TOEFL®

Monograph Series

MS - 20
JUNE 2000

TOEFL 2000 Speaking Framework: A Working Paper

Frances A. Butler
Dan Eignor
Stan Jones
Tim McNamara
Barbara K. Suomi

ETS™ Educational Testing Service

TOEFL 2000 Speaking Framework:
A Working Paper

Frances A. Butler
Dan Eignor
Stan Jones
Tim McNamara
Barbara K. Suomi

Educational Testing Service
Princeton, New Jersey
RM-00-6

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 paper was prepared based on an initial overall framework paper developed for the TOEFL 2000 project by Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000). The paper applies concepts advanced in the overall paper to the modality of speaking. In doing so, this document presents an initial framework for research and development activities for the speaking component of the TOEFL 2000 test. The paper should be viewed as a work-in-progress, as research activities presently underway for TOEFL 2000 will undoubtedly bring about refinements to the contents of this document.

The paper is made up of six parts. Part 1 provides an introduction to the overall document. In Part 2, oral discourse is discussed from a sociological perspective as well as in terms of speech act theory. Part 3 discusses the details of the speaking framework for the TOEFL 2000 test. Identification of the test domain and relevant task characteristics and variables is discussed, along with some of the factors suspected to influence the difficulty of speaking tasks. In Part 4, some of the technical issues involved in eliciting and capturing speech samples are discussed. Part 5 contains a list of relevant research activities that should be pursued as the project progresses. The final part considers ways in which the new TOEFL 2000 speaking component will improve upon the current version of the TOEFL test and the Test of Spoken English (TSE).

Key words/phrases: Speaking proficiency, communicative competence, oral discourse, speaking task design

Acknowledgments

The authors wish to thank Agnes He, Richard Young, and Peter Skehan for generously sharing prepublication manuscripts. A special thank you should be directed as well to Evelyn Hatch for insight and feedback on discourse issues at various stages in the preparation of this document.

Table of Contents

List of Tables

1. Introduction

This document presents the initial framework for the research and development for the speaking component of the TOEFL 2000 test. In the following section (Part 2), we discuss oral discourse from the sociological perspective as well as in terms of speech act theory—the two major traditions in the study of native speaker oral communication. As discussed in that section, the body of research does not provide a strong basis for developing the TOEFL 2000 test of speaking proficiency; therefore, the framework for the assessment represents an effort to combine what is known about oral communication and about language use in academic settings with our growing understanding of the possibilities and limitations inherent in oral language testing.

Part 3 defines the test domain and outlines the characteristics of the tasks (e.g., in terms of situational features, discourse features, visual materials, and test rubric). This section also discusses some of the factors that appear to influence the difficulty of speaking tasks, such as the number of participants and the types of information involved, and summarizes certain task characteristics and performance conditions that affect the listener's attentional resources. Part 4 considers technological issues involved in assessing speaking ability.

Part 5 of this document presents a list of research activities that are needed to guide the test development effort. Part 6, the final section, considers the ways in which the new TOEFL 2000 speaking component will improve on the earlier versions of the test.

2. Conceptualizing Speaking Proficiency

In general, speaking can be defined as the use of oral language to interact directly and immediately with others. For the purposes of the TOEFL 2000 assessment, the concern is with such interactions as occur in academic settings. In such settings, interactions are primarily directed towards acquiring, transmitting, and demonstrating knowledge. In addition, interactions in the academy may concern the organization, management, and regulation of learning activities.

Communicative competence in oral academic language requires control of a wide range of phonological and syntactic features, vocabulary, and oral genres and the knowledge of how to use them appropriately. This framework sets out variables that are central to realizing this construct in an operational test.

It is important to recognize that a test imposes certain constraints on the character of the interactions that are created in the assessment and thus on the validity of generalizations from performances on the test to performance in ordinary interactions outside the test. Success in spoken interaction is determined by (a) the nature of the tasks that the interaction requires and the roles in that interaction; (b) the conditions under which the participants are required to perform; and (c) the resources the individual brings to the interaction.

Normally, (a) and (b) are constructed by the participants during the interaction, but on a test these are created by the test developer and the performances are reacted to by the rater. As neither of these are active participants in the interaction in the kind of semi-direct assessment that is contemplated for the TOEFL 2000 test, the constraints and conditions are not interactively created in the test as they would be in ordinary discourse. Even in direct oral testing, as He and Young (1998) show, the interactions may not provide a firm foundation for generalizations because the oral test is itself a unique oral genre with significantly different patterns of turn-taking and topic control. It is important, therefore, that the validity of the test be demonstrated explicitly, rather than assumed.

The study of native speaker oral communication has been dominated by two traditions that have occasionally found common ground.[1] The sociological tradition, originally deriving from the ethnomethodological approach pioneered by Harvey Sacks (see his early lectures, now published in Sacks, 1995), has focused on how oral discourses are a co-construction of all the participants and has taken interaction as its touchstone. While there have been a variety of approaches to the study of interactive discourse, a common thread has been the notion of joint co-construction. Somewhat independently of other work on genre, the sociological tradition has developed a formal approach to typical types of discourse, seeking to identify key interactional patterns of a number of everyday (telephone calls, for example) and more situation-specific (such as court testimony) discourses. To the extent that the sociological tradition has been concerned with meaning, it has been with the meaning these patterns convey.

The other main research thread, speech act theory, has taken meaning as its major focus and has looked at how speakers use language to convey meaning directly and indirectly. Speech act theory has

[1] The psychological work of Herbert Clark (Clark & Schaefer, 1987, 1989), for example, draws on and integrates both traditions.

tended to take function, not form, as its main organizing principle. The contributions of both traditions to the framework are set out in more detail in Part 3.

Both sociological and speech act concepts have found their way into research on second language speaking, though, with one exception, little of this work has been concerned with speaking in academic contexts. The exception is work on the speaking difficulties of international teaching assistants (ITAs). Interest here has been driven less by any theoretical insights than by the fact that the speaking problems of ITAs present immediate and public problems for university administrations. The literature on speaking in academic contexts is dominated by ITA issues. While Waters (1996) found more studies of speaking than of other skills in his survey of academic needs research in North America, five of the seven studies located by Waters were exclusively about ITAs.

Within the pedagogy of second-language speaking, there is also an important body of research on task-based communication, which only recently has been applied directly to testing (Robinson, 1996; Norris, Brown, Hudson, & Yoshioka, 1998; Skehan, 1998a). The importance of this work for TOEFL 2000 speaking specifications is outlined in the discussion on conditions of performance in Part 3.

