TOEFL®

Monograph Series

MS - 8
APRIL 1997

Testing Speaking Ability in Academic Contexts: Theoretical Considerations

Dan Douglas

Educational Testing Service

Testing Speaking Ability in Academic Contexts: Theoretical Considerations

Dan Douglas
Iowa State University

Educational Testing Service
Princeton, New Jersey
RM-97-1

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

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, 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

In the North American academic context, international students and teaching assistants need to be able to speak proficiently to accomplish a number of tasks/purposes: they need to get around campus, buy books and materials in the bookstore, get meals in cafeterias and restaurants, ask a friend about a homework assignment, get help in locating a book in the library, tell classmates where they are from and how long they've been on campus. They also need to participate in class discussions, ask and respond to questions in classes, give oral reports, get and give help in office hour session, give instructions in labs, conduct tutorials and recitations, give lectures, and so on. In this paper, a theoretical background for the large-scale undergraduate/graduate university admissions, we first discuss the nature of the construct of speaking ability that enables students to carry out the above tasks arguing that speech production and comprehension are systemically integrated, that language knowledge is multicomponential, and that strategic ability is central to the interpretation of context in the test assessment of speaking ability in the context of TOEFL 2000. We then outline research needed before progress can be made.

Acknowledgments

I wish to thank Carol Chapelle, Iowa State University, for her patient discussions of earlier drafts of this paper. She forced me to alter a number of very dearly-held convictions, and I'm afraid the paper is the better for it. I also owe thanks to Carol Taylor and Mary Schedl, Educational Testing Service, for their comments on an earlier draft of the paper. Any remaining flaws are entirely my own responsibility.

Table of Contents

List of Figures

I. Introduction

In the North American academic context, international students and teaching assistants need to be able to speak proficiently to accomplish many tasks and purposes. First, they need basic interpersonal communication skills (BICS) to get around campus, buy books and materials in the bookstore, get meals in cafeterias and restaurants, ask a friend about a homework assignment, get help in locating a book in the library, or tell classmates where they are from and how long they have been on campus. They also need to use cognitive academic language proficiency (CALP) to participate in class discussions, ask and respond to questions in classes, give oral reports, get and give help in office hour sessions, give instructions in labs, conduct tutorials and recitations, and give lectures. The BICS/CALP dichotomy (Cummins, 1980) is useful from an analytical perspective, since it helps us understand the types of choices available to language users. In the academic context, however, learning may take place in either mode. Therefore, it is difficult to say that one or the other should have priority in the assessment of speaking in an academic context. Cummins alone points out that the BICS/CALP distinction is in fact too simplistic and suggests a two-way continuum, with context-embedded language and context-reduced language on one axis and cognitively demanding and cognitively undemanding tasks on the other (Cummins & Swain, 1986). This two-way continuum adds a developmental aspect to the notion of language proficiency, since many language tasks (particularly those of a cognitive/academic nature) require active cognitive involvement at the outset, but as language users become more skillful, the tasks become more automatized and require less active involvement. In new communicative situations, language users must "stretch" their resources to achieve their communicative goals, and it is under these "stretched" conditions that much language testing takes place. The context-embedded/context-reduced dimension refers to the amount of situational and paralinguistic support that language users receive in negotiating meaning interactively. In context-reduced situations, the language user relies almost exclusively on linguistic cues for interpretation, and "may in some cases involve suspending knowledge of the 'real' world in order to interpret (or manipulate) the logic of the communication appropriately" (Cummins & Swain, 1986, pp. 152-153). Context-embedded language is that in which the language user has the opportunity to make use of linguistics, paralinguistic and situational cues in interpretation. These two dimensions are of enormous significance in the interpretation of speaking test performance, and will be discussed in more details later in this paper. However, it is important to remember that any model of speech production and communicative language performance must take the two dimensions of communicative activity into account.

In considering the large-scale testing of speaking ability as part of an overall assessment to inform decision making for undergraduate/graduate university admissions, I will first discuss the nature of the construct of speaking ability by outlining a psycholinguistic model of speech production, including details of communication strategies and communicative language ability. The model will consist of five language-processing components, two knowledge stores, and a strategic component. I will argue that speech production and speech comprehension are systemically integrated, that language knowledge is multicomponential, and that strategic ability is central to the interpretation of context in the test situation.

The second section of the paper will include a discussion of test methods, including context and method facets, discourse and test method, the rating of speaking performances, and consideration of whether to test speaking in isolation from the other language use skills or within an integrated skills assessment model. The main points in this section are that test method facets are the functional equivalents of contextualization cues in natural language use, that method facets should be manipulated to produce

tests for a specific population, that there is a threshold level of method facets necessary to engage test takers in the desired contextual domain, and that context-based tests may be more useful than "general-purpose tests" for making situation-specific judgments about language ability. I will also argue that method facets can be manipulated to engage features of spoken discourse. In this section, I will discuss the rating of speaking performance from the point of view of "fuzzy set theory," arguing that distinctions between score categories are not clearly definable, but that raters can make reliable judgments by being given simple scoring guidelines and "negative feedback" from expert trainers. Finally, I will recommend that the speaking and listening skills be integrated in future tests.

In the final section of the paper, I will summarize the main theoretical points and recommendations for the assessment of speaking ability in the context of TOEFL 2000, and outline research needed before progress can be made on a number of issues.

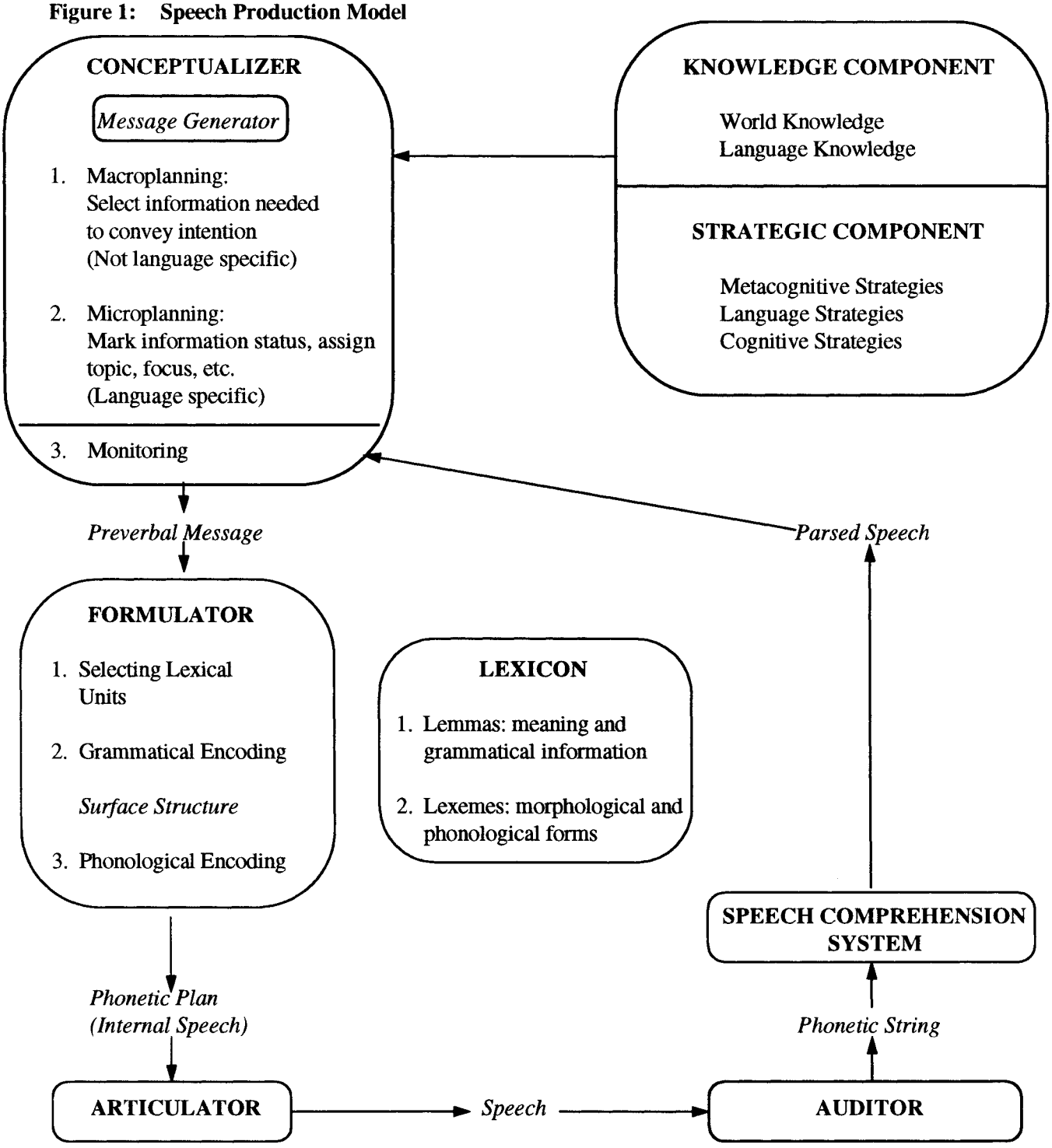
II. Psycholinguistic Model of Speech Production

