TOEFL®

Monograph Series

MS - 9
MAY 1997

Theoretical Underpinnings of the Test of Spoken English Revision Project

Dan Douglas

Jan Smith

Educational Testing Service

Theoretical Underpinnings of the Test of Spoken English Revision Project

Dan Douglas and Jan Smith
Revised with assistance from Mary Schedl,
Gena Netten, and Mark Miller

Educational Testing Service
Princeton, New Jersey
RM-97-2

To obtain more information about TOEFL products 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 Test of English as a Foreign Language program development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts from the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions were invited 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 ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. 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 TOEFL CBT will take 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

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Abstract

The purpose of this paper is to lay a theoretical foundation for the revision of the Test of Spoken English. The revision project was undertaken in response both to concerns expressed by researchers and score users about the validity of the TSE test and to a request by the TOEFL Committee of Examiners to make the TSE test more reflective of current thinking on the assessment of oral language skills. In the paper, we first discuss communicative competence as a basis for understanding the nature of language knowledge, followed by discussions of sociolinguistic and discourse factors which influence spoken language performance. Test method characteristics which influence test performance are also discussed, as are types of evidence necessary for establishing reliability and validity of the revised TSE test. The paper concludes with a discussion of the implications of the theory for the interpretation of examinee performance with regard to academic and professional contexts of language use.

Acknowledgment

The authors would like to thank the Test of Spoken English Committee and staff for their discussion and support during the writing of this theoretical underpinning of the TSE revision project. They are particularly grateful to Mary Schedl, Gena Netten, and Mark Miller for the work they did in revising the document.

Table of Contents

List of Figures

Introduction

The revised Test of Spoken English is a test of speaking ability designed to evaluate the oral language proficiency of nonnative speakers of English who are at or beyond the postsecondary level of education. It was developed on the premise that language is a dynamic vehicle for communication, driven by underlying competencies which interact in various ways so that communication can be effective. The revised TSE test will be validated for use as a general screening test for teaching and research assistant selection and professional certification, primarily in medical and allied health fields.

From the introduction of TSE in 1979 until the formation of the TSE Committee in 1992, the TOEFL Committee of Examiners was responsible for overseeing the test. In carrying out this role, the TOEFL Committee of Examiners regularly suggested and recommended changes in the test format, content, and scoring. In 1988 it held a "blue-sky" discussion for the purpose of coming up with ideas on how to make the TSE test more responsive to current thinking on the assessment of oral language skills. It was clear from these discussions that major revision of TSE should be undertaken. In February 1991, a panel of oral language experts was invited to ETS to consider the theoretical implications of possible TSE revisions. Their recommendations prompted the Committee of Examiners in April 1991 to request that TSE staff create a prototype test from Sections 1, 4, 6, and 7 of the current test and place a high priority on an increase in examinee speech samples.

The prototype test was presented to the Committee of Examiners in October 1991 and was recognized as an improvement over the current test in eliciting communicative speech samples. The Committee suggested that illocutionary (functional) competence and sociolinguistic competence be addressed in the score descriptors. Communicative activities tested in the items should attempt to capture speech events which require more abstract speech, in addition to those which require more concrete speech. In January 1992, a second panel of experts was convened to examine the prototype, suggest modifications and new item types, and discuss the scoring design for the revised TSE test. In response to their suggestions, in April 1992, the Committee of Examiners suggested (1) that TSE level descriptors be better defined, (2) that a description of the construct for the test design be drafted consistent with the score scale and descriptors, (3) that the scoring rubric incorporate the concepts of linguistic competence, sociolinguistic competence, and discourse (textual) competence, (4) that the assessment of the "success of the message" be considered, (5) that positive backwash be kept in mind, and (6) that the question, "What is a particular task testing?" be asked continually. In addition, research into the effect of changing the scales on score reporting to users was recommended.

As can be deduced from this brief history, TSE revisions have been prompted by both pragmatic and theoretical concerns. Generally, revisions have tended to make the test more "user friendly" and to make the items more general for a more varied audience/clientele. However, researchers outside ETS and score users themselves have, over the years, expressed concerns that reflect uneasiness about the validity of the TSE test with regard to its stated purpose. Criticisms have touched upon a number of areas:

- the difficulty some examinees have in speaking into a tape recorder (Bailey, 1987; Plakans & Abraham, 1990).

- the unnaturalness of having only one chance to hear and respond to questions (Plakans & Abraham, 1990).

- the lack of relationship of the content (e.g., describing a bicycle) to academic and professional work (Bailey, 1987; Byrd, 1987; Plakans & Abraham, 1990; Gallego, Goodwin & Turner, 1991).

- the lack of attention to the measurement of listening skills, interpersonal skills, and professional expertise (Plakans & Abraham, 1990).

- the difficulty of using a "native speaker" standard as a scoring template when information is lacking as to how native speakers would perform on the test (Barrett, 1987; Godfrey & Hoekje, 1990).

- the possibility that examinees from different language backgrounds may have different mean scores owing to scorer bias (Barrett, 1987; Constantino, 1988).

Some of the above concerns cannot easily be addressed, given the administrative parameters within which test developers must work: the test must be a semi-direct instrument, delivered by tape, with a time limit of about 20 minutes, and there can be only one test for all audiences. However, it did seem to the TSE Committee that it would be possible to work toward making the relationship between the test content and academic/professional work clearer, by broadening the scoring rubric to address more than just grammatical competence, and by defining level descriptors in terms of communicative language ability. In light of this discussion, then, a framework for revising the TSE test ought to include a thorough specification of the construct of speaking ability in non-test situations and a more thorough specification of the scoring rubric and level descriptors to include a wider range of competencies.

Since its initial meeting in Toronto, August 11-13, 1992, the TSE Committee has worked with ETS staff to draft new formats and test items and create a new scoring rubric for the test. Its writing of a Statement of Theory and Purpose for the revision of the TSE developed into the present paper, which is intended to detail the theoretical background for the new test. The committee has overseen pilot administrations and a formal research study of the new formats, items, and scoring rubric. ETS decided upon an implementation date of July, 1995, for revised TSE.

In this paper, we will first discuss theories of communicative competence, since this concept is central to current understandings of what it means to know a language. Next, we will consider aspects of sociolinguistic and discourse theories which influence language performance: speech acts/functions, coherence, and cohesion. Third, we will take up the topic of test method, because it has become clear in recent years that test performances are the result of an interaction between a test taker's ability (or communicative competence) and the methods employed to measure that ability. Fourth, we will review various types of evidence for reliability and validity and suggest studies that might be undertaken with regard to these aspects of the revised TSE test. Finally, we will examine the implications of the theory presented here for the interpretation of examinee performance on the revised TSE test. While the focus of this paper is on oral language assessment, the implications of the discussions of theories of communicative language ability and of the influence of test method characteristics on performance are important in consideration of language assessment across modalities.

Theories of Communicative Competence

Theoretical definitions

In this paper we use a number of terms from the domain of testing and measurement: *proficiency, ability, knowledge, competence,* and *performance.* It is important to be clear about the meanings of these terms, for some of them are used interchangeably in the literature, while others reflect quite different theoretical positions in the field.

Proficiency and *Ability.* Traditionally, in language testing, the term proficiency has meant the ability to use a language for some purpose, irrespective of how that language was learned (Carroll, 1961). Thus, proficiency and ability are synonymous, although ability may be seen as a learner internal construct, while proficiency is the observable manifestation of it. In this paper, ability is the preferred term, since it is associated with current theories about communicative language ability, though for the sake of stylistic variety, the two terms may be used interchangeably.

Knowledge and *Competence.* These two terms are often used synonymously (e.g., Bachman, 1990), and the term construct is a more precise psychometric equivalent. However, it is important to remember that knowledge is not conceived of as merely something learners possess to some degree or other, but is something to be used. One can thus "know" facts and ideas and their interrelationships, procedures and strategies for using them, and procedures for monitoring and evaluating success (Bennett et al., October, 1991). In this paper, in order to emphasize the notion that knowledge implies use, we have distinguished the terms competence and knowledge by saying that competence includes both *knowledge* of various aspects of language and the *ability* to use that knowledge for some communicative purpose.