There is a broad base of knowledge about testing general second language oral proficiency. The success of the oral proficiency interview (OPI) format, pioneered by the Foreign Service Institute (FSI), has informed almost all extant approaches to assessment, even those that do not or cannot implement a full direct interview, such as the Test of Spoken English (TSE®). There is, however, a growing body of critical research that has called into question the basic assumptions that underlie the OPI approach. The validity of the procedure in terms of its implied view of language ability has been criticized on a number of grounds (Lantolf & Frawley, 1985; Bachman & Savignon, 1986; Bachman, 1988; Shohamy, 1990), for example as being poorly articulated and contradictory (McNamara, 1996). Its view of the development of aspects of language ability—for example, control of linguistic form—has been shown to contradict insights from second language acquisition research (Pienemann, Johnston, & Brindley, 1988). The interview format has been shown to be an institutional genre of its own, unlike conversation in many important ways (van Lier, 1989; Perrett, 1990; Young & He, 1998).

In sum, the available research does not provide a firm foundation for constructing a specific test of speaking as part of second language academic communicative competence. Accordingly, the TOEFL 2000 speaking framework represents an attempt to meld the broad base of knowledge about oral communication and more general knowledge about language needs and uses in an academic setting with the growing understanding of the limitations and possibilities of oral testing. This should provide a workable foundation for developing specifications for test construction and directions for research that will better inform future development.

3. Speaking Framework for the TOEFL 2000 Test

Identifying the Test Domain

In keeping with the context of the three other skill areas being assessed (i.e., listening, reading, and writing), the TOEFL 2000 speaking component will be a measure of oral communication in an academic context. That is, the assessment will ask examinees to produce meaningful speech in situations appropriate to academic life.

Examinees will be asked to demonstrate their oral communication skills across a variety of genres, functions, and situations. Tasks will focus on the middle to upper range of English as a second language/English as a foreign language (ESL/EFL) proficiency. The TOEFL 2000 speaking component will simulate realistic communicative situations and include integrated tasks—for example, ones involving listening and speaking, or reading and speaking.

Identifying Task Characteristics and Variables

In order to identify variables that affect the difficulty of various tasks in the TOEFL 2000 speaking component, it is first necessary to define key task characteristics, as described in the Jamieson, Jones, Kirsch, Mosenthal, and Taylor framework paper (1999). This section examines three characteristics in depth: situational features, discourse features, and test rubric.

Situational Features. Speaking is a social activity and, as such, an assessment of it requires attention to the social setting that the assessment tasks intend to simulate. Situational features provide parameters of social settings that serve to delimit the kinds of tasks that would be appropriate to the TOEFL 2000 test. Some of them, perhaps because they trigger differences in the language used, may be associated with differences in difficulty.

Participants

An interaction has at least two participants. When designing simulations for testing purposes, it is necessary to keep in mind certain characteristics of these participants—namely, their roles and relationships. These characteristics are discussed below.

Roles.

Number: It is important that the number of participants vary from task to task because the number can have an effect on the nature of the interaction and the formality of the language. It is also important to note that there is also at least one "other" in a speaking situation, and all TOEFL 2000 speaking component tasks should have clearly signaled participants in addition to the speaker (examinee). There are four general types of number situations that seem appropriate: (a) one-to-one, (b) a small group, (c) a small audience, and (d) a large audience.

One-to-one. The original ethnomethodological literature dealt primarily with this type of interaction with the telephone call being the prototypic example (see the classic Sacks, Schegloff, & Jefferson, 1974). In this case the participants need not be concerned with who the next speaker is—there is only one possibility. The major interactional problem is to properly give and notice turn-taking signals. A number of one-to-one discourses occur naturally in an academic setting: instructor and student about course content and course requirements; student and student about course content and informal matters; student and service provider; etc.

A small group. The prototypic small group interaction is the dinner party (Tannen, 1984, uses this for example, as does Goodwin, 1981) where the turn-taking task is more complex. A few academic situations have the characteristics of this sort of small group; a group of students discussing course content or working on a group project are probably the most typical.

Although both these types are typified by relatively short turns, participants may be granted (and claim) the floor for longer turns, usually taken up as a story (Eggins & Slade, 1997, have an excellent discussion of this).

A small audience. A typical example of this would be a small seminar or discussion group. Turns are often formally assigned by a leader, and all turns except some questions are typically longer than in one-to-one or small group situations. Micheau and Billmyer (1987) and Weissberg (1993) report research on the spoken language of nonnative speakers of English in this type of setting (see Waters, 1996, for a convenient summary of these studies).

A large audience. The university lecture is the prototype here. Turns for the main speaker are uniformly longer than in any of the other situations, and turns for others are few and often determined by the speaker. Several studies of international teaching assistants are summarized by Waters (1996), but all appear to focus on narrow properties of language, rather than on macro discourse, and so do not provide a full understanding of the lecture situation. Most TOEFL 2000 examinees will never be the main speaker in this situation, though a number of graduate students may be.

Age: All participants will be adults. The typical examinee will be a young adult or late adolescent who will have to be able to interact with same-age peers and older adults (though none need be elderly). Age probably plays a smaller role than status differences, or is subsumed in status differences (see discussion of relationships below), in determining characteristics of the interaction.

Gender: There is a rich, and controversial, literature on the difference between the spoken language of men and women. Studies of whether the differences found among native English speakers also occur among nonnative speakers are scarce. While the tasks used in the TOEFL 2000 test should be those equally available to both genders, the rating system should also be sensitive to the possibility that characteristics that differ by gender are being differentially rewarded.

Ethnicity: If it is possible to have some form of simulated interaction in the test, it is important that the variety of English used by the other speaker(s) be one that is easily understood by most North American native speakers. The rating system should not penalize examinees for "standard" varieties different from the general North American standard (see Davidson & Lowenberg, 1996).

Occupation: The occupations of the participants suggested by the task should be those that a student would interact within an academic setting. This would typically include academic occupations (teachers as well as students) and those administrative/clerical and service occupations common in such a setting. It is unlikely, however, that interactions with all occupations found on a campus would be appropriate (such as those with custodial staff).

Educational Level: The educational level of the participants should be appropriate to the occupations as noted above.

Relationships.

Symmetric: These are typically relations with peers—other students—and should be an important part of the assessment. This would include both relatively informal (e.g., discussions over coffee/tea with other students) and relatively formal (e.g., discussions in a tutorial group) relationships.

Asymmetric: Students are part of a number of relationships where some other participants have higher status and power. Indeed, most relations with academic and administrative personnel in an academic setting are of this type. However, students may be in situations where they have higher status, particularly if they are group leaders. There are also situations of ambiguous or changeable status, such as customer-server relations. All these should be sampled on the test.

Topic

In general, the topics covered in the assessment should be appropriate to an academic setting. Typically, students are called upon to speak about topics they are somewhat knowledgeable about, but they are also expected to talk about subjects they know less well because they are just learning about them.[2] Integrated tasks provide an opportunity for the latter type of topic, as the reading and listening stimuli can be expected to provide some background information the examinee can use in speaking (or writing). Not all speaking tasks need be integrated ones, however, and some topics will have to be those that examinees can be expected to already have some knowledge about. These can be divided into two types: those that are so general that most examinees will be able to talk on them and those that are phrased in such a way as to allow examinees to use their idiosyncratic backgrounds. The latter poses, of course, a particular challenge to the raters. When categorized by the source of the candidate's knowledge, there are three sorts of topics:

- subject matter with information supplied by an integrated task,
- subject matter that is widely known, and
- subject matter known by the individual examinee.