To begin this discussion of the nature of the construct of speaking ability, I will outline a model of speech production. A language performance, including, of course, one produced in a speaking test, is the result of a complex underlying ability. As Snow and Lohman (1989), in a discussion of measurement models, put it: "...a score reflects a complex combination of processing skills, strategies, and knowledge components, both procedural and declarative and both controlled and automatic, some of which are variant and some of which are invariant across persons, or tasks, or stages of practice..." (p. 268). Applying this view to language testing, and in particular to the testing of speaking, a language production model will consist of some number of language processors, knowledge components, and strategic components.

A. Speech Production Model

Speech production is a most complex process. The speaker must monitor what she has just said to determine whether it matches her intention while she is uttering her current phrase and monitoring that, simultaneously planning her next utterance and fitting that into the overall pattern of what she wants to say, and monitoring as well the reception of her performance on a listener (Brown & Yule, 1983, pp. 4-5). For a general overview of the speech production process, see Fromkin (1990) and Garman (1990, Chapter 7). The model presented here is an amalgam of work by Levelt (1989), De Bot (1992), and Chapelle and Douglas (1993), and borrows heavily from them. The model is essentially that developed by Levelt (1989) as a model of monolingual speech production. However, as De Bot (1992) points out, since every monolingual speaker has the potential to become bilingual, and since bi- and multilingualism are the norm in most parts of the world, the basic model of speech production ought to be a bilingual one. Therefore, De Bot sets out to adapt Levelt's model to describe a bilingual speaker (see also the discussion of second-language psychological processes in Schmidt, 1992). As I have further adapted it to accommodate certain aspects of current understandings of communicative competence, the model consists of five *language processors*: a conceptualizer, a formulator, an articulator, an auditor, and a speech comprehension system; two *knowledge stores*: the knowledge component (containing world knowledge and language knowledge), and the lexicon; and one other processor, the *strategic component.* For empirical justification for the basic model, see the excellent reviews in Levelt (1989) and De Bot (1992) and chapters in Harris (1992). Figure 1 outlines the Speech Production Model as revised for this paper.

Figure 1: Speech Production Model



(Adapted from Levelt, 1989 and De Bot, 1992)

Each component of the Speech Production Model is discussed here:

Knowledge Component: This component is where world knowledge and language knowledge are stored. *World knowledge* includes both knowledge about things (content or declarative knowledge) and about processes (procedural knowledge). *Language knowledge* includes organizational knowledge (grammatical and textual) and pragmatic knowledge (illocutionary and sociolinguistic)(cf. Bachman, 1990). Certain aspects of language knowledge are contained in the second knowledge component, the **Lexicon,** to be discussed in more detail later.

These two subcomponents of knowledge — world knowledge and language knowledge — relate to the context-embedded/context-reduced continuum, since it is these types of knowledge that allow language users to recognize features of situational and linguistic context, particularly those which Gumperz (1976) refers to as "contextualization cues" (to be discussed later).

Strategic Component: This includes metacognitive strategies, language strategies, and cognitive strategies. The *metacognitive strategies* are those that direct cognition and behavior: assessing the context, setting cognitive and behavioral goals, establishing plans for accomplishing the goals, and controlling execution. The *language strategies* work specifically with language: assessing the discourse situation and identifying or constructing appropriate discourse domains, setting communicative goals, constructing a linguistic plan for accomplishing the goals, and controlling linguistic execution. The *cognitive strategies* are fundamental processes that control encoding and retrieval functions.

These three types of strategies relate to the cognitively demanding/undemanding continuum; they use world knowledge and language knowledge in assessing situations and establishing goals and plans for dealing with them.

The Conceptualizer: This is where the communicative goals and intentions and the information from the knowledge and strategic components are ordered and adapted in such a way that they can be converted into language. In the planning of messages, two stages can be distinguished: a *macroplanning stage*, involving the selection of information needed to accomplish the communicative goals and intentions, and a *microplanning stage*, in which information is marked for status (given/new), topic, and focus for realizing the communicative goals and intentions. In bilingual individuals, part of the information selected in the microplanning stage is the decision concerning which language a message will be encoded, since, in conversations between bilinguals, language choice expresses communicative intention and, therefore, carries meaning. The macroplanning, information selection stage is not language specific; the microplanning, information-marking stage is language specific. The output of these two processes is a *preverbal message*, which contains all the necessary conceptual information for converting meaning into language but is not itself linguistic.

The Formulator: This component is where the preverbal message is converted into a *phonetic plan* by selecting the right lexical units and applying grammatical and phonological rules. The lexical units are stored in the *lexicon* and consist of two parts: the lemma and the lexeme. *Lemmas* contain pragmatic and stylistic information, syntactic category and function, number, tense, aspect, mood, and case. They are activated by matching their information with the semantic information in the preverbal message. Activation of the lemmas activates syntactic procedures in the Formulator. While the surface structure is being formulated, morphophonological information from the *lexemes* is being activated and encoded in the phonological encoder. In bilingual individuals, some information about the two languages is stored and processed in a single system, especially where both languages are closely related and share many features or where the person's level of proficiency in one of the languages is not very great. Where the languages differ greatly, and/or proficiency in each is high, elements of knowledge of them would be represented and stored separately. The output of the formulator is a *phonetic plan*, which can be scanned internally by the speaker's *speech comprehension system*, the first opportunity for feedback.

The Articulator: This converts the phonetic plan into actual speech. The articulator receives successive chunks of internal speech from the formulator and executes them by coordinating sets of muscles in the respiratory and vocal tracts.

The Auditor: This is the entry point for the monitoring of speech, sending a phonetic representation to the speech comprehension system, where it is parsed and sent to the conceptualizer for comparison with the original goals and intentions.

The model presented here is compatible, I believe, with the model of language use proposed by the ETS Committee of Examiners (COE) (ETS, 1992). The model here is somewhat more specified than the version of the COE model that I have (contained in a working document), particularly with respect to the "verbal-processing component" in the COE model, which is analogous to the Formulator in the model in this paper. The model presented here assumes, as does the COE model, that external academic context forms part of the input to the production process, as do internal affective factors (such as motivation and anxiety), which are discussed in Section III.

B. The Strategic Component

Since it plays such a central role in the processing of communication, I will discuss the strategic component in more detail, following Chapelle and Douglas (1993). The strategic component in this model is comprised of three types of processes: metacognitive strategies, language strategies, and fundamental cognitive strategies. Chapelle and Douglas describe each of these strategies as a *process* (e.g., goal setting) that generates a mental state (e.g., a goal). Resulting mental states are used to guide other processes. Metacognitive strategies direct the language user's interaction with the context. Language strategies (i.e., Færch & Kasper's [1983] communication strategies) are called on by the metacognitive strategies to take

over direction when the problems in the context are identified as communicative ones (i.e., ones in which language is required to execute the necessary plans). The fundamental cognitive strategies are used by both metacognitive and language strategies to handle basic functions such as encoding information from the context and accessing information from knowledge. This three-tiered system is hierarchically arranged so higher-level processes can call on ones at the lower levels. For example, a metacognitive strategy attempting to assess the contextual cues in a situation would call on the fundamental cognitive process to encode contextual input, and (depending on the nature of the input) might also call on the language strategies to decode it.