Performance. We use this term in two ways: in the technical sense of "direct language performance test" and in the more general sense of the observable manifestation of language ability that we wish to measure with our test. In other words, "talk." It is important to emphasize that the revised TSE test is not a direct language performance test, not only because of operational constraints such as tape-recorded prompts, but also because such direct tests tend to blur the distinction between the ability to be measured and the method of measurement. Performance is the result of an interaction between the test taker's *language ability* and the *test methods.* If no distinction is made between ability and method, the performance cannot be interpreted (Bachman, 1990). This is so because, in *direct* performance tests, the performance *is* the construct to be measured. In referring to TSE as a "semi-direct" test, then, we are not only acknowledging that it is not a live, interactive performance test, but also that, ideally, we wish to distinguish between ability and method in the test taker's performance.

Frameworks for communicative language ability

The theoretical construct of communicative competence refers to a speaker's knowledge of and ability to use language. It is the basis of a number of approaches to language teaching and testing that seek to reflect real-life language use. Such approaches include communicative teaching and testing, but also the European notional-functional approach, the American Council on the Teaching of Foreign Languages/ Interagency Language Roundtable (ACTFL/ILR) proficiency movement, Krashen and Terrell's natural approach, the Silent Way, Community Language Learning, and others. In the testing field, while

Educational Resources Information Center (ERIC) lists 74 entries under the heading of "communicative language tests," most of these are discussions of theory and research, and few are actual tests.

Empirical studies conducted by language testers and others reveal that language proficiency is multicomponential (Canale & Swain, 1980; Canale, 1983; Savignon, 1983; Duran *et al.*, 1985; Sang *et al.*, 1986; Bachman, 1990 [Chapter Four]); however, what is extremely unclear is precisely what those components may be and how they interact in actual language use. ETS has entered into this discussion of communicative competence by hosting conferences (Stansfield, 1986) and sponsoring research (Clark & Swinton, 1980; Powers & Stansfield, 1983; Duran *et al.*, 1985; Henning & Cascallar, 1992). Yet, despite all this interest and activity, the concept of communicative competence is not well understood, and there are many fewer communicative tests than there are communicative teaching approaches and materials. Therefore, in order to construct a test of communicative language ability, test designers must come to their own agreement as to which competencies the test will measure and how these competencies will be defined.

The term "communicative competence" has been invoked for nearly three decades now to express the notion that language competence involves more than Chomsky's (1965) rather narrowly defined "linguistic competence." As Hymes (1971, 1972) originally formulated the concept, communicative competence involves judgments about what is systemically *possible*, psycholinguistically *feasible*, and socioculturally *appropriate*, and about the *probability of occurrence* of a linguistic event and what is entailed in the actual accomplishment of it. It is important to remember, too, that for Hymes, "competence" is more than knowledge: "Competence is dependent upon both [tacit] *knowledge* and [ability for] *use*" (Hymes, 1972). This is an important point since, while it joins knowledge and ability as factors underlying communicative performance, it also distinguishes them from it. It is clear from this formulation that communicative competence is not to be confused with "communicative success." Speakers may have sufficient knowledge and ability to address a communicative task and yet, due to factors outside their control, not accomplish the goal.

This is pointed out by Hornberger (1989), who recounts the story of using her Spanish communicative competence to get a driver's license renewed in Peru. She was in fact successful in getting the license, despite many setbacks and frustrations, but makes the following observation:

> . . . it should be pointed out that my communicative competence in these events resides not in the fact of my obtaining the license I set out to get (which would be a kind of performance criterion), but rather in the knowledge and ability that allowed me to suit my language use to the events in which I found myself. Even if my goal of obtaining the license had not been achieved, it would not necessarily have meant that I was not communicatively competent for those events, but rather that . . . one or more of the external factors were insurmountable.
>
> Hornberger, 1989, pp. 228-229

Perhaps the most important aspect of the focus on communicative competence is its distinction between the underlying traits of knowledge and ability and the communicative events and tasks to be addressed. Bachman (1990) points out that the problem of failing to distinguish ability from behavior is that a language performance becomes the trait to be measured, and there is an infinite set of possible performances. Generalization and validation will be impossible. Only by distinguishing competence from performance, by taking into account the interaction between attributes of the test taker and attributes of the test task, can truly authentic tests be constructed. Of course, it may not be possible in practice to discretely disentangle testee ability and test method effects for any individual test taker, since an individual language performance is the result of a perhaps unknowable interaction between the test taker and the test method (see Hudson, 1994). Only in large-scale generalizability studies can the relative effects of ability and test method be distinguished (see Bachman, 1990). Nevertheless, it is important to recognize that what is being attempted with the revised test is not to measure a speaking performance *per se*, but rather to make inferences about test takers' communicative language ability *on the basis of a test performance.*

Others have since reformulated Hymes' notion of communicative competence (Savignon, 1972; 1983; Canale & Swain, 1980; Canale, 1983; Duran *et al.*, 1985; Bachman, 1990). The current, most well-known framework is that of Bachman (1990). It is somewhat amended by Bachman & Palmer (1996), which postulates two components of "communicative language ability," language knowledge and strategic competence. Figure 1 illustrates the components of communicative language ability.

Figure 1: Components of communicative language ability

Language Knowledge

Grammatical Knowledge

Knowledge of Vocabulary
Knowledge of Morphology and Syntax
Knowledge of Phonology

Textual Knowledge

Knowledge of Cohesion
Knowledge of Rhetorical or Conversational Organization

Functional Knowledge

Knowledge of Ideational Functions
Knowledge of Manipulative Functions
Knowledge of Heuristic Functions
Knowledge of Imaginative Functions

Sociolinguistic Knowledge

Knowledge of Dialects/Varieties
Knowledge of Registers
Knowledge of Idiomatic Expressions
Knowledge of Cultural References

Strategic Competence

Goal Setting

Identifying Communicative Situation or Test Task

Assessment

Determining what language components are needed to successfully respond to the communicative situation or complete the test task
Evaluating the correctness or appropriateness of the response

Planning

Selecting required elements from language knowledge

Based on Tables 4.1 and 4.2 in
Bachman and Palmer, 1996

For Bachman, then, **language knowledge** consists of *grammatical knowledge* (knowledge of vocabulary, morphology, syntax, and phonology), *textual knowledge* (knowledge of how to structure and organize language into larger units, particularly in terms of *cohesion* and *rhetorical/conversational organization*), *functional knowledge* (knowledge of how language forms are related to communicative goals, particularly the *ideational, manipulative, heuristic,* and *imaginative* functions), and *sociolinguistic knowledge* (sensitivity to dialects, registers, idiomatic expressions, and cultural references). **Strategic competence**, which Bachman formulated following the work of Faerch and Kasper (1983), includes three components, as reformulated by Bachman and Palmer (1996): *goal setting* (identifying the communicative situation or test task), *assessment* (relating language knowledge to the communicative situation or test task, and subsequently assessing the success of the response), and *planning* (deciding how to utilize language knowledge to successfully respond to the situation or complete the test task).

As Weir (1990) points out, models such as those described above "provide a potentially useful framework for the design of language tests, but it must be emphasized that they are still themselves in need of validation" (pp. 8-9). The next section summarizes an empirical study which lends preliminary support to some of the concepts discussed earlier and speaks directly to the issues of this paper.

An empirical study of the nature of communicative competence

There has not been a great deal of empirical research to date on the construct of communicative competence or communicative language ability (Bachman & Palmer, 1982; Duran et al., 1985). In a fairly comprehensive study of the nature of communicative competence, Henning and Cascallar (1992), in a study which is labeled "preliminary" by the researchers themselves, set out to investigate the "most salient components" of theories of communicative competence, and in particular, their interrelationships. They took the view that there is no single model of communicative competence likely to serve all purposes equally, just as there is no single type of map that will serve all travel needs. Specifically, Henning and Cascallar studied the relationships among 18 aural/oral communication variables, 18 reading/writing communication variables, the TOEFL test, the Test of Written English (TWE®), and the TSE test.

In their analysis of the results, Henning and Cascallar found wide variation in performance across tasks, social register, and pragmatic function, providing, for the researchers, "preliminary support for a view that *communicative language ability* is situation-specific in that it depends largely on the particular communicative context" (p. 12, emphasis added).

Regarding the ability of the original TSE test to predict oral communicative performance, they found, somewhat to their surprise, that grammar and fluency were slightly better overall predictors than comprehensibility. They speculated that this may be due to the possibility that "comprehensibility as a construct is more heavily dependent on abilities present simultaneously in both the producers and the receivers of language communication; whereas grammar and fluency are more highly associated with the producers alone" (p. 16).[1]

[1] There is disagreement among scholars about whether fluency is more associated with production than with reception. On the one hand, fluency can be objectively measured, as was done by Henning and Cascallar, as speech rate; on the other hand, speakers are certainly perceived as "fluent" or "disfluent" by their listeners. The issue deserves further research.