[2] Topic familiarity is one of the performance conditions that influences task difficulty (Skehan, 1998a). For further discussion see page 17.

Whatever the source of knowledge, the topic must also be appropriate to an academic setting. There seem to be four important types of topics that occur in an academic setting:

- academic subjects—the standard content of lectures and texts;
- organization of learning activities—discussions of learning strategies and negotiations over procedures and tools;
- rules of academic life—largely bureaucratic discourse over the formal requirements of courses and academic regulations; and
- daily living events that occur on a campus—service encounters (bookstore, medical services) and informal discussions with friends.

Setting

The general framework (Jamieson et al., 1999) identified three broad setting types in an academic milieu: instructional, study, and service settings. These are simply listed here, with a few comments.

Instructional locations.

Lecture hall. As few students will actually be expected to lecture, the most typical speaking activities in this setting would be questions about the content of a lecture.

Seminar room/discussion group. Here students might be expected not only to ask questions but also to make short presentations or to provide sizable responses to a leader's questions.

Laboratory. The most likely speaking tasks here would be to ask questions and seek clarification and to explain findings to a lab instructor.

Study locations.

Library. The typical library speaking activity is to seek advice or help with locating particular information.

Instructor's office. Usually students seek clarification of subject material or rules when they visit an instructor, but sometimes, as in tutorial situations, they may have to make a presentation about their work and respond to questions about it.

Study room. Students may discuss their class material together, seeking and providing clarification.

Service locations.

These are generally characterized by students seeking information or materials.

Purpose

The purpose of the tasks is covered under the following section on discourse features.

Register

It is difficult to characterize register because there is no widely agreed upon framework for labeling different types. As used in the general framework (Jamieson et al., 1999), such variables are intended to characterize the degree of formality of language, much in the sense that Labov (1966) described the social structure of language. Rather than enumerate categories, it seems more sensible to identify situational features that effect the formality of the language. These include:

- the degree to which the language is expected to satisfy formal rules of grammar;
- the degree to which there is some focus on the quality of the language;
- whether spoken contributions are phrasal or sentential;
- how much time to plan utterances is available; and
- whether the larger discourse is constructed interactionally or is largely scripted.

Many of these formality features also play roles in conditions of performance, discussed later in this section.

*Discourse Features***.** The discourse features describe the domain being assessed by the speaking component of the TOEFL 2000 test and are discussed according to generic/pragmatic features and structural features. These features are highly interrelated and are intended to stress the importance of interaction in the assessment of oral language ability. For operational test purposes, they can also be viewed as features of the response since they reflect the productive nature of this TOEFL 2000 test component. (In the listening and reading components of the test, these features are captured under text materials where they help describe test prompts.)

Since there is no single agreed-upon position on discourse, the existing literature serves as a point of departure for articulating the critical role of discourse in the assessment of speaking ability in an academic setting. As research is systematically conducted as part of the TOEFL 2000 test development process, it will be possible to gain a better understanding of the role that discourse features must play in test specification, task development, and score interpretation.

Generic/Pragmatic Features

The TOEFL 2000 test framework looks to Halliday (1973) for categories that relate to the purposes for engaging in tasks and identifies the following six as relevant for international students using English in a North American university: heuristic, instrumental, regulatory, personal, representational, and interactional (Jamieson et al., 1999, p. 15).

In the context of the TOEFL 2000 speaking component, we find it useful to consider the functional analysis of discourse in terms of generic/pragmatic features existing at micro and macro levels. These two levels provide a means for organizing and categorizing a range of language uses for test development purposes.

Micro level: Shorter turns in interaction. One possible source of theory on the micro-interactional aspect of discourse is speech act theory (Austin, 1962; Searle, 1969), which focuses directly on interaction and the types of functions captured in an exchange or series of exchanges. The speaker's or writer's intent is at the heart of the exchange. Searle (1969, 1979), building on Austin, offers five sets of functions: directives, commissives, representatives, declaratives, and expressives. The Council of Europe, drawing on both Austin and Searle, provides a notional-functional syllabus for language teaching (van Ek, 1976) which yields a slightly different set of functions. The six major functions in the Council of Europe scheme are (p. 25):

- imparting and seeking factual information,
- expressing and finding out intellectual attitudes,
- expressing and finding out emotional attitudes,
- expressing and finding out moral attitudes,
- getting things done (suasion), and
- socializing.

Within this scheme, each function is divided into subfunctions.

An alternative approach to micro-analysis of interaction comes from the tradition of ethnomethodology and conversation analysis, which has identified turn-taking conventions and a number of other structural features of interaction, such as adjacency pairs and preference sequences. Essential to this approach is the notion of on-line co-construction of discourse.

Macro level: Extended discourse. A number of approaches to extended discourse exist. The rhetorical function approach focuses on the identification of traditional patterns of rhetorical organization, or patterns of exposition (Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996), that are most relevant to the academic use of language at the university level. Such functions include definition, narration, description, comparison and contrast, procedural/process, etc. These functions are well established in the literature (Shaughnessy, 1977; D'Angelo, 1980; and others; c.f., the rhetorical properties provided in Appendix D of the framework document, Jamieson et al., p. 57). They might be used alone or in combination in a given communicative situation to achieve a specific goal. As Hatch (personal communication, Nov. 7, 1997) indicates, "Giving a report, a speaker might need to describe materials used, narrate what happened, compare and contrast methods used, and do other parts of problem solving. On the other hand, a task might ask for only one type of organization (e.g., procedural: how to access the Internet)."

Essentially similar features of discourse have also been discussed under the heading "genre." The concept of genre stems from work in traditional literary studies, but has appeared in a range of work on or relevant to spoken language, including the ethnography of speaking, particularly in the work of Hymes (1972), literary theory (Bakhtin, 1986), variation theorists in sociolinguistics such as Labov and Waletzky (1967), the Sydney school of systemic-functional linguistics including Martin (1993), Eggins & Slade (see, especially, Eggins & Slade, 1997), and critical discourse analysis (Fairclough, 1995). However as Slade (1996) puts it:

> There are many different approaches to the analysis of genre. What each approach has in common is the recognition that there are, in both spoken and written language, different text-types or genres with their own internal structures, which accord with different social goals. (p. 9)

A number of accounts of generic types exist in the literature. These include:

- face-to-face service encounters,
- service encounters over the phone,
- spoken pedagogic discourse,
- narratives, including those elicited from a sociolinguistic interview,
- opinion texts, and
- gossip.