Metacognitive Strategies. Consistent with the psychologist's view of metacognition (e.g., Sternberg, 1977), we can identify basic metacognitive strategies which direct cognition and behavior. These general metacognitive strategies are directly responsible for performance in situations which do not require language (such as packing a lot of suitcases in a small car or completing items on an embedded-figures test) (Chapelle & Green, 1992). Although the metacognitive strategies are not directly responsible for communicative language use, they do have access to language knowledge which they can use in non-communicative settings (such as doing crossword puzzles or foreign-language drills). From the language tester's point of view, test situations do exist where language knowledge may be called upon but that are not *communicative* situations. Whenever language is involved in a problem situation, the metacognitive strategies must call upon the language strategies, which return a result to the metacognitive level. Four types of strategies operate at the metacognitive level:

1. Assessing the context. From the perspective of language production, this metacognitive strategy is very important since it provides the language user with an interpretation of the context, the first "orientation" to the situation. This assessment allows the subject to create a "mental model" of the situation by perceiving contextual cues and interpreting them with reference to memory of similar situations he has experienced. This process results in the subject's "mental model": his understanding of the situation and his role in it. This mental model then guides subsequent processes, including goal setting.

2. Setting goals for cognition and behavior. The mental model is used by the goal-setting process, which determines the goal — what the person's objective will be.

3. Cognitive planning. That goal is the input for the cognitive-planning procedure, which constructs a plan for accomplishing the goal.

4. Control of execution and attention. The subject must then control the execution of the plan.

Language Strategies. The previous discussion is concerned with noncommunicative cognitive activities. Chapelle and Douglas (1993) hypothesize a separate set of strategies to work specifically with communication: setting communicative goals and carrying out communicative plans (cf. Færch & Kasper,

1983). This is clearly an important set of strategies for language testers to be aware of. Details of language strategic processes are given here:

1. Assessing the discourse context. When the metacognitive processor identifies a situation as communicative, it directs the language strategies to assess the communication situation for "cues" to identify the appropriate "discourse domain" (Selinker & Douglas, 1985; Douglas, 1992): the language user's assessment of the communicative context/situation and her role in it. The subject's discourse domain, which may or may not be in concert with what the testers intended, affects subsequent strategies.

2. Setting communicative goals. The discourse domain is used by the goal-setting process to determine what the person's communicative objective will be. These goals are input to the conceptualizer, which is responsible for selecting and ordering the information needed to carry out the goal.

3. Linguistic planning. The strategic component also directs the macro- and microplanning processes in the conceptualizer.

4. Control of linguistic execution. Finally, language execution strategies direct the marshalling of language knowledge, including that from the lexicon, in formulating the surface structure and the phonetic plan.

There has not been much research on the use of cognitive language strategies by second language users. What there is focuses on the relationship between intended communicative function (e.g, narration, description, instruction) and linguistic form (see, for example, Yule & Tarone, 1990). Cohen and Olshtain (1993) employed "think aloud" techniques to investigate the strategic processes used to produce speech acts and found three strategic styles in a group of fifteen advanced learners producing apologies, complaints, and requests. One type, called "metacognizers" by Cohen and Olshtain, seemed to be highly aware of their metacognitive processes, planning and monitoring their utterances. The second type, "avoiders," systemically avoided problem areas (such as uncertainty about whether to call a teacher by name). The third type, "pragmatists," were characterized by trying out alternative methods of achieving the communicative goal. This study highlights the fact, as Cohen and Olshtain point out, that "not all speaking tasks are created equal — that there are tasks which make far greater demands on learners than do others" (p. 50). Language testers need to be aware of this fact and should consider the language-processing demands that test tasks are likely to make.

To further confuse the issue of strategic processing, however, Cohen (1993), in a later study, uncovered a second language-processing phenomenon called "reprocessing" (Lambert & Tucker, 1972) among immersion learners. On the surface, it seemed that schoolchildren studying mathematics in a second language were comprehending math problems, processing information and carrying out math operations, and producing answers in the second language. Think-aloud protocols, however, suggested strongly that the children were *reprocessing* in their first language much of the information received through the second. In other words, rather than processing the math problems in the second language, the subjects were either

performing on-line translations into their first language or operating in the second language until encountering perceptual difficulties, and then switching to first-language processing. The upshot of this finding is that surface speech comprehension and production may or may not reflect underlying second-language processing. More research is needed in this area to help testing researchers understand more completely the nature of the underlying construct of speech production. This is highly relevant to validity concerns, as well as the interpretations test users make of learner performance.

Another aspect of language strategic ability is relevant to academic language use. Bardovi-Harlig and Hartford (1990), for example, found that academic advising sessions are marked by an incongruence in the status of the two interlocutors (adviser and student). They found that both native and nonnative students were able to negotiate status-challenging speech acts (e.g., disagreeing) in such situations but that nonnatives lacked status-preserving strategies (i.e., preserving adviser-ego) and, were thus less successful at challenging than the native speakers.

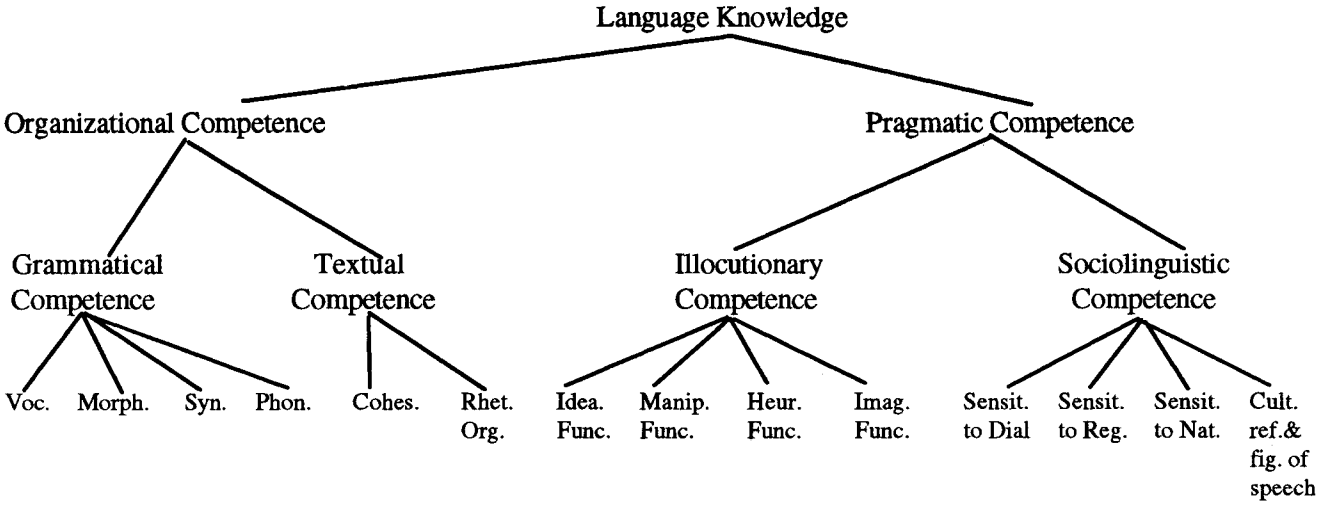
C. Language Knowledge

I have discussed in detail the nature of one aspect of test-taker attributes, the strategic component, and will now take up a detailed discussion of a second crucial attribute, language knowledge. As empirical studies conducted by language testers and others in recent years has made clear, language proficiency is multicomponential (see, for example, Canale and Swain, 1980; Canale, 1983; Savignon, 1983; Duran et al., 1985; Sang et al., 1986; Bachman, 1990 [Chapter Four]; Olaofe, 1992). Precisely what those components may be and how they interact in actual language use, however, is extremely unclear. (See Henning and Cascallar, 1992, for a preliminary study of the interaction of components of communicative competence.)

The term "communicative competence" has been invoked for nearly three decades now to encompass 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. For Hymes, "competence" is more than knowledge: "Competence is dependent upon both [tacit] *knowledge* and [ability for] *use*" (Hymes, 1972, p. 282; brackets and emphasis in original).

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), which postulates three components of "communicative language ability": language competence, strategic competence, and psychophysiological mechanisms. These are roughly analogous to language knowledge, the strategic component, and the articulator in the model being proposed here. In this section, we are primarily interested in the component of language knowledge. Figure 2 illustrates the components of language knowledge.

Figure 2: Components of Language Competence



(Based on Bachman, 1990, p. 87)

In the model presented here, language knowledge consists of *grammatical competence* (knowledge of vocabulary, morphology, syntax, and phonology), *textual competence* (knowledge of how to structure and organize language into larger units: rhetorical organization; and how to mark such organization: cohesion), *illocutionary competence* (knowledge of the functions of language: ideational, manipulative, heuristic, and imaginative), and *sociolinguistic competence* (sensitivity to dialects, registers, naturalness, and cultural references and figures of speech).