They concluded their discussion with three recommendations, which are paraphrased below:

1. Any valid assessment of communicative competence can take place only within a well-articulated framework for the elicitation and rating of communicative language abilities, and care must be taken to ensure that appropriate contexts are defined before communicative language tests are devised.

2. The evidence strongly suggests that traditional measures of knowledge and use of appropriate language structure, as represented in the structure components of the TSE test, are not empirically unrelated to ratings of communicative competence, and we should not abandon concern for assessment of language structure.

3. Based on the importance of the strategic domain in accounting for variance in the communicative competence construct, the assessment of communicative language ability should ideally include the measurement of such abilities as fluency of cognition and expression, density of information transfer, and compensatory strategies.

It must be remembered that the study upon which these interpretations and recommendations are based was a preliminary one, and that much more research of this type is needed before recommendations of a general nature can be made. However, the study is a valuable one for the theory of communicative competence, suggesting as it does the importance of context of situation, language competence, and strategic competence in the construct of communicative language ability.

Implications of communicative competence theory for TSE revision

In considering the theories of communicative language ability in concert with the operational constraints outlined above, the designers of the revised Test of Spoken English decided to focus on the four types of **language knowledge** proposed by Bachman and Palmer, but employing slightly different terminology[2] *linguistic competence, textual competence, functional competence,* and *sociolinguistic competence*) and on **strategic competence**. These competencies are defined, in the context of the revised TSE test, as follows: *linguistic competence* is the knowledge of and ability to use phonological, morphological, syntactic, and semantic structures of the language. *Textual competence* involves the speaker's knowledge and ability in the areas of organizing information in a coherent manner and making effective use of cohesive devices to help the listener follow the organization of the response. *Functional competence* is the speaker's ability to select language functions to reasonably address the test task. *Sociolinguistic competence* involves the speaker's ability to demonstrate an awareness of audience

[2] The terminology employed in the revised TSE documentation, "linguistic" instead of "grammatical" and "competence" instead of "knowledge," reflects decisions made early in the development process, after which changes in terminology would have been administratively difficult to make. The intent, however, is to employ the concepts of *communicative language ability* in the spirit of Bachman, 1990 and Bachman & Palmer, 1996.

and setting by selecting socially and culturally appropriate language and register (level of formality). *Strategic competence* is evident in the speaker's ability to address the test task, to supply alternative responses to the question, to deal effectively with communication breakdowns, and to compensate for gaps in language ability.

The designers of the revised TSE test believe that the ability to communicate effectively involves the interaction of such competencies, but acknowledge that at present there is no effective method available to determine how one competence interacts with another. Each competence is a part of a speaker's ability to communicate effectively, but no one item on the test can effectively assess them all, and most items will assess more than one competence. The goal of the revised TSE test is to be sure that examinees provide a language sample which is extensive enough to judge all the language competence areas selected for the test.

Sociolinguistic and Discourse Theories

Existing theories of sociolinguistics and discourse provide a number of useful concepts which can be used to structure the test items on the revised TSE test and thus shape desired examinee responses. These concepts include *speech acts* or *functions, coherence,* and *cohesion.*

Speech acts/functions

Austin (1962) first introduced the notions of *locutionary, illocutionary,* and *perlocutionary* acts to distinguish between the form of an utterance, its intended purpose, and its effect. Searle (1965, 1969, 1976) further developed Austin's work by analyzing illocutionary or speech acts into the five macro-classes of *representatives, directives, commissives, expressives,* and *declarations.* These distinctions allow speakers to represent their beliefs about the world, direct others to perform actions, commit themselves to performing actions, express psychological states, and perform ritual acts. Searle (1975) revealed the existence of a continuum of directness in which individual speech acts — (all of which contain the same speaker intention) — may be placed, ranging from "Can you pass the salt?" to "I don't think you salted the potatoes." Labov and Fanchel (1977) elaborate on this work in their examination of therapeutic speech by placing indirect speech acts in conversational context, and Grice (1975) proposes a cooperative principle governing the selection of speech acts within a particular context.

The notional/functional syllabus outlined by van Ek (1975) and Wilkins (1976) specifies a method for organizing second and foreign language teaching according to the functional (speech act) and notional (topic) needs of language learners. This method allows language teachers to provide input based on the situational needs of language learners by focusing on the purpose of particular utterances spoken within particular settings. Although originally a separate effort from the work on speech act theory, the notional/functional syllabus provided a way to apply the work being done by discourse analysts to the language classroom in the 1970s and early 1980s.

Van Ek uses the term "threshold level" to identify the minimal level of language functions required of second language users. Second language learners who are at or beyond the threshold level are able to express at least one exponent (example of linguistic structure) for each of the functions in the following categories: imparting and seeking factual information (*identifying, reporting, correcting*), expressing and finding out about intellectual attitudes (*agreeing, accepting, stating, inquiring*), expressing and finding out about emotional attitudes (*expressing pleasure, hope, fear, want*), expressing and finding out about moral attitudes (*apologizing, approving, expressing regret*), getting things done (*suggesting, requesting, advising, inviting*), socializing (*greeting, introducing, congratulating*). More advanced speakers may have many exponents for each function, which can be used differentially according to the demands of a particular situation.

Coherence and cohesion

Brown and Yule (1983) relate *coherence* in spoken text to the listener's assumption that the vocabulary and grammatical structures they hear are connected in some way. It is impossible for the speaker to be completely explicit about every aspect of the message, so all listeners must fill in gaps in the context of a spoken text, but speakers can make this task easier. In English, this can be done by selecting appropriate appropriate exponents of functions which can be easily interpreted within the context, relating aspects of the message to the listener's knowledge of the world as well as the context of what is being said, and providing predictable cues for the listener to use in applying inferencing strategies. The listener must participate in the process by simultaneously processing individual components of the utterance while making predictions about their composite meaning. In addition, the listener must use available background and specialized knowledge related to the situation to determine how and why particular structures were used and fill in missing information conveyed through inference.

Halliday & Hasan (1976) define *cohesion* in spoken text as an instance where the interpretation of one component of the discourse is dependent on that of another. *Textual cohesion* can take the form of *reference, substitution, ellipsis, conjunction,* or *lexical ties. Referential cohesion* is concerned with the way in which the same thing or class of things are identified more than once within a text. Substitution involves the replacement of noun, verb, or clause by grammatical structures such as *one, do,* or *so,* while ellipsis refers to the omission of a noun, verb, or clause in a subsequent utterance. *Conjunctive cohesion* refers to the use of additive, adversative, causal, and temporal structures which tie concepts together within the utterance. Examples include simple conjunctions such as *and* or *but,* and more complex adverbials, conjunctions, and phrases such as *yet, though, however, because, for this reason, after that,* and *previously. Lexical cohesion* refers to the replacement of vocabulary items by other items from the same lexical class. All of these phenomena help to tie the structure of a text together so that the listener can relate information in one part of the utterance to information in another part of the same utterance.

Implications of sociolinguistic and discourse theory for TSE revision

In light of the previous discussion of sociolinguistic aspects of language use, the revised TSE test will employ items, or tasks, which are specified in terms of seven contextual characteristics: test interviewer, audience, setting, topic, purpose, function, and visuals. The *interviewer* poses the tasks on the test, and will be modelled on a friendly, face-to-face interviewer speaking in the first person. The *audience* will be specified, sometimes implicitly, and will be the interviewer, friends, colleagues, people one would ordinarily meet in daily life, or people with shared interests. The communicative *setting* will include such features as physical location, a medium of communication, and/or a time. The *topics* of the test tasks, while they may not be specifically academic or professional in nature, will be intended to elicit responses that will engage the testees' communicative language ability. Appropriate topics will include current events, politics, the environment, education, culture, technology, food, books, and entertainment. A *purpose* for the communication will be specified, sometimes implicitly, to give the testees' some reason to carry out the task, apart from the purpose of language assessment. Communicative *functions* will be specified, explicitly or implicitly, for each task. These are discussed in more detail below. Finally, each TSE form will contain four *visual contexts*: a map or spatial plan, a chart or graph, a picture sequence, and a schedule.