While the concept of genre is useful in defining the speaking variable, it is worth noting that the above list represents as yet a rather slim research base, and further research will be required. In addition, a number of writers (Paltridge, 1997; Swales, 1990; Jefferson & Lee, 1992; Fairclough, 1995) stress that genres are ideal text types, prototypes, or templates that people orient to but do not necessarily conform to. Although these underlying abstract structures exist, participants negotiate their way through such structures and regularly disrupt them. It will thus be important that the test not take too prescriptive a view of generic structure. Similarly it will be important to include what Bakhtin (1986, p. 78) calls "flexible, plastic, and creative" genres such as personal narrative or expression of opinion in addition to those with a more rigid generic format such as formulaic exchanges in everyday situations.

Structural Features

A number of features of performance will be reflected in the characteristics of the response. These will include features of discourse, lexico-grammar, and phonology/prosody and their communicative effect. Such features are likely to include:

- accomplishment of task (in terms of discursive requirements, coherence, etc.),
- sufficiency of response in terms of length and complexity,
- comprehensibility, including control of phonological and prosodic features,
- adequacy of grammatical resources,
- range and precision of vocabulary,
- fluency (this may not represent a single category; it is currently the focus of considerable discussion, even controversy: c.f., Fulcher, Skehan, and see below), and
- cohesion.

Not all of these features will be relevant to the evaluation of every task, and performance on each may vary relative to other categories from task to task. In other words, performance on different aspects of the task may not vary uniformly but differentially, depending on the attentional focus required of the task, task demands, etc. Thus, particular features may be emphasized for particular tasks, depending on their character and associated performance conditions. Rating will be conducted using appropriate analytic scoring schemes, as discussed later in this section.

Content Visuals. As Douglas pointed out (1997), it is logically and practically impossible to test speaking *apart* from some other skills: the examinee must receive input in some form, whether written or spoken (p. 25). Visual stimuli are an economical and effective way of introducing a situation or a topic of conversation without providing the candidate with a lot of vocabulary (Underhill, 1987). One way of thinking about the role of visuals is in terms of the extent and type of cognitive structuring of the task they provide. This may usefully be conceptualized in Vygotskyan terms (for further discussion of this perspective, c.f., McNamara, 1997).

The most commonly used visuals in a survey of existing ESL tests were maps, single pictures or photographs, multiple-picture sequences, and charts or graphs. The quality of the visual as well as the question or directive itself both would seem to contribute to the difficulty of the task. Giving directions is considered to be a Level 2 task on the Foreign Service Institute (FSI) scale; however, if the map that accompanies the question contains many details and the sites between which the directions should be given are hard to find, this most likely would negatively affect the examinee’s performance on this task. Likewise, to describe a graph that is about an esoteric topic and requires specialized knowledge to interpret would be a more difficult task than describing a simpler graph on a more generalized topic.

The usual test purpose for a picture sequence is to elicit the function of narration (FSI Level 2). Additional tasks at different levels, such as describe (Level 2), hypothesize (Level 3), and state and support an opinion (Level 3), could accompany the picture sequence. Poor quality drawings/photographs or a confusing story line can affect the difficulty level of these tasks. Research into the same task with and without visuals might help to expand our understanding of the difficulty levels of certain question types.

Test Rubric. The test rubric deals with issues related to examinee output in response to the tasks in the TOEFL 2000 speaking component. This section describes the formats of the responses to be elicited with this component, in terms of type and length of response. In addition, this section describes how quality of performance on the various tasks in the speaking component will be assessed, in terms of levels of scores to be assigned, method of scoring, criteria for scoring, and mode of scoring.

Response Formats

Type of response. Three possible response formats can be considered for the TOEFL 2000 speaking component: (a) read aloud, (b) elicited sentence repetition, and (c) constructed response. The first two of these, read aloud and elicited sentence repetition, provide for a more restricted assessment of speaking but lend themselves to computer scoring. With these formats only certain structural features of speaking can be assessed. These features might be scored as present or absent (i.e., right/wrong) or they might be scored by level (see below).

Constructed response formats, on the other hand, allow for a much richer and more varied assessment of speaking, whereby most to all of the structural features of speaking can be assessed. For these formats, scoring would most certainly be done by level, including whether features are simply present or absent (singularly or in combination).

Length of response. In order to make a judgment about a person's speaking ability, it is necessary to have a speech sample to assess. The larger the sample, the more evidence of the person's oral proficiency is presented and is available to be assessed. Issues of face validity and content validity must be taken into consideration in deciding how large a speech sample is necessary in order to give a meaningful score (see Bachman, 1990).

A survey of existing ESL speaking tests showed that a range of 8 to 35 minutes is allotted for examinee speech depending on the purpose and format of the assessment. Some tests have a threshold level which must be achieved in order for the examinee to have an opportunity to respond to additional questions (Underhill, 1987). The format of the FSI and ACTFL Oral Proficiency Interviews (OPI), direct measures of speaking ability, is a one-on-one (or two-on-one) "conversation" lasting 15 to 30 minutes depending on the proficiency level of the examinee (Lowe & Liskin-Gasparro, 1983). The TSE takes about 20 minutes to administer, of which 10 to 12 minutes are available for examinee speech. On the TSE the individual questions are allotted 30, 45, 60, or 90 seconds of response time, depending on the task. The length of time allocated to each task is determined by the level of the question (based on the equivalent task in an ACTFL or FSI OPI) and feedback from pilot testing.

The fact that the TOEFL 2000 test will be administered by computer raises many questions about how the available technology can best be used to assess an examinee's speaking proficiency. Speech recognition systems are not yet capable of accepting all the various forms and pronunciations of words found in open conversation, so at present they cannot provide a direct estimate of the communicative skill level of speakers. However, indirect measures of communicative skill can be derived from performance on tasks like cloze exercises and elicited sentence repetition (Bernstein, 1997). Certain variables, such as pronunciation or grammar, might be able to be captured and scored by speech recognition functions, and these options should be explored.

On the other hand, in recent years the shift in emphasis to language as a tool for communication, and not as an end in itself, has led to the inclusion in oral assessments of such considerations as length of utterance, complexity, rate, flexibility, appropriacy, repetition, and hesitation, as well as the "traditional" categories of accuracy in grammar, vocabulary, pronunciation, fluency, and content (Underhill, 1987). Scoring variables such as coherence and cohesion can be measured only when the sample is long enough to contain several "paragraphs" of speech. Revision of the TSE in 1995 included the elimination of more mechanical short-answer tasks in favor of extended discourse involving more abstract communicative language. The TSE revision team expected that examinee scores on the revised test would be better indicators of communicative ability. They also anticipated that these changes would have a positive washback effect on ESL instructors to emphasize communicative teaching and learning activities in their classes (Douglas & Smith, 1997).

In addition, it has been noted (Douglas, 1997) that a measure of "grammar knowledge," for example, in a communicative situation, may not be strictly possible, since we know so little about how grammar interacts with other components of communicative language ability. Research needs to be done to ascertain the feasibility and desirability of being able to isolate any single component of language ability.