Certain language information is contained in the *Lexicon*. Aspects of language knowledge such as lexical meaning, syntactic category, and syntactic function are part of the *lemma* of each lexical unit. Knowledge about the morphology and phonology of each lexical unit is part of the information contained in the *lexeme* of the lexical unit. Metacognitive and language strategies are employed to direct the use of these types of language knowledge in the planning and execution of speech production.

D. Overview and Summary of the Model

In the model, production takes place from left to right, and the next processor will start working on the output from the current processor even if this output is incomplete. Production is also incremental; as soon as the information that goes with one part of the utterance is passed on to the formulator, the conceptualizer does not wait for that chunk to go through the whole system but immediately starts on the next part. Thus, various parts of the same utterance will be at different processing stages: when the first part is being produced by the articulator, the last part may not have left the conceptualizer. Consequently, the different components are at work simultaneously, or in parallel processing.

This process works rapidly. At an average rate of 150 words a minute, a bilingual speaker with a total vocabulary of, say, 60,000 words in his two languages, must choose the correct item two to five times a second (i.e., one item every 200 to 400 milliseconds)(Garman, 1990; De Bot, 1992). Lexical retrieval is seen as "passive"; that is, no actual active search is carried out. Rather, as meaning characteristics from the preverbal message are "read" by the Lexicon, lexical candidates automatically present themselves until enough characteristics are processed to eliminate all except one item.

The model of speech production outlined so far has the following characteristics:

1. The model distinguishes between knowledge components and processors. There are two types of knowledge: world knowledge and language knowledge. These provide information to the processors for interpreting context, setting goals, establishing plans, and executing the plans.

2. A strategic processing component is responsible for interpreting context, setting goals, and directing the establishment of plans and executing them.

3. When a communicative goal/intention is established, a processing component, the conceptualizer, selects information from the knowledge component necessary for meeting the goal and organizes it into a preverbal message.

4. The next processor, the formulator, employs information from the lexicon to select lexical units and encode them grammatically and phonologically to produce a phonetic plan. This plan, in the form of internal speech, may be conveyed by the speech comprehension system for self-monitoring by the conceptualizer.

5. This phonetic plan is realized in the form of speech by the articulator. This overt speech may be processed by the auditor and sent to the speech comprehension system for monitoring.

6. The model is intended to serve for the description of the speech production of both monolinguals and bilinguals, although the latter is the unmarked case. Some of the bilingual's language knowledge is stored together and some separately, depending on both the "distance" between the languages and the level of proficiency of the bilingual.

E. Implications of the Model for Testing Speaking Ability

1. The Strategic Component and the Interpretation of Context

Testees, when confronted with a language-testing situation, interpret context in much the same way as in any other situation they may find themselves in. The difference between a language test as situation and a natural language situation is that a language test is contrived, and all situational cues must come from its

method facets" (discussed in the next section of this paper). Therefore, it is incumbent upon the test developers to provide, in the form of method facets, sufficient cues to engage the testees in the intended discourse domain. More will be said about how this may be done in the next section, but it is clear that testees *do* interpret context, that they use these interpretations to set communicative goals, and that these goals do influence their subsequent language performance. In testing situations where the context is insufficiently specified, testees will create their own interpretations of the situation on the basis of previous experience with tests, advice from friends, and so on, and their performance will be impossible to interpret. Douglas and Selinker (1985) argued that, when testees approach a test, three possibilities exist concerning the interpretation of the context: (1) they will engage a discourse domain that already exists in their knowledge base if they recognize a sufficient number of cues in the test context; (2) they will create a temporary domain to deal with a novel situation, based on whatever world and language knowledge they can bring to bear; or (3) they will flounder, attempting to make sense of a context that provides insufficient information for interpretation.

It seems to me that there are three degrees of specification of context in current testing practice. At one end of a continuum are tests which are highly specified contextually, known as "specific-purpose" or "field-specific" tests; at the other end are tests, called "general" tests, which are underspecified in situational or paralinguistic context, and sometimes even in linguistic context. In the middle of the continuum are tests (although not many of these exist) which are situationally and linguistically contextualized but broader in scope than more field-specific tests. I will refer to such tests as "pan-technical" tests and suggest that it is this level of contextualization that is most appropriate for large-scale language tests interpretable for decision making in academic admissions. The term "pan-technical" was coined by Josh Ard (personal communication, January, 1983) to better describe a level of vocabulary in technical language use, known more widely as "subtechnical" vocabulary: that which characterizes technical language situations, yet is not field specific. It includes such lexical items as *bulb, correlation, filter, table, tube,* and *verify.* I want to suggest that the pan-technical notion be extended to include that of "academic language": language ability required for coping with communication in the academic milieu. Some research has been conducted on the nature of language proficiency in an academic context (see, for example, Adamson, 1993, for a summary; see also Bridgeman & Carlson, 1983, and Powers, 1985), but much more is needed before we can understand the nature of academic language use.

When considering the implications for language testing of the strategic component of the speech production model, a related point needs to be emphasized. This point is that it is possible (and is often the case) that a test can elicit a language performance that does not reflect changes in topic or content. Thus, researchers have frequently been baffled by a failure to find differences in performances between "general" language tests and "field-specific" language tests (cf. Clapham, 1993; McNamara, 1990; Read, 1990). This occurs when the application of metacognitive strategies in the assessment of the context does not result in the recognition that a situation is particularly field specific and, thus, does not engage an existing discourse domain in that field. In such a case, the testee will either employ an existing domain for "taking a language test" which is no different than that employed for any other language test or will create a temporary domain to deal with the situation that reflects the intended field-specific area. An example is discussed in Chapelle & Douglas (1993), in which they consider the implications of applying a cognitive/interactive model to the interpretation of test performance data. They cite a study by Smith

(1992) in which the researcher compared performance on a "general" test of speaking proficiency (SPEAK) and revised tests of the same format but with field-specific topical content (mathematics, physics, and chemistry). In other words, Smith manipulated one aspect of context — topic — but held constant all others. Chapelle and Douglas argue that, overall, Smith did not test field-specific language ability because her revised tests did not provide enough contextualization cues to engage a domain that was much different from that engaged in taking the "general" speaking test. This is not a criticism of Smith's study, since she was specifically investigating the effect on performance of changing only the *topic* of test items, not others (such as rubric or environment). Her study does help us better understand the nature of context and its interpretation as they relate to the testing of speaking ability.

The final issue I want to mention regarding the strategic component of the speech production model is that of the necessity of taking account of cognitively demanding versus cognitively undemanding tasks in test construction. The cognitively demanding/undemanding continuum (Cummins & Swain, 1986) is related to the degree to which a task requires active cognitive involvement (as many academic tasks do) as opposed to more automatized activities (such as basic social routines). Testers need to take account of the degree of cognitive involvement required in the test tasks in interpreting performance results. Tasks that are new to the testee will be more cognitively demanding than those that are more familiar and will elicit different types of performance.

2. The Implications of the Knowledge Component for Testing Speaking

The Speech Production Model presented here distinguishes between world knowledge and language knowledge. The distinction between them has been problematical for language testers, since there is the difficulty of distinguishing between them in interpreting test results. Many researchers have studied this problem. For example, Alderson and Urquhart (1985), Clapham (1991), Douglas and Selinker (1991 and 1993), Hale (1988), Smith (1992), Tedick (1990), and Zuengler and Bent (1991) all have found significant interactions between world (or background) knowledge and language test performance, particularly when test content and tasks were sufficiently specified and when subjects' level of language proficiency was sufficiently high to be able to use the situational information. Thus, it appears that under some conditions, world knowledge makes a difference to language test performance. On the other hand, some studies suggest that, under other conditions, such nonlinguistic knowledge does not influence language test performance to any significant degree (e.g., Clapham, 1993; McNamara, 1990; Read, 1990). The problem of language testing is to understand what the conditions are that influence test performance. Until such features are understood and controlled, valid interpretation of language test results will be impossible.

From another perspective, however, it may be that the world knowledge/language knowledge distinction does not reflect reality — that, in natural language use, the distinction does not hold up. From this perspective, part of the problem for test performance interpretation is that, while tests may not require testees to engage much world knowledge at all (depending rather on context-reduced, linguistic cues to elicit performance), test users wish to make interpretations of scores in terms of context-embedded situations (such as academic performance or teaching). This brings us to the second dimension of language proficiency, the context-embedded/context-reduced continuum. If there is to be a congruence between the

elicitation of language performances and the interpretation of these performances, it seems to me that there needs to be a congruence between the types of knowledge the test requires and the types of knowledge demanded by the situation for which the test results are to be used. In other words, it may be a mistake to use the results of a context-reduced test to make interpretations about a context-embedded situation such as academic study. Language test developers need to be aware of this aspect of the relationship between world knowledge and language knowledge.