The purpose of all these characteristics is to provide appropriate and rich sociolinguistic and discourse features, in terms of test task characteristics, and to enable testees to engage their communicative language ability in responding to the test tasks. (For an explanation of test task characteristics and their relationship to language ability, see the following discussion of the influence of test method on test performance in Section IV.)

The TSE Committee decided to employ *language functions* as the central structural framework for the revised Test of Spoken English on the assumption that such functions are universal but that their realization is language/culture specific.[3] The revised TSE test will employ language functions which do not depend upon an interactive response in the context of the Test of Spoken English. These functions include *describing information presented graphically*; *narrating from visual materials; summarizing; giving directions based on visual materials*; *giving instructions*; *recommending, advising, or suggesting*; *giving and supporting an opinion; comparing or contrasting; hypothesizing, predicting, or speculating; persuading, apologizing, or complaining.* Each task on the TSE test is intended to elicit at least one of these functions, and each form of the test must elicit each function at least once and not more than twice. Functional competence on the TSE test will be evaluated according to the "reasonableness" of the speaker's choice of functional exponents in addressing the task within the context specified by the test item.

Coherence will be evaluated on the revised TSE test according to the speaker's ability to choose culturally appropriate ways of organizing functions to directly or indirectly convey intended meaning and to refer effectively to available world and background knowledge. *Cohesion* on the Test of Spoken English will be evaluated according to the speaker's ability to use cohesive markers of reference, grammatical substitution, ellipsis, conjunction, and lexical substitution to tie together utterances within a text.

[3] Some empirical support for this assumption may be found in a preliminary study of some discourse features of responses to a prototype version of the revised Test of Spoken English (Lazaraton & Wagner 1994).

Influence of Test Method on Test Performance

The notion of context as it applies to performance on language tests is operationalized in language tests as *test method facets* or *characteristics.* Bachman (1990) presents succinctly the relationship between language ability and test method as they influence test performance:

> While it is generally recognized that [test specification] involves the specification of the ability domain, what is often ignored is that examining content relevance also requires the specification of the test method facets.
>
> Bachman, 1990, p. 224

In this paper we assume that any factor changed in the test method can lead to changes in examinee perceptions of the situational context, and thus lead to changes in performance on the test.

Contextual factors

For the past 15 years or so, a recurring theme in applied linguistics and language testing has been the role of context in the acquisition of the language code itself, in the development of communicative ability, and in language test performance. (see Ellis & Roberts, 1987; Bachman, 1990; Larsen-Freeman & Long, 1991). "Context" has been defined in various ways by researchers, and has been associated with such notions as *situation, setting, domain, environment, participants, task, purpose, content, topic, schema, frame, script,* and *background knowledge.* The number of factors relevant in the study of the effect of context on language performance has been growing, and Hornberger (1989) argues that "the very essence of a communicative event [is] that it is situated in a real, physical, cultural, historical, and socioeconomic context" (p. 228), and that all these aspects of context feature in the acquisition of communicative competence.

Features of context relevant to consideration of spoken discourse include the following from Hymes (1974): persons participating (interlocutors) or listening (audience), including their relative status, states of knowledge, and purposes; topic; setting (place, time, physical relations of the interactants); channel (speech in this case); code (language, dialect, style); form (genre); event (the larger event in which the communicative event in question takes place, such as a university class, a church service, a convention); key (the evaluative or emotional tone of the communicative event, such as anger, joy, dispassion, outrage); and purpose (the result of the communicative event intended by the participants).

In language testing, contextual features are realized as *test method* characteristics: features of tests such as the *environment* (personnel and location), *organization* (the salience, sequence, and relative importance of the sections of the test), *instructions* (clarity and explicitness), *format* (means of delivery), and *language* (level of vocabulary, degree of new information, degree of abstractness, topic, genre). These method characteristics "can be seen as analogous to the features that characterize the context of situation, or the speech event" (Bachman, 1990, p. 111). There is no doubt that context, or test method, does play a role in influencing language test performance; the problem has been arriving at a common understanding of the nature of method characteristics, and then determining specifically how various characteristics influence performance. One problem has been the merging of the notion of context with the its interpretation by language users.

Selinker and Douglas proposed the notion of discourse domain to clearly distinguish the external features of context from the internal interpretation of them by the test taker (Selinker & Douglas, 1985; Douglas & Selinker, 1985, 1989). The discourse domains hypothesis is of help in understanding individual variation in interpretation; that is, language test takers might respond to method characteristics differentially, and, from the point of view of the "language community" or a "native speaker," perhaps erroneously. The result might be, for example, interpreting a listening comprehension item as "work talk" when in fact the test writers had intended to present a casual conversation of a more social nature. This mismatch in domains might lead to serious misunderstanding, frustration, or simply a difficult time getting the "gist" of the item.

Another source of variation, it can be hypothesized, would be in making the link between the internal context and units of communicative language ability. Once a test taker settles upon an interpretation of the item, relevant aspects of communicative language ability would be brought to bear in response, whether for comprehension or for production. However, a learner might not possess the grammatical, psycholinguistic or sociolinguistic "wherewithal" to carry out the required linguistic task, to employ effective strategies for dealing with the situation, or to retrieve appropriate language items. Any of these difficulties might produce flawed test responses.

Interlanguage variation and speech accommodation theory

Interlanguage (Selinker, 1972) is the generally accepted term for the internal second language system that a learner has constructed at any given temporal point. It distinguishes learners' second language development both from their first languages and from the target language. The first researcher to provide data to show that contextual factors which are external to the speaker are associated with shifts between styles of *native-speaker* speech was Labov (1972). In the case of *second language* learners, Ellis (1986) cites two causes of contextual variation: (1) differing linguistic environments and (2) the situational factors of task, topic, and interlocutor. Tarone (1988) surveyed current research on the investigation of these situational factors and lists up to 50 studies (e.g. Dickerson, 1975; Dickerson & Dickerson, 1977; Schmidt, 1977; Beebe, 1980) which demonstrate that second language users vary significantly in proficiency according to who is being spoken to, what task is called for, and what is being spoken about. Tarone (1989) goes on to suggest that situational factors of context often interact within a particular situation. Young (1989) further points out that examining any one contextual variable without controlling for the affects of others produces unreliable and inconclusive results.

Tarone (1979, 1982, 1983) asserted that the cause of interlanguage variation was the degree of attention paid to *language form*, but beginning with Tarone (1985), modified her position to state that the nature of the discourse required by a particular task interacts with speaker attention to both *form* and *function* to cause variation in the accuracy and choice of linguistic structure. Rounds (1991) examined contextual variation according to task using the Speaking Proficiency English Assessment Kit (SPEAK®) test scores from her institution and found that only the grammar scores conformed with the assumption that decreased cognitive demand causes a decrease in attention to form. Rounds suggests that communicative pressure interacts with attention to form, causing a decrease in accuracy in tasks that involve a high cognitive load.

Speech accommodation theory, as outlined by Beebe and Giles (1984) (supported by empirical evidence in Beebe, 1977; 1980; Beebe & Zuengler, 1983), asserts that interlanguage speakers modify their speech in relation to that of native speaker interlocutors due to the need for social approval and communicative efficiency. Bell (1984) claims that the speaker's awareness of a listener's assessment of his or her social status determines the extent to which this linguistic accommodation occurs. Yule and Tarone (1990) suggest that speakers must also take into account how much their interlocutors know about the world in order to best communicate necessary information to them.

It is widely agreed that situational context has a significant effect on the interlanguage produced by second language learners, but the available theoretical frameworks have not been successful in separating the effects of task from topic, topic from interlocutor, and interlocutor from task. Agreement is also lacking on how broadly or narrowly to define the categories of task, topic, and interlocutor.