We believe that the TOEFL 2000 examinee should have the opportunity to produce a total of at least 10 to 15 minutes' worth of speech for assessment, and that the sample should provide opportunities to judge the examinee's linguistic, sociolinguistic, discourse, and functional competencies. Because of the practical considerations of a very large volume of examinees, research needs to be done to ascertain how the test might include a combination of short-answer tasks that focus on linguistic competency variables and could be scored by computer with other tasks requiring answers of extended discourse which most likely would need to be rated by humans.

Quality of Performance

Levels of scoring. As mentioned earlier, responses to the assessment tasks can be scored as right/wrong or present/absent, which is referred to as dichotomous scoring; alternatively, they can be evaluated for different levels of correctness, which is referred to as polytomous scoring. At present, it is envisioned that most of the TOEFL 2000 speaking tasks will be polytomously scored.

Speaking tasks that will be scored on multi-level scales lend themselves to partial credit scoring, and from the perspective of item response theory (IRT), to the use of the polytomous partial credit or generalized partial credit IRT models. Further, use of the simpler partial credit model lends itself to many-faceted Rasch modeling (Linacre, 1993), whereby other facets besides task difficulty, such as rater severity, can be considered in the rating process. This is usually done via the computer program FACETS (Linacre & Wright, 1993) or, more recently, the computer program ConQuest (Wu, Adams, & Wilson, 1998); see Bachman, Lynch, and Mason (1995) for an application that made use of FACETS. However, recent work by Tang and Eignor (1997), who used TSE data, seems to indicate that, at least with TSE-like tasks, use of the more complex generalized partial credit model would be necessary.

Finally, speaking tasks that are simply scored right/wrong or present/absent are also amenable to IRT analyses using one of the current IRT models for dichotomously scored items. Since guessing will not be possible on these tasks (i.e., options will not be provided as in multiple-choice questions), either the one-parameter logistic or Rasch model or the two-parameter logistic model are likely candidates.

Method of scoring. Turner and Upshur (1996) discuss the two general "families" of rating scales that have been used in language assessment when scoring is done using multi-level scales. The first family of scales is more or less generic in nature (i.e., the same rating scale(s) can be applied to different tasks); this family includes both holistic and analytic scales. A holistic scale of speaking ability would describe a number of levels of overall speaking ability which are intended to be applicable to any sample of speech. The score level descriptors might refer to any of a number of criteria and expected levels of performance on these criteria, but these are generally incorporated into one overall description at each level. Analytic scales, on the other hand, have more than one overall scale, so that it is possible to separately rate a number of different attributes or criteria. The separate analytic ratings may then be combined to provide a more general measure of speaking ability, but they need not be. In some situations, the same set of analytic scales is applied to different tasks; in other situations, a common subset of scales is applied to the tasks along with some additional scales that are unique to each task. The first approach is by far the most common with analytic scoring (see Bachman et al., 1995, for an application of analytic scoring).

Experience with the TSE holistic scoring procedure suggests that such an approach masks the apparent difference in difficulty among TSE speaking tasks. That is to say, even though the TSE tasks appear to differ in difficulty (see Powers, Schedl, Wilson-Leung, & Butler, in press), the holistic procedure and generic rubric produce score means on the various tasks that are close to equivalent, thereby masking difficulty differences. Since for the overall proficiency scales to be developed in the TOEFL 2000 project, position on the scale is related to task difficulty, the holistic approach seems questionable. The use of analytic scales is unlikely to cause this problem to the same degree as does use of the holistic approach.

With the analytic scaling approach, the number of levels or rating points for each scale is typically developed based on a logical analysis of the possible levels of separation afforded by using the particular scales with the tasks. Also, nothing precludes the use of analytic scales for assessing a set of tasks from having differing numbers of levels within the different scales.

The second family of rating scales described by Turner and Upshur (1996) includes the primary trait and multitrait rating scales. According to Turner and Upshur (1996), "These [scales] are less reflective of general theories [about the development of language abilities] than the holistic and analytic scales, but are more influenced by an analysis of task demands" (p. 56). Such scales are designed to indicate how well an individual satisfies the requirements of a given task. Primary trait scoring uses one score scale with a number of levels which are specific to the task. Turner and Upshur provide an example of a primary trait scale from Cohen (1994) for the following task:

> "Some people believe that a woman's place is in the home. Others do not. Take one side of the issue. Write an essay in which you state your position and defend it." (p. 57)

The primary trait levels then indicate: (a) whether the writer has, in fact, taken a position (level 1); (b) whether reasons are given (level 2); and (c) whether or not each of the reasons has been elaborated (level 3).

Multitrait scales are similar to analytic scales in that performance is rated using more than one scale. They are similar to primary trait scales in that each score scale is specific to the particular task.

The task-specific levels of either primary trait or multitrait scales can be defined either logically or empirically. In most applications to date, the specific levels have been defined logically. Fulcher (1987, 1996; see also Skehan, 1998a, in press) has expressed concern that logically defined scales reflect what theorists think happens in communicative situations, and not what actually happens. Turner and Upshur (1996; see also Upshur & Turner, 1995) have investigated an empirically based procedure for defining the specific score levels of primary trait scales. This approach, referred to as EBB scaling by Turner and Upshur (the scale is Empirically derived, requires Binary choices by raters, and defines the Boundaries between score levels), while labor intensive, should be further explored for TOEFL 2000 speaking tasks if a primary trait approach is chosen. However, the viability of this and other approaches from a financial perspective would need to be assessed before it could be implemented. North (1995, 1996) provides an excellent discussion of a number of issues in the development of empirically based rating scales (see also North and Schneider, 1998).

Criteria for scoring. When using either the primary trait or multitrait approaches, criteria for scoring are task specific. For the holistic and analytic approaches, where the resulting scales are generic, it makes more sense to talk generally about scoring criteria. These criteria are either combined into some sort of general statement at each scale point with the holistic approach or defined in separate scales with the analytic approach.

Bachman et al. (1995) used five analytic scales that measured the following components of speaking ability: pronunciation, vocabulary, coherence, organization, and grammar. An expanded list now under consideration for the speaking component of the TOEFL 2000 test includes: pronunciation, vocabulary, cohesion, organization (coherence), grammar (grammatical accuracy), comprehensibility, and fluency. In addition, length of responses may be considered, provided it can be broken down into meaningful levels to define the points on the (analytic) scale.

Mode of capturing responses and scoring.

> Capturing responses: It is assumed that the speaking section of the TOEFL 2000 test will be an indirect measure of speaking ability in that the prompts will not be directly delivered by an interlocutor nor will the responses to the tasks be scored immediately by judges. It is assumed that the tasks themselves will be delivered via computer. The oral response can be captured by the computer. If the visual record is seen as important, then the speech act associated with the response will need to be either captured by the computer or videotaped. The amount of hard disk space that will have to be used in the computer capture of the visual record may constrain which method will be used.
>
> Human/computer scoring: Given the current state of speech recognition technology (see Burstein, Kaplan, Rohen-Wolf, Zuckerman, & Lu, 1997), it is assumed that most responses to speaking prompts will be scored by human raters. Developments in this field will need to be carefully monitored.
>
> When the state of speech recognition technology is such that it can be considered for the scoring of speaking tasks, a model similar to the one under consideration for writing tasks might be feasible. With this model, the speech sample would be scored once by a human rater and once by the computer. Another model worthy of consideration, which would make use of analytic scales, is one where the computer would be used to score features such as sentence length, while humans would be used to evaluate features such as intelligibility.