Another aspect of the knowledge component that has implications for test development is the division of language knowledge into four subcomponents: grammatical knowledge, textual knowledge, illocutionary knowledge, and sociolinguistic knowledge. The question is whether a speaking test, or a test of any of the language skills, should test each of the four areas. That is, should there be test sections, items, or tasks that require the testee to display ability in each knowledge subcomponent, or should such differentiation be a part of the scoring/rating procedure (on the assumption that all four areas are "in play" all the time in communication)? Before an answer can be given to this question, much more research is needed on the nature of language knowledge and the interaction among its components.

3. Implications of Monitoring for Speaking Tests

The Speech Production Model presented here suggests that both "internal speech" and articulated speech are subject to monitoring, using the same processing components that are used for comprehending the speech of interlocutors in interactive communication. This, in turn, suggests that (1) it would be good if a test could provide for assessing self-correction and message adjustment, and (2) the speaking and listening skills might usefully and naturally be tested together. This issue will be discussed in detail later.

III. Test Methods

In this paper, I have discussed the attributes of the test taker. In a language test, these attributes interact with aspects of the test method to produce a test performance interpretable as evidence of language ability for a particular purpose or purposes. This section of the paper will deal with test method from the point of view of context and how it is realized in test method facets, discourse as it relates to method, the rating of speaking performances, and the question of whether to test speaking in isolation or integrated with other skills. For a general overview of methods of testing spoken language, see Plakans and Abraham (1990), Underhill (1987), Walker (1990), and Weir (1990).

A. Context and Test Method

For the past fifteen years or so, a recurring theme in applied linguistics has been the role of context both in the acquisition of the language code itself and on the development of communicative ability (see Ellis and Roberts, 1987, and Larsen-Freeman and Long, 1991, for discussion). "Context" has been defined in various ways by researchers and has been associated with such notions as "situation," "setting," "domain," "environment," "task," "content," "topic," "schema," "frame," "script," and "background knowledge." Kramsch (1993) describes five dimensions of context: linguistic, situational, interactional, cultural, and intertextual. She reminds us that in each of its dimensions, context "is shaped by people in dialogue with one another in a variety of roles and statuses" (p. 67). Contexts, viewed in this way, are not stable; interlocutors are constantly changing and reconstructing them. For this discussion, I will collapse the distinctions among the five dimensions of context because these distinctions are irrelevant to the main thesis, which is that context is a social/psychological construct that results from attention to a variety of external and internal features.

Context does play a role in influencing language performance; the problem is arriving at a common understanding of the nature of context and what constitutes it, then determining specifically how various contextual features influence language use in tests. In considering the various aspects of context, scholars have listed many features, such as those in Hymes' SPEAKING mnemonic:

> **S**ituation: the meaning the participants attribute to the physical and temporal setting, psychological and cultural scene
> **P**articipants: the role(s) they think they should take in the interaction
> **E**nds: outcomes and goals they attribute to the exchange
> **A**ct sequence: message form and content they think they should attend to
> **K**ey: the tone and manner they think appropriate
> **I**nstrumentalities: channels, codes, and registers they think are appropriate
> **N**orms: norms of interaction and interpretation they think are called for
> **G**enres: categories of speech events they think they are engaged in
>
> (Hymes, 1974)

Such situational factors as those set out by Hymes derive from a number of different sources: societal and community values, social situations, role relationships, personal interactions, and linguistic resources. I have listed them here as a reminder that the notion of context is grounded in the complex interaction of

physical, social, and psychological factors. In a paper that presents a complex challenge to applied linguistics researchers, Hornberger (1989) describes an extended speech event in which she was a participant, involving her attempts to obtain a duplicate driving license, using her English/Spanish interlanguage. She employs Hymes' framework in her analysis and refers to many situational variables that might seem irrelevant to language acquisition and use: the altitude of the Peruvian city she was in, the location of offices (e.g., on the roof of a building), an electrical blackout, the closing of a bank, a driver's license obtained eight years previously, the theft of the license at a religious festival, ongoing transportation strikes, and the presence of terrorists. Hornberger argues, however, that the point of her analysis is that "the very essence of a communicative event [is] that it is situated in a real, physical, cultural, historical, and socio-economic context" (Hornberger, 1989, p. 228), and that these all featured in the advancement of the acquisition of her own communicative competence in Spanish. To paraphrase Kramsch (1993), the challenge to language testers is how to shape and control the context of the test as both an individual speech event and as a social encounter "with regard to its setting, its participant roles, the purpose of its activities, its topics of conversation, its tone, modalities, norms of interaction, and the genre of its tasks" (p. 67).

The influence of affect on test performance should be mentioned here. Affective factors (such as motivation and anxiety) in test takers do influence their performance on tests. Exactly how is not so clear. I will not review here the research on these two factors but will point out that context and test methods are important in the increase of motivation and the reduction of anxiety. Test developers need to keep both these attributes in mind as they manipulate test environment, rubric, and input facets. As Bachman puts it, "...I refuse to accept the criticism that taking tests is necessarily an unpleasant experience. Humanizing the facets of the testing environment is perhaps the single most effective means for reducing test anxiety" (1990, p. 318).

In the study of context and its effect on second-language performance, then, I would like to take account of those features of external context that participants attend to in constructing an interpretation of context (with the proviso that we may never be able to know with any precision what those features are because such assessment is dynamic, internal, highly personal, and to some degree unconscious). Gumperz (1976) has referred to "contextualization cues" — changes in voice tone, pitch, tempo, rhythm, code, topic, style, posture, gaze, and facial expression — culturally conventional, highly redundant signals that people who are interacting attend to in their mutual construction of context. For example, a study of contextualization cues used, and not used, by an elementary school teacher (Dorr-Bremme, 1990) suggested that the pupils attended to the teacher's use of explicit formulations, paralinguistic shifts in speech rate and volume, and framing words. When the teacher employed such cues (largely unconsciously), the transition from context to context occurred smoothly. When the cues were absent, however, the context remained unestablished for some time, and the interactions were marked by uncertainty, confusion, and chaos on the part of the children and the teacher. Researchers can analyze communicative events of groups with respect to certain contextualization cues to see what effect the presence or absence of the cues seems to have on the development of the discourse. Because of individual variation in attention and schematic expectations and of redundancy in the cuing system, however, we can never be sure that individuals are all attending to the same cues or that the absence of one type of cue is the factor that leads to particular observed twists and turns of discourse.

This issue of redundancy in contextualization cues is considered in more detail here. Erickson and Shultz (1981) suggest that redundancy makes it more likely that people who are interacting will "get the message" that something new is happening, in spite of "differences in interactional competence, whether due to difference in culture, to personality, or to level of acquisition of competence...and despite differences in individual variation in focus of attention at any given moment..." (p. 150). The redundancy in contextual cues means it will be difficult to determine exactly what features of a context are critical. There is probably a "threshold," a cumulative effect of cues, that determines for an individual what is going on, and this will vary both from individual to individual and from time to time. Recent research with field-specific language tests (Clapham, 1992; Smith, 1992; Douglas & Selinker, 1991, 1993) suggests that field-specific language ability will be engaged only when there is sufficient contextualization (i.e., when the test is sufficiently specified for context). The definition of "sufficient" is a matter for future empirical research.

To make further progress in understanding the nature of communicative ability in language tests, then, language test developers need to operationalize the notion of context in their tests. As a way of doing so, we can follow Bachman (1990), where he presents conclusions on the method effect in testing theory:

> While it is generally recognized that [specification of the task or test domain] 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. 244)

I would argue (cf. Douglas and Selinker, 1991) that any factor one changes in the test environment — personnel, physical conditions, time, organization, instructions, level of precision, propositional content, etc.— can lead to changes in testees' perceptions and assessment of the communicative situation and, thus, to changes in interlanguage performance on a test. Douglas and Selinker (1991) have related the concept of discourse domains in second language acquisition (SLA) to that of contextualization cues in ethnomethodology (Gumperz, 1976) and thus to the concept of method facets in language testing (Bachman, 1990). From this perspective, method facets in language tests are the functional equivalent to contextualization cues in natural language use. By manipulating these facets, the tester can control and shape the context of the test.