Studies of the manipulation of test method characteristics

Several empirical studies of speaking ability and oral interaction have involved, directly or indirectly, the manipulation of test method characteristics. In a study of SPEAK test topics, Smith (1992) found a significant difference in performance between a general topic test and a mathematics field-specific test, in favor of the field-specific test, although, at the same time, she found no differences between performance on the general topic test and either a chemistry or a physics field-specific test. She speculates that this finding may be due to a somewhat higher overall English proficiency level among the subjects whose specialty was math. In analyzing a "profession-specific" test of speaking ability, McNamara (1990) found that raters tended to ignore a genre-related "appropriateness" criterion in favor of a "resources of grammar" criterion in assessing "communicative effectiveness," and concludes that even when tests are designed to be communicative, raters may not make use of the full range of data available in making their judgments. Zuengler and Bent (1991) investigated patterns of conversational participation between native speaker and nonnative speaker interlocutors, and found that relative control of content knowledge had a significant influence on conversational involvement. They conclude that "relative content knowledge should be recognized as an important dynamic affecting many NS-NNS conversations" (p. 410). Douglas and Selinker (1993, 1991) have explored this area in two empirical studies, following Smith (1989), in which the goal was to investigate the effect of manipulating method characteristics (Bachman, 1990), such as the instructions, vocabulary, contextualization, distribution of information, level of abstraction, topic and genre, on outcomes.

These studies point out the complex nature of the interaction between aspects of context and language test performance. It would appear that there is some "threshold" of contextual specificity (Bachman [personal communication] has suggested the term "situational authenticity") required for a second language learner to respond differentially to context, or in Selinker and Douglas's terms, to trigger the engagement of a relevant discourse domain in an interlanguage learner. There is also a similar threshold of interlanguage development necessary before the learner can effectively respond to differences in context. In other words, it seems that a certain number of situational features and contextualization cues, in the form of test method characteristics, are necessary in order for different contexts to make a difference in test performance[4].

Implications of test method characteristics for TSE revision

Test method facets most directly influence test taker performance in sociolinguistic competence and strategic competence. Evaluating sociolinguistic competence requires attention to register and the ability to use language appropriately within a context. Evaluating strategic competence requires attention to the cognitive strategies that test takers use to *enhance* the effectiveness of communication, *compensate* for gaps in language ability, or *repair* breakdowns in communication. Enhancement strategies make an already adequate response even more effective. Compensation strategies fill gaps in language skills that would otherwise lead to communication breakdowns. Repair strategies may or may not succeed in addressing serious communication difficulties.

The revised TSE test examines strategic competence by focusing on communicative tasks that call for the application of general information within a variety of contexts. These contexts test for success of the message by eliciting language in relation to a range of audiences, settings, topics, purposes, functions, and visuals, which vary the communicative requirements for particular test items. Although it will not be possible to isolate the effects of the different contextual variables from each other, it is expected that the range of situations presented in each version of the test will tap a similar range of situational abilities in the examinee. Examinees will have the opportunity to demonstrate their ability to successfully address a communicative task or, if necessary, to make effective use of strategies to compensate for their inability to directly address the task.

The revised test cannot address the difficulty some examinees have in speaking into a tape recorder or the unnaturalness of having only one chance to hear/read and respond to questions. Nevertheless, it is crucial that test items be designed with a "bias-for-best" criterion in order to be as fair as possible to examinees within the specified physical constraints of the test. For this reason, the revised TSE test will make every effort to clearly present the situational requirements and accompanying expectations of examinee performance for each test item. In addition, the revised TSE instructions will encourage maximal examinee performance by providing both written and spoken prompts for each communicative task.

Because of the relationship between the external and internal context of the test for the individual examinee, test designers will pay particular attention to the ways in which *test organization, instructions,* and *format* may affect examinee performance. Each section of the test will provide *contextualization cues* through a variety of test task characteristics to alert the speaker to the situational context of the item

[4] In addition, there is evidence that context is more of a factor in the production and interpretation of language at higher levels of proficiency (Clapham, 1991). If this turns out to be the case, then contextualization is especially important in a test such as the Test of Spoken English, which is designed for learners at higher levels of proficiency.

and to encourage the engagement of available discourse domains in order to provide an effective level of response. An example of this would be to provide spoken/written dialogue as a prompt to examinee responses or spoken/written description to flesh out the identity of the person to whom examinee speech is to be delivered. As a consequence of these revisions, examinees will need to be informed of the revised nature of the Test of Spoken English and its new requirements for appropriate responses to the context of each item.

Theoretical Issues in Test Scoring and Rating

In one way or another, the spoken performances of the test takers must be rated. It is almost axiomatic that, because language use is a multicomponential phenomenon that requires interlocutors to negotiate meanings, no two listeners hear the same message. This aspect of language use is a source of bias in test scores, and leads language test developers to limit severely which features of a performance they require raters to attend to in making their ratings. They hope that by focusing attention only on such features as pronunciation, grammar, fluency, and comprehensibility, for example, raters will not be influenced by the manifold other features of the discourse. For example, Underhill (1987) observes that it may not be feasible for raters to keep track of more than three or four of these criteria at one time. Yet our current view of the complexity of communicative language use would seem to compel at least the attempt to evaluate language performances in light of the broad range of factors that underlie the performances.

Canale (1984) has articulated the challenge to language testers presented by the broadening of our understanding of factors that constitute language knowledge and affect language use:

> Just as the shift in emphasis from language form to language use has placed new demands on language teaching, so to has it placed new demands on language testing. Evaluation within a communicative approach must address, for example, new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures to emphasize interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgmental nature.
>
> Canale, 1984, p. 79

Underhill (1987) notes that in addition to the "traditional" categories of accuracy in grammar, vocabulary, pronunciation, fluency and content, "the shift in emphasis to language as a tool for communication, and not as an end in itself" leads to the inclusion of such considerations as length of utterance, complexity, rate, flexibility, appropriacy, repetition, and hesitation (p. 96). How can test developers take into account this complex range of features, while at the same time avoiding increasing the cognitive load on raters to an unacceptable level?

A solution to this problem has been the development and use of impressionistic, or holistic, ratings of language performance. Weir (1990) describes impressionistic scoring as that which "entails two or more markers giving a single mark based on their total impression of the [performance] as a whole" (p. 63). Weir provides a cogent review of studies comparing analytical rating schemes for writing (later applying this discussion to the testing of speaking), which award a score to each of a number of criteria, with holistic ratings. He concludes that "holistic evaluation is obviously to be preferred where the primary concern is with evaluating the communicative effectiveness of candidates' [production]" (p. 67), and particularly recommends a "band marking" system, such as that employed by the English Language Testing Service (Carroll, 1980, pp. 137-139). Weir suggests, however, that a problem with such a scale, which describes "typical" levels of achievement, is that "it does not cater for learners whose performance levels vary in terms of different criteria" (p. 67). On the other hand, it could be argued that language itself is multicomponential, with various types of knowledge and abilities underlying performance in unspecifiable interaction. In such a situation, a scale that describes typical knowledge and ability criteria at each level may be the best that can be done. Ingram and Wylie (1993) argue for the use of such a band scale in the International English Language Testing System (IELTS):

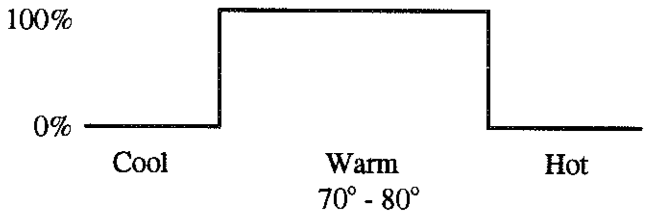
> . . . if a theory of language can link the sociocultural, semantic, and linguistic systems (Halliday, 1978), if the test tasks require candidates to integrate elements from the various systems (which the tasks must do if they reflect actual language use), and if patterns of coincidence or coemergence of elements from the various systems can be observed in learners' use (which the analysis of the interlanguage of thousands of learners in the development and use of such scales suggests), then it does seem appropriate to combine different types of criteria in the scale descriptors.
>
> Ingram & Wylie, 1993, pp. 221-222

The key to combining criteria in holistic band descriptions for speaking test performance is the rating process. The boundaries between categories will necessarily be vague, described by such terminology as almost always, generally, somewhat, comprehensible, or incomprehensible. It will not be possible to define such terms so concretely that every rater will agree precisely on what they mean. However, it will be possible, despite this uncertainty, for raters to make judgments about performances and arrive at a level of agreement that will satisfy psychometric reliability requirements. This is so because language users routinely deal with such uncertainty in natural language use, manipulating such concepts as warm, cool, dirty, hard, and comprehensible by means of "fuzzy logic." Spolsky (1989) acknowledges this "fuzziness" in his discussion of the conditions for second language learning, and develops an analogy with "expert systems," which are computer models designed to emulate human experts in solving complex problems such as communication. Such expert systems, Spolsky points out, "are designed to deal with cases of uncertainty, with fuzzy situations. They work by gathering as much information as possible and then making a decision" (p. 223). The following discussion of "fuzzy set theory" is intended to explore this notion in a bit more detail. The goal is to better understand the role of human raters in the test scoring process.