Integrated Tasks. The portion of the TOEFL 2000 test that involves integrated skills is a very important one for assessing an examinee’s speaking proficiency because it is logically and practically impossible to test speaking apart from some other skills. In the academic environment that is the context for TOEFL 2000, students are constantly in circumstances in which they listen to or read something and then are required to respond either orally or in writing. Although there are situations like a political address or an academic lecture where one person speaks and others only listen, the stimulus for even these speech acts has undoubtedly come from something in either written or spoken form. Therefore, the integration of skills in the TOEFL 2000 measurement has strong face validity.

Speech is normally an interactive mode of communication in which the participants switch roles between listener and speaker frequently within a conversation. Because of practical limitations, it is unlikely that the TOEFL 2000 speaking component will be a direct measure of speaking ability. However, even in a semi-direct test of oral proficiency, speech cannot exist in a vacuum; the examinee must receive input in some form, either spoken or written, in order to know what task he/she is being required to perform.

In the TSE, the candidate both hears the instructions and reads them in the test book and then records the answer on cassette tape for future assessment. Because of their relevance to an academic context, many of the functions that the TSE assesses might appear on the TOEFL 2000 speaking component. These functions include describing, persuading, narrating, hypothesizing, summarizing, comparing, giving directions or instructions, recommending, and stating and supporting an opinion. Any of these functions could be linked to either written or spoken stimulus material. For example, the examinee could read a passage or listen to a short lecture and then be asked to make an oral summary of it. An examinee could listen to an extended conversation between two students about going on a field trip for an astronomy class, look at a map in the test book, and then be asked to leave directions to the planetarium on another student's voice mail system. Many of the integrated tasks that have been suggested by the TOEFL 2000 writing team would also probably work with spoken responses.

Finally, it should be noted that because many of the tasks envisioned for assessing speaking will not be able to be scored immediately, this portion of the test could not be delivered in the same adaptive fashion as certain sections of the current computer-based TOEFL test. However, research could be carried out on adapting speech tasks based on examinee performance on the listening tasks, under the assumption that examinees who do poorly on these ought not to get extremely difficult speaking tasks.

Factors Influencing Task Difficulty. The difficulty of an assessment task, as well as the quality of the performance as judged against specific criteria, can be shown to be affected by the cognitive demands of the task and the conditions under which individuals carry out the task. Recent work in Britain by Skehan and his colleagues, building on earlier work by Brown, Anderson, Shilcock, and Yule (1984), has attempted to explore the ways in which varying qualities of task ("task characteristics") and conditions of performance ("task implementation conditions") influence various measures of output on speaking tasks. Skehan (1998a, p.174) recognizes the potential of this work for defining an ability/difficulty continuum using item response theory (IRT) approaches. Table 1 summarizes research findings to date as reported by Skehan which have explored factors affecting task difficulty in some detail. For example, in Brown et al.'s research, narrative tasks in response to picture stimuli were found to differ in difficulty depending on how many people or narrative elements were involved in the pictures: a car crash involving a truck and a car was easier than a crash involving three cars that looked rather similar.

Table 1
Factors Influencing Task Difficulty

Easier condition	**Harder condition**
Small number of participants, elements	Greater number of participants, elements
Concrete information and task	Abstract information and task
Immediate, here-and-now information	Remote, there-and-then information
Information requiring retrieval	Information requiring transformation
Familiar information	Unfamiliar information

Based on Skehan, 1998a: 174, Table 7.2

A further range of influences on difficulty is summarized in Table 2. The impact of these factors is proposed tentatively as it has yet to be confirmed by solid research (for example, recent research suggests that the impact of visual factors is more complex than is proposed here. See more on content visuals below). "Surprise elements" in the table below means that new information (e.g., as a stimulus to a narrative) is provided in the course of carrying out a task, which implies that the speaker has to change tack without warning; this limits the possibility of producing a rehearsed performance.

Table 2
Other Possible Factors Influencing Task Difficulty

Factor	**Likely Impact on Difficulty**
As time pressure increases	Increase
As opportunity for control (e.g., to ask for clarification, to delay, to rephrase, to modify—but not completely change—the task goals) increases	Decrease
As surprise elements increase	Increase
As visual support increases	Decrease

Based on Skehan, 1998a: 176-177

Other aspects of performance conditions may affect task difficulty in a more complex way, in that tasks which elicit particular cognitive processes may have measurable impact not on task difficulty as a whole but on particular aspects of performance. The following components of cognitive processing in performance on speaking tasks have been identified in Skehan's research:

1. Tasks and performance conditions which direct attentional resources to form and rule. These tasks and conditions may induce either "risk-avoiding" or "risk-taking" behavior, yielding variation in measures of accuracy and complexity of language, respectively.

2. Tasks and performance conditions which focus attentional resources on meaning and real-time processing, yielding variation in measures of fluency.

Findings from research to date on the impact of task characteristics and performance conditions respectively on particular aspects of performance are summarized in Tables 3 and 4. The aspects are *fluency* (as measured by such things as pausing, silence, repetition, false starts, reformulations, etc.), grammatical *accuracy* (based on proportion of error), and *complexity* (based on measures of subordination, for example, number of clauses per T-unit or c-unit).

In Table 3, the ascending arrows represent increases in measures of the relevant component: for example, increasing familiarity of task material may lead to performances which are measurably more fluent, but not more accurate or more syntactically complex. “Differentiated outcomes” refers to problem-based tasks in which there are a number of possible solutions to the problem; for example, where candidates are asked to give opinions about the best course of action. It seems that when speakers are thinking about the various possibilities, they have fewer attentional resources available for monitoring correctness, and their language resources are “stretched,” so to speak, to express the complex meanings they intend.

Table 3
Task Characteristics and Their Impact on Attentional Resources

Nature of Task	Resulting Change	Aspect
Differentiated outcomes	⇑	Complexity
Well structured tasks	⇑	Accuracy
Familiar material	⇑	Fluency
On-line computation involved	⇑	Complexity

In Table 4, the impact of varying conditions of planning is considered. Candidates may be given detailed advice as to how to plan: for example, they may be told to focus on correctness, or on the overall organization of their response to the task; or they may be given planning time but no advice on how to use it (“undetailed planning time”).

Table 4
Performance Conditions and Their Impact on Attentional Resources

Performance Condition	Resulting Change	Aspect
Undetailed planning time	⇑	Accuracy
Different lengths of planning time	⇑ ⇑	Fluency Complexity

This research has complex implications for assessment. The factors set out in Tables 1 and 2 could act as difficulty drivers in the definition of an ability/difficulty continuum. The factors set out in Tables 3 and 4 could be used to establish the specifications for particular tasks, with care being taken to ensure that the task demands and performance conditions do not engender a competition for attentional resources. These factors could also be used to establish appropriate rating criteria for each task depending on the character of the task and its performance conditions. A complex program of research relevant to the goals of the TOEFL 2000 project is thus called for. An initial project exploring the potential of the work of Skehan under formal test conditions has begun (McNamara, Elder, & Iwashita, 1998).