Bachman (1990) outlines in detail the facets of test method, which include the following:

environment

equipment, personnel, time, and location

rubric

organization and salience of parts; time allocated to each part; and language, explicitness, and specification of instructions, including domain-specific reasons for carrying out the test tasks

input

format of presentation; subtechnical as well as technical language; context-embedded/reduced discourse; level of precision; topic; genre; grammatical, cohesive, and rhetorical organizational features; and pragmatic and sociolinguistic domain-related features

expected response

format of response, nature of language of response, and restrictions on response

relationship between input and response

reciprocal, nonreciprocal, adaptive

(cf. Bachman, 1990, p. 119)

With regard to the focus on test method as a distinguishing factor in performance, we should consider what Bachman (1990) calls "the dilemma of language testing" (p. 287): language is both the object and the instrument of measurement. This means that it may be difficult to distinguish between the test method and the ability we wish to measure. For Bachman, a way out of the dilemma is to understand more explicitly both the nature of language ability and the nature of test methods, so that we can minimize the effect of test method in the interpretation of results as indicators of language abilities. The perspective in this paper has a different focus. I argue that, instead of attempting to *minimize* the method effects, we need to *capitalize* on them by designing a test for a specific population: international applicants for admission to North American colleges and universities. This will involve constructing a test that contains instructions, content, genre, and language directed toward that population. The goal is to produce a test that will provide information interpretable as evidence of spoken communicative language ability in an academic context.

In summary, I offer the following suggestions concerning testing methods as guidelines for context-based language testing:

1. Any factor testers change in the test domain can lead to changes in a testee's perceptions and assessment of the communicative situation and, thus, to changes in interlanguage performance on the test.

2. Rather than attempting to minimize the effects of test method facets on the interpretation of results, testers should use them to design tests for particular populations.

3. Context-based language tests can be constructed by manipulating many test method facets, including test environment, rubric, input, response, and the relationship between the input and response.

4. There is likely to be a threshold level of context-based method facets, or "situational authenticity" (Bachman, 1991), necessary for domain engagement.

5. Context-based tests may provide more useful information than general-purpose tests when the goal is to make situation-specific judgments about subjects' communicative language ability.

B. Discourse and Test Method

As the discussion of the Speech Production Model in Section II makes clear, consideration of speaking ability assumes the presence of an interlocutor, a co-participant in the speech event (Hall, 1993). The speech that language users produce has pragmatic consequences for hearers (Searle, 1969; Austin, 1962). While speech may be produced when no listener is present (such as when talking to a telephone answering machine, or practicing an after-dinner speech in front of a mirror), an audience or interlocutor is nearly always assumed. The discourse features of spoken language differ from those of written language primarily because spoken language is transitory, immediate (in the sense of on-line), and modifiable in response to listener feedback. (For an overview of the features of conversational discourse, see Ervin-Tripp, 1993.) Features of these differences include the following, according to Brown and Yule, 1983:

> Syntax of spoken language is less structured than that of written: more incomplete sentences, relatively little subordination, passive, it-clefting, wh-clefting.
>
> Clausal relations in spoken language tend to be marked by simple connectors such as *and, but,* and *then.*
>
> Spoken language is typically less explicit than is written: *I'm so tired* [because] *I walked all the way home.*
>
> Spoken discourse tends to be structured in short chunks of speech, each chunk rarely containing more than two premodifying adjectives or more than one predicate per referent: *Old man Ross + he's a little guy + very short + and uh a beard + and he's sort of bald.*
>
> Spoken sentences tend to be structured as topic-comment sequences: *The dog + did you let him out?*
>
> Spoken vocabulary tends to be somewhat generalized: *a lot of, got, do, thing, nice, stuff, place, things like that.*
>
> Spoken syntactic forms tend to get repeated: *I ate breakfast + I ate lunch + I ate dinner + I went to bed.*

> Spoken discourse is marked by a large number of "fillers": *well, uhm, I think, you know, if you know what I mean, of course, and so on.*
>
> (Brown & Yule, 1983, pp. 15-17)

Biber (1992) conducted a theory-based factor analysis of discourse complexity in spoken and written registers. The results of his study characterize a spoken academic register (that of prepared speeches) on five dimensions: infrequent use of reduced forms, frequent use of referential elaboration, frequent use of integrated (informationally dense) structures, infrequent use of framing elaboration (the expression of personal attitudes, justifications, and feelings), and moderate use of passives. More statistical research of this type, as well as more qualitative types, is needed to understand more fully the nature of what it means to speak proficiently in an academic context.

Spoken discourse is also characterized by what Goffman (1976) considered to be universal system constraints on communication. In other words, all languages share certain discourse features, following Hatch, 1992:

> *Channel open and close signals*: language/culture-specific ways of signaling that communication is about to begin and about to end.
>
> *Backchannel signals*: ways of signaling that the message is being followed, such as nodding, smiling, *uhmhum, uhhuh, yeah, right,* or not followed, such as *huh? what?*
>
> *Turnover signals*: signals that allow for smooth exchange of turns: slowing tempo, vowel elongation, falling intonation.
>
> *Acoustically adequate and interpretable messages*: There are ways of adjusting messages to accommodate noisy conditions, dialect differences, differences in level of proficiency, differences in register.
>
> *Bracket signals*: There must be ways to show that parts of a message are not part of the main message: *As the main topic of today's lecture* [looks directly at audience] *Is this microphone on? OK. Now, the main topic...*
>
> *Nonparticipant signals*: There must be ways for nonparticipants to signal that they wish to participate in a speech event, such as repeating part of the ongoing talk: *California? I'm from California...*, and for nonparticipants to remain so, such as avoiding eye contact.
>
> *Preempt signals*: There must be accepted ways for speakers to interrupt each other in emergencies or when repair or clarification is needed.
>
> *Gricean norms*: Grice's (1975) maxims of conversational cooperation must be observed: quantity, quality, relation, manner.
>
> (based on Hatch, 1992, pp. 8-35)

In addition to these system constraints on discourse, Goffman (1976) postulates a set of "ritual" or social constraints that may be characterized by the general feature of appropriacy. For example, when a speaker engages in a bid to open a conversation, she expects that her bid will be well received and reciprocated; speakers are expected to allow others a fair length of time for a turn, and to receive a fair turn; a speaker attempting to join a conversation expects to be allowed to do so; a preempting signal is expected to be judged as reasonable and not rude; when a speaker is unclear, she expects her interlocutor to signal that there is a problem. These ritual constraints are intimately tied to a speaker's image of herself and of others, and reflect these images. Speakers want to present themselves as modest, competent, socialized beings and to attribute the same features to their interlocutors. Language/culture-specific ways of signaling these features are very much a part of what it means to know a language and to be judged a competent speaker of it. The research of Tyler (1992), Tyler and Bro (1992), and Williams (1992) strongly suggests that the nonnative discourse structure of international teaching assistants in academic situations can lead to perceptions of noncomprehension among their audience. Discourse ability is, therefore, an important aspect of what it means to know a language.

With regard to discourse and test method, we need to consider how test methods constrain spoken discourse. A study of the discourse of answering machine talk (Gold, 1991) is worth looking at, since it was carried out specifically to investigate the following relevant question: "What do people do with their talk when there is no interlocutor present to give them feedback and negotiate meaning and yet they are encouraged to speak as if there were?" (p. 245). Gold analyzed the recorded messages received on her answering machine from a small population, during a short period of time. However, she did uncover some interesting features of answering machine talk (AMT). She calls attention to an underlying *irony* in the task of leaving an answering machine message: callers are asked to speak as if they are talking to a someone, or at the very least a something, when in fact they know they are not. This irony is reflected in callers' use of questions not framed as if they are to be answered when the answering machine owner returns home, in asking and then answering their own questions, in the use of humor to mask the absurdity of the task, and in the use of suprasegmental features such as voice pitch and timbre, rhythm, pauses, and intonation to create a dialogic effect (e.g., leaving "space" for a reply even though none is expected). Gold also finds features of written discourse in the AMT: the use of the machine owner's name in the salutation (*Hi, Ruby* instead of simply *hello* or *hi* as would be the case in face-to-face or normal telephone talk); stating the date, time, and/or location of the call as in a letter; and stating the name of the caller (although all these callers were well known to the author by voice). Gold concludes that AMT is a unique form of discourse in which callers must "call upon all of their repertoire, in the form of dialogic, monologic, written and school-based conventions in order to cope with the irony of the task and create new conventions for use in AMT" (p. 254).