Fuzzy set theory

In standard set theory, an object or state was said either to belong to a set, say, "warm," or not belong to the set, which might mean either "cool" or "hot" (Kosko & Isaka, 1993). Figure 2, below, illustrates this state of affairs:

Figure 2: Standard Set



Thus, if an object were within the 70° to 80° range, it would be categorized as "warm." At the same time, an object outside that range would be either "cool" if it were below 70° or "hot" if it were over 80°. In other words, in standard set theory, an object or state is either 100% in the set, or 100% outside it.

In language, as in many other areas of life, however, things are not so clear cut. A problem in assigning ratings to communicative speaking test performances is that the boundaries of judgment are vague and hard to define. Suppose we wanted to assign scores in bands indicating *almost always comprehensible, generally comprehensible, somewhat comprehensible, generally incomprehensible*, and *almost always incomprehensible.* In assigning language performances to the above categories, an individual performance belongs to a set only to some extent. For example, a performance may be 75% "somewhat comprehensible" and 25% "generally incomprehensible." In such a case, it can become a matter of policy to assign the performance to the "majority" category, "somewhat comprehensible." A more intractable problem would arise when a performance was judged to be 50% in one category and 50% in another; i.e., the glass is either half full or half empty. However, in natural language use, this apparent dilemma may be vacuous. In other words, a human rater, if "allowed" to respond to a speaking performance naturally, can learn to avoid the "half full/half empty" dilemma because the human neural network responsible for interpreting communicative input can "push" the performance into one category or the other (Kosko, 1991). Humans do this naturally by weighing the multidimensional characteristics of the performance in a way that may be different from that of other raters, yet "close enough" to have allowed that rater to have dealt successfully with language input all her life. It is also true that test takers do not exhibit uniform, incremental increases in their ability to speak a second language across all tasks and in all contexts. In other words, there is an enormous amount of uncertainty involved in the assigning of ratings of speaking ability. Yet it can be done, because in fact it is done.

In an interesting small-scale study, Barnwell (1989) found that naive native speakers (of Spanish), using guidelines in the form of the ACTFL oral proficiency scales, but without training in their use, were able to agree on which of four tape recorded Spanish interviews represented the best performance, the second best, and so on. Although Barnwell found that the naive raters varied quite a bit in the absolute ratings they gave each performance, he also found clear evidence of patterning in the ratings; there was not a random scatter of ratings. Thus, though interrater reliability was not high for the untrained raters, Barnwell suggests that "Further training, consultation, and feedback could be expected to improve reliability radically" (p. 158). It seems clear that given simple rules or guidelines, such as may be found in many existing rubrics for rating spoken performances, raters can make use of "negative evidence" provided by feedback and consultation with expert trainers to calibrate their ratings to a standard. They need to be encouraged to make use of their native speaker ability in assigning performances to score bands (see, for example, Hadden, 1991).

In summary, the point of this discussion is to help understand why it is so difficult, indeed impossible, to entirely specify the defining features of score categories such that raters will be able to place a performance unerringly into the correct one. Raters need some simple rules or guidelines for judging input and negative evidence regarding their output in order to make consistent ratings of spoken language performances. To this end the revised TSE test will employ a holistic band scale rating system, based on the interaction of speaker knowledge and ability attributes, and attributes of the test task. Each task/item on the TSE test will be rated on a five-point scale that guides raters in considering the effectiveness of the communication, the test taker's success in performing the intended task, and the use of communication strategies, taking into account features of linguistic, textual, functional, and sociolinguistic competence.

Evidence for Reliability and Validity

A number of issues need to be dealt with in *reliability* and *validity* studies that will accompany the development and introduction of the revised TSE test. This discussion is directly tied to the ETS document, "Developing measurement approaches for constructed-response formats" (Bennett et al., 1991). "Constructed-response formats" are those in which the test taker's response is generated rather than selected from a short list of options. The Test of Spoken English is a constructed-response test. Tests of this sort present new and special challenges to researchers and analysts: what kind of information should be extracted from constructed responses, how can scoring rubrics be designed to provide this information, how can this information be interpreted and used, how can IRT models be adapted to constructed-response formats? These issues influence the methods used to evaluate test reliability and validity.

The TSE test is also a performance-based assessment instrument, but as Linn, Baker, and Dunbar (1991) point out, that does not automatically make it more valid than a multiple-choice, paper-and-pencil test. They suggest that, as we expand forms of assessments, we also "expand the criteria we use to judge the adequacy of assessments" (p. 16). The criteria they discuss include the following: (1) evidence about the *intended and unintended effects* of the assessment on the ways teachers and students spend their time, (2) evidence about *biases against racial/ethnic minorities* in both the selection of performance tasks and the scoring of responses, (3) the *generalizability of performance* on the test to non-test tasks, (4) investigation of the *cognitive complexity* of the test tasks and the nature of the responses they elicit, (5) a determination that the tasks selected for the test are *worthy of the time and effort* test takers are asked to devote to them and that the *breadth of coverage* is adequate for generalizations to be made, (6) that score reports and interpretations are *meaningful* to the test takers, teachers, and score users, and finally, (7) that *cost-effective data collection and scoring procedures* be considered as part of the validation process. The collection of validity evidence for the TSE test must be an ongoing process that touches upon the areas mentioned earlier. Now we provide some suggestions for validity studies, and discuss some preliminary studies that have already been carried as part of the test development process.

Reliability

The issue of reliability is concerned with distinguishing score variance due to the ability being measured from score variance due to error. Reliability studies will logically include both *intra-rater* and *inter-rater* measures to determine the amount of error due to rater inconsistency. However, it is essential to obtain more detailed information about the sources of variance in the ratings in order to generalize the test taker's performance on the test to performance in other contexts. Thus, reliability studies should be conducted that will investigate the *sources of variance* (e.g. abilities, methods, examinees' personal attributes, and error) to estimate the relative sizes of the different sources of variance ("variance components"). Such studies should be conducted under both trial and operational conditions and employ analysis of variance (ANOVA) procedures to estimate the main effects of the sources of variance and the interactions among them (Bachman, 1990, Chapter 5).

Related to the question of the sources of variance in the test scores is the assumption that maximum information from test tasks is gained from those tasks with difficulty levels that are appropriate to the ability level of the test taker. For this reason, *item response theory* (IRT) procedures have been used to estimate measurement error for items at each ability level. However, the models most often used for IRT analyses (e.g. the Rasch model) are most appropriate for test items, such as multiple-choice items, that can be assumed to be independent and are scored right or wrong. For constructed-response test analysis, newer models have been developed, and should be considered for use in TSE reliability studies (Bennett et al., 1991). To carry out reliability studies in the TSE revision project, test development staff will have to work now with input from those who are going to evaluate the data the test will produce.

A preliminary study of the then current TSE test with a prototype of the revised test was carried out in 1994 (Henning, Schedl, & Suomi 1994). Part of the focus of the study was on comparing the reliability of the prototype with that of the existing TSE test. Medical professionals (*n*=158) and international teaching assistants (*n*=184) were given both versions of the test, which were then rated by 40 trained raters. Interrater reliability for the original and the prototype of the TSE test were both calculated as .899. It was concluded then, that the prototype test was as reliable as the original TSE test.

Validity

TSE validation studies must provide evidence for the appropriateness of the interpretations made from test performance and their uses (Messick, 1989), and are, of course, logical extensions of reliability and generalizability studies. We want to know to what extent TSE scores can be interpreted as evidence of the test taker's ability to use English in professional and social environments. A validation research agenda for the revised TSE test ought to include a number of related components (Bennett et al., 1991):

1. *Qualitative studies* of how good TSE performances differ from less successful ones, what strategies test takers at different levels appear to follow, and what misconceptions lead to problems. This would allow for a better understanding of how test takers at various levels might respond to particular test tasks, which in turn would allow for more meaningful interpretation of scores. This would provide evidence for construct validation.

   A preliminary study of some discourse features of the revised prototype TSE test was conducted in 1994 (Lazaraton & Wagner, 1994). The investigators examined the responses to the prototype test of six native speakers and 12 nonnative speakers. They found that the nonnative speakers performed the same communicative functions as the native speakers did, especially at the higher band levels.