4. Technical Issues Affecting the Measurement of the Speaking Construct

In order to ensure that the TOEFL 2000 test provides the best assessment of an examinee’s English oral language proficiency, it will be necessary to research each step of the test development process, including developing the prompts, delivering the test items, obtaining a speech sample, rating the speech sample, and distributing the score to the score users. Each step entails technical considerations which must be addressed.

Speaking is an interactive skill, and so even though the TOEFL 2000 speaking component by necessity will have to be an indirect measure of speech administered by computer, the more the tasks can simulate interactivity, the more true the test will be to the construct. Given the rapid development of the computer field, it has been predicted that what is considered to be “high-end” equipment today will become commonplace in three to five years. In terms of delivery of the test items, the hardware currently available in 1998 in this category has very high-quality sound and video capability. The improvements in technologies for the compression, storage, and transmission of audio and video images will help to make their use practical at the network level. On the other hand, although it might be argued that video presentation of the items would be more realistic and “interactive,” research is needed to weigh the advantages and disadvantages of such a delivery system. Certainly the cost and time involved in creating a very large pool of video items will have to be weighed against the verisimilitude factor.

On the other end of the interactive scale, existing speech technologies, such as automatic speech recognition (ASR), can be used to assess pronunciation of short segments of speech by matching the examinee’s utterances with preprogrammed models. Although this question type is not very communicative, it might be useful for measuring “threshold” proficiency. Certainly future developments with ASR should be followed closely.

With regard to obtaining the speech sample, microphones and videocameras in the computer for recording the speech and video images of the examinee are becoming more readily available. Whether or not it is necessary or even helpful to have a video image of the examinee responding to questions is something that will need to be researched.

As for rating the speech sample, because of the limitations of ASR (discussed above), humans will need to be involved in scoring the assessment. Therefore, it must be possible to transmit the speech/video sample to the raters. The techniques being used for on-line scoring of written essays might be adapted for scoring speech samples. One advantage of having the speech sample scored on line is that the raters do not have to gather in one location and can work on a more flexible schedule. If, however, this technique means that a lot of special equipment would be needed for raters to access the speech/video files, this could have an impact on the cost of the test.

With regard to reporting of scores, at the moment the capacity does not exist to do “real-time,” automated scoring for longer speech segments. Since it is envisioned that a number of the speaking items will be integrated with items testing the other skills, it will have to be determined how the lack of immediate scoring on the speaking segment would affect the reporting of scores for the rest of the TOEFL 2000 test. One advantage to having the speech sample on disk is that it could then be accessible to admissions officers and other university personnel who might want to hear a real speech sample as opposed to just receiving a score number for an examinee.

It seems obvious that the best way to meet these technological challenges to the TOEFL 2000 project is through a strong collaborative effort by researchers, test developers, technical support staff, and outside consultants as necessary.

5. Research Agenda

In preparing the framework document for the TOEFL 2000 speaking component, team members were unable to locate review articles in a number of important areas that clearly would have helped inform the development of the framework. Hence, as detailed below, we call for a number of literature reviews to be conducted. These literature reviews will undoubtedly help refine certain elements contained in the framework.

The speaking team plans to follow the sequence of task design and development steps outlined in the research agenda for the writing team. Hence, research activities related to final task design, defined in the second section below, will need to be conducted during the cycle outlined by the writing team. We also specify (see the third section below) a number of topics that, if researched, would help to inform task design. While these are important to the final task design process, they are not as crucial as the activities listed in the second section.

Finally, in discussions with the Natural Language Processing (NLP) group at ETS about ways in which that group might support the development of the TOEFL 2000 speaking component, a number of intermediate to long-range research topics were identified. These topics are briefly listed in the final part of this section and depend on the NLP group first establishing a platform whereby speech data can routinely be collected. Before such work is initiated, a careful review of work performed by other technology-oriented researchers in the field of speaking will need to be undertaken.

Literature Reviews

We anticipate the need for the following literature reviews:

1. a review of the literature on speaking needs[1, 2]—range of general vocabulary needed;
2. a review of recent research on register;[1]
3. a review of recent research on oral genres;[1]
4. a review of recent research on cognitive demands of tasks; and[1]
5. a survey of rating scales used in existing speaking tests.

Research Activities Related to Task Design

Other research activities that will need to be undertaken include the following:

1. Confirm that variables identified by Skehan et al. (1998) on performance conditions that are contained in the framework do predict task difficulty.
2. Confirm that variables contained in the framework under discourse features, particularly structural features, predict task difficulty.
3. Perform research that would lead to a decision whether to use primary trait or analytic scales for the speaking assessment. Can primary trait scoring be made cost effective?

[1] Include native and nonnative speakers

[2] Separate review for undergraduates and graduates

4. Perform research on the extent to which content visuals increase or decrease task difficulty.
5. Perform research on the extent to which interactivity can be incorporated in speaking task design.
6. Perform research on the role that rater cognition plays in determining task difficulty.

Other Research Activities to Inform Task Design

Additional research activities include:

1. A needs analysis of speech production required of undergraduate and graduate students (pending outcome of A1);
2. Given the level of interactivity likely to be featured in the speaking component, an analysis of the consequences for the validity of the test;
3. Research on whether an initial threshold-level assessment of speaking could be created, based partly on read-aloud and elicited repetition activities;
4. A comparison of the effects of capturing responses on video vs. audiotapes; and
5. Research on incorporating the work on attentional resources into the analytic scoring process.

Research Agenda for the NLP Group

The Natural Language Processing (NLP) group at ETS will need to undertake the following research activities:

1. Speech genre characterization studies via manual analysis,
2. Speech genre characterization studies via automated tools,
3. Written versus spoken language characterization studies,
4. Nonnative, minority, and gender-based language characterization studies,
5. Creation of an auditory idiom knowledge base,
6. Automated speech sample scoring studies,
7. Limited domain concept-to-speech synthesis research,
8. Independent speech synthesis research,
9. NLP-based automated speech recognition research,
10. Statistical language model-based automated speech recognition research, and
11. Combined knowledge source automated speech recognition research.

6. Speaking Assessment in TOEFL 2000

The current TOEFL test does not include an oral communication component. Although the TSE is offered, this assessment of general speaking ability is designed primarily for teaching assistants and medical professionals. Thus, the speaking component of the TOEFL 2000 test is a highly important development. The integrated skills portion of TOEFL 2000 will simulate realistic communicative situations since it is logically and practically impossible to test speaking apart from some other skill (the examinee must receive input in some form, whether written or spoken). This new section of the test will meet score users' expressed need for information about examinees' English oral language proficiency in an academic context.