The analogue to answering machine talk is performance on the current semidirect *Test of Spoken English* (TSE®), and the comparison highlights the crucial influence of test method on language production. In implementing new test delivery methods, developers must take into account their effects on the nature of the discourse of responses. The challenge is to consider how test methods can be manipulated to engage features of natural spoken discourse, or at least to control for those produced by the method facets. For example, instructions should be designed to give test takers plausible reasons for carrying out the required speaking tasks; the input format should be dialogic to the degree possible; the discourse in the input and

expected response should be context embedded (i.e., contain linguistic, paralinguistic, and situational cues); and a spoken genre should be employed, containing grammatical, cohesive, and rhetorical features of spoken discourse, as discussed previously. The relationship between the input and the expected response should be reciprocal; that is, the speaker's message should have the capability of reducing uncertainty in the listener, which in turn will allow the listener to fashion a message in response that reflects the change in information. Thus, there should be provision for the test taker to formulate a communicative goal or intention and to see that goal affect the interlocutor.

Finally, another study, Hartford and Bardovi-Harlig (1992), suggests that institutional discourse itself, in the form of academic advising sessions, is a unique form of discourse, different from natural conversation. This suggests that, not only do test method facets influence the nature of the discourse of the test, but also that context of situation interacts with method to create unique forms of discourse. Both context and method must be carefully considered in test production and the interpretation of performance as indicative of academic speaking ability. Other studies of discourse in an academic setting have focused on international teaching assistants (ITAs). For excellent overviews of the problems of assessing and training ITAs, see Plakans and Abraham (1990) and Hoekje and Williams (1992).

C. Rating

The spoken performances of the test takers must be rated in some way. It is almost axiomatic that, because language use is a multicomponential phenomenon, requiring interlocutors to negotiate meanings, no two listeners hear the same message. This aspect of language use is a source of bias in test scores. It leads language test developers to severely limit which features of a performance they require raters to attend to in making their ratings. They hope that, if raters focusing attention only on pronunciation, grammar, fluency, and comprehensibility, for example, the many other features of the discourse will not influence them. There is mounting evidence that this is a vain hope.

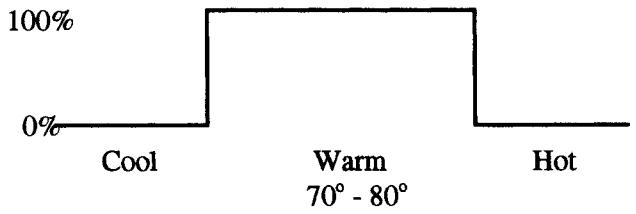
For a start, scores that adhere strictly to those aspects of the communication that are desired could probably be obtained only by so severely restricting the allowable aspects of language that raters take into account in assigning ratings that the naturalness of the language use process will be seriously distorted. This would require a retreat from integrative, communicative tests and a return to discrete point measures. This has been done before, with little apparent success. A measure of "grammar knowledge," for example, "in a communicative situation," may not be strictly possible, since we know so little about how grammar knowledge interacts with other components of communicative language ability. To attempt to isolate any single component of language ability, while desirable, may be fruitless, given our current level of understanding. Once again, more research is needed. We need to know more about how raters arrive at judgments, what aspects of the discourse they attend to in making their ratings, and how different raters arrive at similar ratings for perhaps very different reasons.

Rating is, in my view, the other side of the speech production coin: speech comprehension. Raters are comprehenders, and they function as an audience for the test taker's performance. Instead of being severely limited in what aspects of a spoken performance they attend to in making their ratings, they need to be allowed (indeed, encouraged) to perform as normal listeners in making sense of the speaker's message.

Thus, raters will be encouraged to make use of their own linguistic and world knowledge in establishing communicative goals and intentions by means of metacognitive and language strategies, to formulate an interpretation of performance data. Raters must be assisted in attending to all relevant aspects of the performance, including its grammatical, textual, illocutionary, and sociolinguistic features, in light of the context established by the method facets. The difficulty that this entails is an uncertainty on the part of raters as to the categories of judgment they must adhere to in making their ratings. 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. Despite this uncertainty, however, it will be possible 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, comprehensible,* by means of "fuzzy logic." To help test developers understand somewhat the nature of rating using vaguely defined categories, a discussion of fuzzy set theory is offered next. It is by no means a full treatment of the topic but is intended to help justify an approach to language testing that is more in line with natural behavior than is perhaps currently the case.

Traditionally, 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." Figure 3 illustrates this state of affairs:

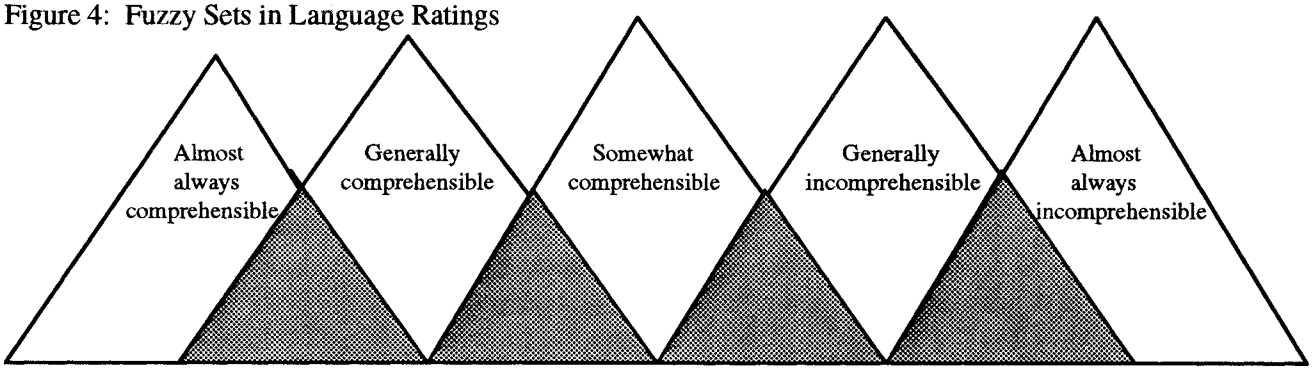
Figure 3: Standard Set



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, for example, is that the boundaries of judgment are vague and hard to define. For example, suppose we wanted to assign scores in bands indicating "almost always comprehensible," "generally comprehensible," "somewhat comprehensible," "generally incomprehensible," and "almost always incomprehensible." Figure 4, illustrates the problem:

Figure 4: Fuzzy Sets in Language Ratings



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." It can become a matter of policy to assign such a 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 (e.g., the glass is either half full or half empty). In natural language use, however, 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. Humans do this naturally by weighing the multidimensional facets of the performance in a way that may be different from that of other raters, yet "close enough" to be a professional response. In addition, 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 American Council on the Teaching of Foreign Language (ACTFL) oral proficiency scales, but no 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, although 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). I would not advocate the use of naive raters in future test development on the basis of this study. It does seem clear, however, that if raters are given simple rules or guidelines (such as may be found in many existing rubrics for rating spoken performances), they can use "negative evidence" provided by feedback and consultation with expert trainers to calibrate their ratings to a standard. They need to be encouraged to use their native speaker ability in assigning performances to score bands. (See, for example, Hadden, 1991 for a discussion of factors in ESL teacher and nonteacher ratings of second-language performance.)

The use of computers to analyze, and even score, speech production is nothing new (cf. Molholt & Presler, 1986; Rekart & Dunkel, 1992). A computer using "fuzzy logic" to score spoken language performance, however, is something new. A "fuzzy logic" computer using silicon "neural networks" can "learn" to emulate this human ability by responding to input according to simple rules, then being told by a supervisor when its response differs from the desired output. The "teacher" corrects the responses to sample input until the network responds correctly every time. Computer rating of spoken language performances might even be possible in the near future (cf. Kosko, 1991, and Kosko & Isaka, 1993).

In summary, the point of this discussion is to help developers of test scoring rubrics understand why it is impossible to specify entirely the defining features of score categories such that raters will be able to place a performance unerringly into the correct one. To make consistent ratings of spoken language performances, raters need some simple rules or guidelines for judging input and negative evidence regarding the correctness of their output.

