right now." The clerk is rather unhelpful: "Well, it's up to you. 'Recommended' means recommended, not required." Student asks if she can buy two now and get the rest next week. Clerk says maybe, but they might be sold out by next week, and anyway, they send the unsold books back to the publisher after two weeks. Student says she'll take the two required texts and the dictionary now and come back for Clark later. Clerk adds up total, money changes hands, student leaves with books.

Rubric			
Instructions			
	Channel		both aural and visual
	Structure		two tasks: marking which books are required, which recommended; limited production questions
	Time allotment		unlimited
	Scoring		responses are scored for content, not grammar, style
Question format			
	Channel		aural only
	Form		both language, non-language (pictures of books)
	Time allotment		real-time processing of questions
Item-text interaction			
	Type of information requested		concrete, implicit
	Type of match		process: remember, integrate, generate
	Plausibility of distractors		plausible response
Response format			
	Channel		written
	Type		selected, limited production
	Time allotment		unlimited response time (test taker controls rate of question presentation)
Rules for scoring			
	Criteria for correctness		judging relative importance of information; subordinate/superordinate relations; recognizing paralinguistic cues; comprehend main idea, details; extrapolate
	Procedures for scoring		right/wrong, rating scale
Situation prompt			
	Participants		student, service personnel: bookstore clerk
	Topic	Campus related	required textbooks
	Setting	Service location	bookstore
	Purpose		listen for information
	Situation visuals		content: sign—"English 101—Required and Recommended texts" setting: bookstore participants: student, clerk text type: face to face conversation
Text			
	Format		
		Channel	both aural and visual
		Content visuals	replicate: sign—some books required, some recommended
		Form	language
		Gestures	demonstrative: pointing to textbooks symbolic: raising hands, palms up–"it's up to you"
		Length	500 words, 5 minutes
	Pragmatics		
		Functions	recommend/suggest/advise/persuade
		Text type	conversation
		Degree of planning	unplanned

Sample task 3: Movie analysis (audio input)

In this task, test takers listen to a discussion between two instructors (one is Hispanic) about the merits and faults of a movie. One gives it a "thumbs up" and the other a "thumbs down," and they explain and justify their points of view with reference to specific scenes in the film. After listening to the discussion, test takers first answer some multiple-choice comprehension questions, then write a paper comparing and contrasting the two views of the movie.

Instructions		
	Channel	both aural and visual
	Structure	two tasks: MC comprehension questions, writing task
	Time allotment	unlimited
	Scoring	writing task rated for communicative effectiveness, register appropriacy, cohesion/coherence
Question format		
	Channel	both aural and visual
	Form	language
	Time allotment	limited processing time
Item-text interaction		
	Type of information requested	abstract, implicit
	Type of match	process: remember, integrate, generate
	Plausibility	plausible response, two or more
Response format		
	Channel	written
	Type	selected and extended production
	Time allotment	unlimited response time
Rules for scoring		
	Criteria for correctness	follow topic development; analyze tone; recognize genre markings; recognize paralinguistic clues; infer links between ideas
	Procedures for scoring	right/wrong, rating scale

Input			
Situation prompt			
	Participants		two instructors
	Topic	Academic	humanities/arts
	Setting	Study location	instructor's office
	Purpose		listen to integrate
	Situation visuals		setting: stills of instructors in office participants: two middle aged instructors text type: academic discussion
Text			
	Format		
		Channel	both aural and visual
		Content visuals	replicate: poster for movie illustrate: stills from movie showing certain characteristics
		Form	language and non-language: stills from film
		Gestures	symbolic: thumbs up, thumbs down illustrator: occasional "normal" gestures
		Length	600 words, 6 minutes
	Grammar		
		Vocabulary	pan-technical, non-technical fairly frequent lexical bundles
		Phonology	long pauses; lots of fillers, contractions, interruptions; stress, intonation; accent: one standard, one hispanic
		Syntax	highly involved; fairly non-narrative; moderately situation dependent; moderately argumentative; fairly non-abstract
	Discourse		
		Propositional density	moderate ratio of new/total
		Propositional structure	conversational, turntaking
		Propositional complexity	fairly non-linear
	Pragmatics		
		Functions	explain/justify
		Text type	academic discussion
		Degree of planning	unplanned

Appendix D: Construct-irrelevant Factors That May Affect Difficulty

The fact that most response formats require listeners to use other language skills implies that it is not possible to design a "pure" measure of listening (Brindley, 1998; Henning, 1991). In the TOEFL 2000 test, however, every attempt will be made to reduce the impact of the other skill, unless that other skill is being measured in an integrated task. For example, if answering a question requires reading options or writing a response, every attempt will be made to keep answers brief.