   Another qualitative contribution to the preliminary TSE validation program was a study (Hudson, 1994) of the congruence between the draft theoretical underpinnings (an early draft of the present paper) and the draft test specifications that were developed from the theory. A number of discrepancies between the theory and the specifications were revealed in this study and both the theoretical paper and the specifications were modified and clarified as a result.

2. Investigation of the dimensions of *holistic scoring* of contextualized tasks. It may be that raters making holistic judgments of performance on the proposed contextualized tasks on the revised TSE test will differ in what aspects of the task

they consider in making their judgments, and in the weights they assign to the dimensions. Such variation will affect interpretations that can be made based on test scores, and needs to be understood. This type of study is closely related to reliability, and may well be carried out as part of reliability investigations.

3. Studies investigating the *correlation* of the revised TSE test to the previous versions, to other TOEFL program tests, and to other speaking measures. We need to know how the new TSE test "fits" existing measures of language abilities and what place it holds in the overall assessment of language ability. Such studies may provide evidence of validity by comparing the revised test to currently used criterion measures.

   The Henning, Schedl, Suomi study referred to earlier also looked the correlation of the prototype TSE test and the then-current version with the Oral Proficiency Interview (ACTFL 1986). It was found that where the current test had a correlation of .748 with the OPI, the prototype had a correlation of .819, an encouraging improvement.

4. Studies of the language use of the test takers, and of its evaluation by their interlocutors, in their professional and social environments. Successful test development requires a thorough specification of language use in academic and professional contexts, as well as a method of generalizing performance on a single test to field-specific language use. In order to relate test performance to real life contexts, information is needed about the language needs of the examinees, the procedural work that they carry out by means of language, and how the people they interact with (e.g. patients, colleagues, students, professors) view their language abilities. These types of studies provide content evidence of construct validity.

This research agenda can be seen as important both in the development phases of the revision project, and in providing long-term goals for ongoing research within the TSE program.

Summary of Implications of the Theoretical Underpinnings for Scoring and the Interpretation of Examinee Performance on the Revised TSE Test

The incorporation of current theory in communicative language ability, contextual factors, sociolinguistics, and discourse analysis into the design and scoring rubric of the revised Test of Spoken English allows test users to interpret test results as a measure of communicative ability in spoken English, rather than a measure of linguistic accuracy. This revision in test design and scoring will improve the effectiveness of the revised TSE test in predicting examinee communicative language ability without sacrificing ease of test administration and scoring procedure. Because more mechanical tasks have been eliminated in favor of extended discourse involving more abstract communicative language as well as concrete language, it is expected that examinee scores will be better indicators of their communicative ability.

An ideal speaking test would involve direct interaction of examinees with interlocutors in communicative situations. Operational constraints, however, require the TSE test to be administered via recording equipment with the examinee responding orally to prerecorded prompts. Although the test offers the opportunity to demonstrate speaking ability in situations that require a variety of oral responses, the test is not interactive and therefore cannot test for all aspects of the linguistic, functional, sociolinguistic, textual, and strategic competencies.

The revised TSE test is intended for use in assessing the construct of speaking ability. Although the test designers recognize that creating field-specific tests could provide a more accurate prediction of how examinees might perform in a specific academic or professional domain, it is also true that such test development would be financially and administratively prohibitive. Therefore, the revised TSE test focuses on communicative tasks calling for the application of general information within a variety of contexts. The TSE test alone, with no corroborating assessment measures, should not be used as the sole predictive measure of academic or professional performance. A score on the revised TSE test must be evaluated in conjunction with other types of information about the examinee for decision-making purposes, not as the only indicator of the examinee's ability to successfully communicate in a particular academic or professional environment.

Test items on the revised TSE test are developed in accordance with a strict specification of language function in relation to a particular task, topic, and interlocutor. Item responses are judged on the basis of linguistic accuracy, sociolinguistic appropriateness, coherence and cohesion, functional effectiveness, and strategic success in meeting the requirements of the particular communicative situation represented by the item (or successfully compensating for a lack of available language resources in the accomplishment of the task). The performance of examinees on each item is rated holistically by trained raters, who represent native speaker listeners unfamiliar with nonnative English, according to a set of descriptive criteria that reflect the multifaceted nature of language.

Level descriptors for holistic item scoring are defined in terms of linguistic, textual, functional, sociolinguistic, and strategic competence. Each level of the holistic scale is differentiated by the degree to which examinee performance on the item approximates the performance of a communicatively competent nonnative speaker. The criterion-referenced holistic scoring rubric for the revised TSE test provides five levels of rater judgment with a range of scores from 20 to 60 for each item. The items of each examinee's performance are scored by two raters. Each rater's item scores are averaged and the resulting overall scores are compared. If the averaged scores differ by more than an acceptable level (to be determined empirically), a third rater is called on to determine which of the two original ratings will be discarded. The final score is an average of two ratings rounded to five points.

Results are reported to score users as band scores accompanied by band descriptions of the most common characteristics of examinees that score within each particular band. The revised TSE test does not provide diagnostic information about the examinee's language proficiency in terms of discrete areas of strength and weakness in English. Instead, the band scores provide a holistic evaluation of an examinee's overall communicative language ability.

Successful implementation of the revised TSE test requires piloting prototype tests which make use of the new item types and revised scoring procedures. Successful implementation will also depend on educating the users of the test about what it means to go beyond measuring linguistic competence to measuring an individual's ability to use spoken language to communicate effectively within a particular context. Because the revised TSE test assesses the degree to which examinees achieve success in addressing communicative tasks in addition to maintaining linguistic accuracy, test developers anticipate that changes in the TSE test will encourage English language instructors to emphasize communicative teaching and learning activities in their classes, a positive development.

References

American Council on the Teaching of Foreign Languages. (1986). Oral proficiency interview. Chicago, IL: ACTFL.

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

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L. F., Kunnan, A., Vanniarajan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency tests. *Language Testing 5,* 2: 128-159.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*: 449-465.

Bailey, K. M. (1987). Test of Spoken English. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 84-86). Washington, D.C.: TESOL.

Barnwell, D. (1989). "Naive" native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*: 152-163.

Barrett, R. P. (1987). The SPEAK Test: Some comments by a former user. *NAFSA Newsletter, 38, 7.*

Beebe, L. (1977). The influence of the listener on code-switching. *Language Learning, 27*: 331-340.

Beebe, L. (1980). Sociolinguistic variation and style shifting in second language acquisition. *Language Learning, 30*: 433-448.

Beebe, L., & Giles, H. (1984). Speech accommodation theories: A discussion in terms of second language acquisition. *International Journal of the Sociology of Language, 46*: 5-32.

Beebe, L., & Zuengler, J. (1983). Accommodation Theory: An explanation for style shifting in second language dialects. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition* (pp. 195-213). Rowley, Mass: Newbury House.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*: 145-204.

Bennett, R., Enright, M., & Tatsuoka, K.. (October, 1991). *Developing measurement approaches for constructed-response formats.* Princeton, NJ: Educational Testing Service.

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

Byrd. P. (1987). Being seduced by face validity: Linguistic and administrative issues in videotaped teaching simulation testing. In N. V. N. Chism & S. B. Warner (Eds.), *Institutional responsibilities and responses in the employment and education of teaching assistants: Readings from a national conference* (pp. 355-357). Columbus, Ohio: The Ohio State University Center for Teaching Excellence.

Canale, M. (1983). From communicative competence to language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-28). London: Longman.

Canale, M. (1984). Testing in a communicative approach. In G. A. Jarvis (Ed.), *The challenge for excellence in foreign language education* (pp. 79-92). Middlebury, VT: The Northeast Conference Organization.

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

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30-40). Washington, D.C.: Center for Applied Linguistics.

Carroll, B. J. (1980). *Testing communicative performance.* Oxford: Pergamon Press.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: The MIT Press.

Clapham, C. (1991). *The effect of academic discipline on reading test performance.* Paper presented at the 13th Annual Language Testing Research Colloquium, Princeton, NJ.

Clark, J. L. D., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (Research Report No. 7*)*. Princeton, NJ: Educational Testing Service.

Constantino, M. (1988). *Performance of international students on the SPEAK Test.* Unpublished manuscript. University of Pennsylvania, Philadelphia, PA.

Dickerson, L. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly, 9*: 401-407.