We expect that the introduction of an oral communication component will have a positive washback effect on the ESL teaching community. By using constructed-response items, which are less likely to be coachable, in the TOEFL 2000 speaking component, we will encourage students to learn to communicate orally—not to learn a skill simply to do well on a test.

Finally, unlike the current TOEFL test, which yields a single score, the TOEFL 2000 speaking component will provide descriptive information about the examinee's level of proficiency in each skill area assessed. These descriptions will be developed through analyses of specific variables that affect the difficulty of the assessment tasks.

References

Austin, J. L. (1962). *How to do things with words.* New York: Oxford University Press.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*, 149-164.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

Bachman, L. F., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal*, *70*, 380-390.

Bakhtin, M. M. (1986). *Speech genres and other late essays* (V.W. McGee, Trans.). Austin: University of Texas Press.

Bernstein, J. (1997). Speech recognition in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment – Proceedings of LTRC 96* (pp. 534-537). Jyväskylä, Finland: University of Jyväskylä.

Brown, G., Anderson, A., Shilcock, R., & Yule, G. (1984). *Teaching talk: Strategies for production and assessment.* Cambridge: Cambridge University Press.

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

Clark, H. H., & Schafer, E. F. (1987). Collaborating on contributions to conversation. *Language and Cognitive Processes, 2,* 19-41.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259-294.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed). Boston: Heinle and Heinle.

D'Angelo, F. (1980). *Process and thought in composition* (2nd ed.). Cambridge, MA: Winthrop.

Davidson, F., & Lowenberg, P. (1996, December). *Language testing and world Englishes: A proposed research agenda*. Paper presented at the 3rd Conference of the International Association of World Englishes, Honolulu, HI.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series Report No. 8). Princeton, NJ: Educational Testing Service.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English Revision Project* (TOEFL Monograph Series Report No. 9). Princeton, NJ: Educational Testing Service.

Eggins, S., & Slade, D. (1997). *Analysing casual conversation.* London and Washington: Cassell.

Fairclough, N. (1995). *Critical discourse analysis*. London: Longman.

Fulcher, G. (1987). Test of oral performance: The need for data-based criteria. *EST Journal, 41,* 287-291.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13,* 208-238.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. (TOEFL Research Report No. 54). Princeton, NJ: Educational Testing Service.

Halliday, M. A. K. (1973). *Explorations in the functions of language.* London: Arnold.

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

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A.W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam and Philadelphia: John Benjamins.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication.* New York: Holt, Rinehart & Winston.

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.

Jefferson, G., & Lee, J. (1992). The rejection of advice: Managing the problematic convergence of a "trouble-telling" and a "service encounter." In P. Drew & J. Heritage (Eds.), *Talk at work*. Cambridge: Cambridge University Press.

Labov, W. (1966). *The social stratification of English in New York City.* Washington, DC: Center for Applied Linguistics.

Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experiences. In J. Helm (Ed.), *Essays on the verbal and visual arts. (Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society).* Seattle, WA: University of Washington Press, 12-14.

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *The Modern Language Journal, 69,* 337-345.

Linacre, J. M. (1993). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (1993). *A user's guide to FACETS: Rasch Model computer program.* Chicago: MESA Press.

Lowe, P., & Liskin-Gasparro, J. (1983). *Testing speaking proficiency: The oral interview.* Washington, DC: Center for Applied Linguistics.

McNamara, T. F. (1996). *Measuring second language performance.* London and New York: Addison Wesley Longman.

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics, 18,* 446-466.

McNamara, T. F., Elder, C. A., & Iwashita, N. (1998). *Investigating predictors of task difficulty in the measurement of speaking proficiency*. TOEFL 2000 Research project. Princeton, NJ: Educational Testing Service.

Martin, J. (1993). Genre and literacy – modeling context in educational linguistics. *Annual Review of Applied Linguistics 13*, 141-172.

Micheau, C., & Billmyer, K. (1987). Discourse strategies for foreign business students: Preliminary research findings. *ESP Journal, 6,* 87-97.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* (NFLRC Technical Report #18). Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

North, B. (1995). Scales of language proficiency. *Melbourne Papers in Language Testing, 4,* 60-111.

North, B. (1996). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. Unpublished doctoral thesis, Thames Valley University.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*, 217-262.

Paltridge, B. (1997). *Genre, frames and writing in research settings*. Amsterdam: John Benjamins.

Perrett, G. (1990). The language testing interview: A reappraisal. In J. H. A. L. de Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities*. Clevedon, Avon, UK: Multilingual Matters.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10,* 217-243.

Powers, D. E., Schedl, M. A., Wilson-Leung, S., & Butler, F. A. (in press). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing.*

Robinson, P. (1996). Task-based testing, performance-referencing and program development. *University of Queensland Working Papers in Language and Linguistics, 1,* 95-116.

Sacks, H. (1995). *Lectures on conversation.* Oxford: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simple systematics for the organization of turn-taking for conversation. *Language, 50,* 696-735.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts.* Cambridge: Cambridge University Press.

Shaughnessy, M. (1977). *Errors and expectations.* New York: Oxford University Press.

Shohamy, E. (1990). Language testing priorities: A different perspective. *Foreign Language Annals, 23,* 385-394.

Skehan, P. (1998a). *A cognitive approach to language learning.* Oxford: Oxford University Press. [See esp. Ch. 7, "Processing perspectives on testing," pp. 153-183.]

Skehan, P. (1998b). Task-based instruction. In W. Grabe (Ed.), *Annual Review of Applied Linguistics,18*, 268-286.

Skehan, P. (in press). Task characteristics, fluency, and oral performance testing. In P. Skehan (Ed.), *Thames Valley Working Papers in Applied Linguistics*, Vol. 5, pp. 1-18.

Slade, D. (1996). *The texture of casual conversation: A multidimensional interpretation*. Unpublished doctoral thesis, University of Sydney.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tang, K. L., & Eignor, D. (1997). *Concurrent calibration of dichotomously and polytomously scored TOEFL Items using IRT models* (TOEFL Technical Report No.13). Princeton, NJ: Educational Testing Service.

Tannen, D. (1984). *Conversational style: Analyzing talk among friends.* Norwood, NJ: Ablex.

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *Australian Review of Applied Linguistics*: Series S, No. 13. *The Language Testing Cycle: From Inception to Washback* (pp. 55-79). Melbourne: ARAL.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

van Ek, J. A. (1976). *The threshold level for modern language learning in schools.* London: Longman.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context.* (TOEFL Monograph Series Report No. 6). Princeton, NJ: Educational Testing Service.

Weissberg, B. (1993). The graduate seminar: Another research-process genre. *SP Journal*, *12*, 23-35.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER Conquest*. *Generalised Item Response Modelling Software.* Melbourne, Victoria: ACER Press.

Young, R., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam and Philadelphia: John Benjamins.

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

58713-005536 • Y60M.750 • 253716 • Printed in U.S.A.