Other test method factors which could influence difficulty include computer familiarity (Taylor, Kirsch, Eignor, & Jamieson, 1999), time permitted for testing, the testing environment, and the sequencing of the test (Bachman, 1990; Bachman & Palmer, 1996).

Additional factors which could affect difficulty and are unrelated to our construct definition include characteristics of the individual: background knowledge such as topic familiarity, personality characteristics such as motivation and anxiety, strategies such as note-taking, and general cognitive ability (Beatty & Payne, 1984; Hale & Courtney, 1994; Rubin, 1994). Effort needs to be made to keep influences such as these to a minimum, since they represent sources of "construct-irrelevant variance" (Messick, 1989).

Appendix E: Academic Listening and Reading

Because listening and reading are both receptive skills, unlike the productive skills of speaking and writing, there are many similarities between them. Similar mental processes are thought to underlie reading and listening in that the subskills associated with these two skills often overlap (Devine, 1978; Brindley, in press). Chapelle et al. (1997) characterize listening and reading in terms of nearly identical lists of features under each of the following types of language competence: sociolinguistic, linguistic, discourse, and functional. Similarities between reading and listening have been reported in many empirical studies using correlational techniques; however, the degree of similarity has not been sufficient to consider them the same construct. The correlation between the listening and reading sections of the 1995 TOEFL was .64. Devine (1978) reported findings from several studies with correlations between listening and reading of about .70. Studies correlating speed of reading with listening also reported coefficients of about .70 (Duker, 1964). This pattern of moderate correlations indicates that there are differences between listening and reading.

Differences between listening and reading can be categorized in terms of text characteristics (linguistic), interaction with the text (processing time), and what is remembered from the text. Grammatical differences are found between spoken and written texts (Biber, 1988, 1992, 1995; Devine, 1978; Lynch, 1998; Rubin, 1980). For example, Biber has identified syntactic features such as *that* deletions, contractions, pro-verb *do*, *it*, and demonstrative pronouns as being more prevalent in spoken registers than written, whereas relative clauses, nominalizations, and high levels of lexical specificity are much more common in written registers than spoken.

Another type of linguistic difference between listening and speaking involves discourse and the marking of idea units. Spoken language features temporal characteristics such as pauses and changes in speed which often provide clues for the chunking of information. Written text does have some compensatory aspects; a partial analogue to prosodic features is punctuation, although our limited set of punctuation marks does not reflect all of the nuances possible with speech. In contrast with speech, segmentation of the written message into words and sentences is correctly indicated in written text and is not a task that must be performed by the reader. In addition, certain devices that are used solely in written texts can help specify the larger structure of the message, such as demarcation of paragraphs (Rubin, 1986).

Two other differences between spoken and written text are processing time and interactions. Speech is typically temporary, whereas the reader can go back to text previously read or look ahead to get a sense of the structure of the material. Speech is less clearly segmented, disfluent when spontaneous, more colloquial, subject to phonetic modification, and phrasally or clausally based (Lynch, 1998). The reader is usually alone with printed material; he or she can neither ask it questions nor pick up signals apart from the print. The listener, on the other hand, (usually) has the speaker right there and can interrupt in order to ask for clarification; he or she also has the advantage of being able to study facial expressions. The listener can also pick up signals from the speaker's management of stress, pitch, and juncture patterns (Devine, 1978).

Still another difference involves the type of information that listeners and readers remember. Lund (1991) found that readers recalled more information and in greater detail than listeners; listeners recalled proportionately more main ideas and did more inferencing work

Test of English as a Foreign Language
P.O. Box 6155
Princeton, NJ 08541-6155
USA

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

Phone: 609-771-7100
E-mail: toefl@ets.org
Web site: http://www.toefl.org

58713-005961• Y90M.750 • 253720 • Printed in U.S.A.