Dickerson, L., & Dickerson, W. (1977). Interlanguage phonology: Current research and future directions. In S. P. Corder & E. Roulet (Eds.), *The notions of simplification, interlanguages, and pidgins and their relations to second language pedagogy.*(pp. 18-29). Geneva: Librairie Droz Neufchatel.

Douglas, D., & Selinker, L. (1985). Principles for language tests within the 'discourse domains' theory of interlanguage. *Language Testing,* 2: 205-226.

Douglas, D., & Selinker, L. (1989). Markedness in discourse domains: Native and non-native teaching assistants. *Papers in Applied Linguistics, 1*, 69-82.

Douglas, D., & Selinker, L. (1991, March). *SPEAK and CHEMSPEAK: Measuring the English speaking ability of international teaching assistants in chemistry.* Paper presented at Language Testing Research Colloquium, Princeton, NJ.

Douglas, D., & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 235-256). Alexandria, VA: TESOL.

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

Ellis, R. (1986). *Understanding second language acquisition.* Oxford: Oxford University Press.

Ellis, R., & Roberts, C. (1987). Two approaches for investigating second language acquisition. In R. Ellis (Ed.), *Second language acquisition in context* (pp. 3-30). London: Prentice-Hall International.

Faerch, K., & Kasper, G. (1983). Plans and strategies in foreign language communication. In K. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20-60). London: Longman.

Gallego, J. C., Goodwin, J., & Turner, J. (1991). ITA oral assessment: The examinee's perspective. In J. D. Nyquist, R. D. Abbott, D. H. Wulff, & J. Sprague (Eds.). *Preparing the professoriate of tomorrow to teach: Selected readings in TA training* (pp. 404-412). Dubuque, Iowa: Kendall/Hunt Publishing Company.

Godfrey, D., & Hoekje, B. (1990). *The use of native speaker norms in evaluating nonnative speaker oral proficiency.* Unpublished manuscript, West Chester University.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics, 3,* Speech acts (pp. 41-58). New York: Academic Press.

Hadden, B. (1991). Teacher and non-teacher perceptions of second-language communication. *Language Learning, 41*: 1-24.

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

Halliday, M. A. K. (1978). *Language as social semiotic.* London: Edward Arnold. (Cited in Ingram & Wylie, 1993).

Henning, G., & Cascallar, E. (1992). *A preliminary study of the nature of communicative competence.* TOEFL Research Report No. 36. Princeton, NJ: Educational Testing Service.

Henning, G., Schedl, M., & Suomi, B. (1994). *Analysis of proposed revisions of the Test of Spoken English.* TOEFL Research Report No. 48. Princeton, NJ: Educational Testing Service.

Hornberger, N. H. (1989). Tramites and Transportes: The acquisition of second language communicative competence for one speech event in Puno, Peru. *Applied Linguistics, 10,* 214-230.

Hudson, T. (1994). *A conceptual validation study of the theory to test specification congruence of the revised test of spoken English.* Unpublished report submitted to the Test of Spoken English Committee, Educational Testing Service, Princeton, NJ.

Hymes, D. (1971). Competence and performance in linguistic theory. In R. Huxley & E. Ingram (Eds.), *Language acquisition: Models and methods* (pp. 3-24). London: Academic Press.

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

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach.* Philadelphia, PA: University of Pennsylvania Press.

Ingram, D. E., & Wylie, E. (1993). Assessing speaking proficiency in the International English Language Testing System. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 220-234). Alexandria, VA: TESOL Publications.

Kosko, B. (1991). *Neural networks and fuzzy systems.* New York: Prentice-Hall.

Kosko, B., & Isaka, S. (1993). Fuzzy logic. *Scientific American, 269.1*: 76-81.

Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia, PA: University of Pennsylvania Press.

Labov, W., & Fanchel, D. (1977). *Therapeutic discourse: Psychotherapy as conversation.* New York: Academic Press.

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research.* London: Longman.

Lazaraton, A., & Wagner, S. (1994). *The revised TSE: Discourse analysis of native speaker and nonnative speaker data.* Unpublished report submitted to Test of Spoken English Committee, Educational Testing Service, Princeton, NJ.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7,* 52-76.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition). New York: Macmillan.

Plakans, B., & Abraham, R. (1990). The testing and evaluation of international teaching assistants. In D. Douglas, (Ed.), *English language testing in U. S. colleges and universities* (pp. 68-81). Washington, DC: National Association for Foreign Student Affairs (NAFSA).

Powers, D. E., & Stansfield, C. W. (1983). *The Test of Spoken English as a measure of communicative ability in the health professions: Validation and standard setting.* TOEFL Research Report No. 13. Princeton, NJ: Educational Testing Service.

Rounds, P. L. (1991). *The contribution of SPEAK Test subscores to understanding of ITA's variable oral proficiency.* Paper presented at the third National Conference on the Training and Employment of Graduate Teaching Assistants.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J., & Roeder, P. M. (1986). Models of second language competence: A structural approach. *Language Testing, 3,* 54-79.

Savignon, S. (1972). *Communicative competence: An experiment in foreign language teaching.* Philadelphia: Center for Curriculum Development.

Savignon, S. (1983). *Communicative competence: Theory and classroom practice.* Reading, Massachusetts: Addison-Wesley Publishing Company.

Schmidt, R. W. (1977). Sociolinguistic variation and language transfer in phonology. In G. Ioup & S. H. Weinberger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system* (pp. 365-377). Rowley, Massachusetts: Newbury House Publishers.

Searle, J. R. (1965). What is a speech act? In M. Black (Ed.), *Philosophy in America* (pp. 221-239). Ithaca: Cornell University Press.

Searle, J. R. (1969). *Speech acts.* Cambridge: Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), Speech acts. *Syntax and semantics, 3,* (pp. 59-82). New York: Academic Press.

Searle, J. R. (1976). The classification of illocutionary acts. *Language in Society, 5*: 1-24.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*: 209-231.

Selinker, L., & Douglas, D. (1985). Wrestling with 'context' in interlanguage theory. *Applied Linguistics, 6*: 190-205.

Smith, J. (1989). Topic and variation in ITA Oral Proficiency: SPEAK and field-specific tests. *English for Specific Purposes, 8*: 155-168.

Smith, J. (1992). *Topic and variation in the oral proficiency of international teaching assistants.* Doctoral dissertation. University of Minnesota.

Spolsky, B. (1989). *Conditions for second language learning.* Oxford: Oxford University Press.

Stansfield, C. (Ed.). (1986). Toward communicative competence testing. *Proceedings of the second TOEFL invitational conference.* Princeton, NJ: Educational Testing Service.

Tarone, E. E. (1979). Interlanguage as chameleon. *Language Learning, 29*: 181-191.

Tarone, E. E. (1982). Systematicity and attention in interlanguage. *Language Learning, 32*: 69-82.

Tarone, E. E. (1983). On the variability of interlanguage systems. *Applied Linguistics 4*: 143-163.

Tarone, E. E. (1985). Variability in interlanguage use: A study of style shifting in morphology and syntax. *Language Learning, 35: 373-404.*

Tarone, E. E. (1988). *Variation in interlanguage.* London: Edward Arnold.

Tarone, E. E. (1989). Of chameleons and monitors. In M. Eisenstein, (Ed.), *The dynamic interlanguage: Empirical studies in second language variation* (pp. 3-15). New York: Plenum Press.

Underhill, N. (1987). *Testing spoken language.* Cambridge: Cambridge University Press.

van Ek, J. A. (1975). *The threshold level for modern language learning in schools.* Essex: Longman.

Weir, C. J. (1990). *Communicative language testing.* Englewood Cliffs, NJ: Prentice Hall.

Wilkins, D. (1976). *Notional syllabuses.* Oxford: Oxford University Press.

Young, R. (1989). Ends and means: Methods for the study of interlanguage variation. In S. M. Gass, C. Madden, D. Preston, & L. Selinker (Eds.), *Variation in second language acquisition: Psycholinguistic issues* (pp. 13-21). Clevedon, Avon, England: Multilingual Matters.

Yule, G., & Tarone, E. E. (1990). Eliciting the performance of strategic competence. In R. C. Scarcella, E. S. Anderson, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 179-194). New York: Newbury House Publishers.

Zuengler, J., & Bent, B. (1991). Relative knowledge of content domain: An influence on native-nonnative conversations. *Applied Linguistics, 12*, 397-415.

Cover Printed on Recycled Paper

58701-14726 • Y46M.75 • 253707 • Printed in U.S.A.