D. Discrete or Integrated Skills?

The final issue in this section is whether speaking should be assessed "discretely" or integrated with other skills in the test battery. We recognize the distinction between "integrated" and "integrative." The latter term has been used for many years to refer to tests which do not distinguish between the formal components of language competence (cf. Oller, 1979) and contrasts with "discrete point" tests, which do attempt to assess formal knowledge components separately. The present discussion is about whether we can reasonably test two or more *skills* at one time and whether separate scores can or should be given for the different skills. Several issues must be considered. First, it is logically and practically impossible to test speaking *apart* from some other skills: the test taker must receive input in some form, whether written or spoken. Thus, facets of the input will influence the "pure" speaking ability measure. Second, as is clear from the Speech Production Model discussed earlier, speech production and speech comprehension are two sides of the same coin: the speech comprehension system is an integral part of the production model. Third, as discussed previously, the raters of the speaking performance make their judgments as competent native speakers do by using their own comprehension skills, a further potential source of bias in the assessment of speaking ability. Therefore, it is my view that speaking should be tested alongside the listening skill. The real question, it seems to me, is whether to give separate scores for speaking and listening.

Brown and Yule (1983) are quite pessimistic about the possibility that listening can even be assessed. They point out that in listening comprehension, unlike reading comprehension, there is no "text" available to compare against the evidence of comprehension. This is because in listening, the "text" is a representation in the test taker's head, not directly available for checking. They note that in tests of listening, there is often a very large amount of input (in the form of preparation materials), and very little output. Only by providing a very large number of items, then, can reliability and discrimination be achieved. They conclude that listening tests also "fail to provide us with any genuine insights into the processes by which listeners come to understand what they hear" (p. 148). The TOEFL 2000 paper on testing of listening comprehension will no doubt offer arguments that are far more optimistic. At present, however, it seems to me that listening and speaking are theoretically and practically very difficult to separate. I recommend that serious consideration be given to integrating them, both methodologically and

psychometrically. That is, I believe we should consider an oral/aural skills test, where the test taker uses his or her communicative language ability to produce and comprehend meanings in a variety of tasks and receives a single score reflecting the performance.

Not much previous research or experience with the integration of the speaking and listening skills exists, as Weir (1990) makes clear. The most common approach to integration has been to provide a "story line," or thematic strand, through the test, so that material the test taker hears might become material for an oral task, or content that is presented in written form might be integrated into a writing task. The British Associated Examining Board *Test in English for Educational Purposes (TEEP)* is an example. In Paper II of *TEEP*, candidates are first required to read a passage and write short answers to questions about it. Next, they listen to a short interview on tape, following a written outline consisting of a number of questions. They make notes while they are listening and use the notes to answer the questions. Finally, they are asked to write a summary of relevant information from the reading passage and the taped interview. The speaking component of *TEEP*, however, is not related to the thematic material of the rest of the test. A profile score is given in each of the four skill areas.

Another British test, the *International English Language Testing System (IELTS)*, administered by the British Council and the University of Cambridge Local Examinations Syndicate, integrates reading and writing skills: a passage in the reading subtest is used as material for a task in the writing subtest. The listening and speaking subtests are not integrated thematically. Like *TEEP, IELTS* reports a band score for each separate skill: reading, writing, listening, and speaking.

Weir (1990) notes two advantages to integrated-skills tests: authenticity and the use of context. The authenticity issue concerns the simulation of reality: students in academic situations have to read a variety of texts, listen to lectures, and discuss work with instructors and peers, all in preparation for a writing assignment. The integration of skills means that extended contextualization can be achieved: the text taker does not have to continually switch discourse domains. On the other hand, Weir sees problems with the integration of skills: full authenticity is probably not achievable; the psychometric requirement of local independence of items and tasks cannot be maintained in integrated tests; such tests are difficult to produce, take, score, and interpret; and little is known about the relative advantage of enhanced validity of integrated tests versus the potential loss in reliability. As is customary, Weir, too, calls for more research.

My recommendation that a single oral/aural score be reported is based partly on my concern, following Brown and Yule (1983), that it may be impossible to make a valid assessment of listening comprehension in any case, but also on a question about the usefulness to score users of separate speaking and listening scores. The justification for reporting separate scores is that it allows receiving institutions to decide what type of instruction to provide candidates who fall short of required proficiency levels. It remains to be seen whether this justification will remain valid for TOEFL 2000. I believe it is likely that instructional programs will no longer differentiate between the two skills (see, for example, Murphy, 1991); institutions may no longer be interested in diagnosing language deficiencies and providing for remediation. Research will be needed on the directions of teaching and admissions policies in coming years.

IV. Summary

A. Main points

1. Two-dimensional continuum: context-embedded/reduced on one axis, cognitively demanding/undemanding on the other.

2. Speech production depends on five language processors (conceptualizer, formulator, articulator, auditor, comprehension system), two knowledge stores (knowledge component, lexicon), and a strategic component (metacognitive, language and cognitive strategies).

3. Knowledge component related to context

4. Strategic component related to cognitive demands of task

5. Language strategies interpret context, set communicative goals, establish communicative plan, execute plan.

6. Conceptualizer's macroplanning stage selects information for meeting communicative goal; microplanning stage is language specific in bilingual individuals and marks information for use: preverbal message.

7. Formulator selects lexical units, encodes them grammatically and phonologically to produce a phonetic plan.

8. Both the phonetic plan and actual speech may be monitored by speech comprehension system.

9. Model here is compatible with that of COE.

10. Metacognitive strategies have access to language knowledge in noncommunicative situations.

11. Language strategies direct setting communicative goals, carrying out communicative plans.

12. Language users differ in preference for strategic styles.

13. Surface production in a second language may not reflect underlying processing in that language.

14. The interaction among the components of language knowledge is not well understood.

15. Tests differ with respect to degree of specification of context, from highly specified "field-specific tests" to underspecified "general tests." "Pan-technical tests" are those at a middle level of specification.

16. Specification of context requires more than simple changes in topic or content: there is a threshold level of contextual features needed to bring about a perceived change in context.

17. Tasks that are new to test takers make greater cognitive demands than those that are familiar.

18. If there is to be congruence between elicitation of language performances and the interpretation of those performances, there must be congruence between the types of knowledge required by the test tasks and the types of knowledge demanded by the situation for which the test results are to be used.

19. Context in natural language is signaled by conventionalized linguistic and paralinguistic cues that have their functional counterpart in test method facets.

20. Method facets can be manipulated to design a test for a specific population: international applicants for admission to North American colleges and universities.

21. Context-based tests will provide more useful information than "general-purpose tests" when the goal is to make situation-specific judgments about test takers' communicative language ability.

22. In speaking performances, an audience or interlocutor is normally assumed, the performance is transitory, immediate, and modifiable, and these factors bring about differences between spoken and written discourse.

23. Spoken *academic* discourse is marked by the relatively infrequent use of reduced forms and the expression of personal attitudes and feeling, the frequent use of referential elaboration and informationally dense structures, and the moderate use of passives.

24. Both basic interpersonal communication skills and cognitive academic language proficiency may be used in academic situations.

25. There are both universal system constraints and universal "ritual" constraints on spoken communication that are realized differently in different languages and are a part of what it means to be judged a competent speaker of a language.

26. Situational context interacts with test methods to create unique forms of discourse.

27. Attempting to isolate single components of language ability for rating may be fruitless, given our primitive state of knowledge about the interaction of such components in communication.

28. Raters should be encouraged to perform as normal listeners, making sense of a speaker's message.

29. Fuzzy rating categories cannot be specified sufficiently to allow raters to agree precisely on what they mean.

30. Raters will be able to confidently and reliably place performances in judgment categories by being given simple guidelines and feedback from expert trainers concerning their agreement with a desired outcome.

31. "Fuzzy logic" computers can probably be "trained" to rate speaking performances.

32. Speaking and listening skills are in principle and in practice difficult to separate, speech production and comprehension systems are closely related, and test takers' speaking performance and raters' comprehension performances merge the two aspects of language.

B. Recommendations for TOEFL 2000

1. Pan-technical tests: middle level of contextualization

2. Sufficient cues to engage discourse domain

3. Nonthreatening, though "stretching"

4. Instructions, content, genre, language of test directed at specific population: international applicants to North American colleges and universities

5. Methods manipulated to engage features of spoken discourse, including self-correction and message adjustment.

6. Raters need simple guidelines and feedback from expert trainers to place performances into "fuzzy categories."

7. "Fuzzy logic" computers may be able to rate speaking.

8. Speaking and listening tested together, with a single oral/aural score given

9. Computer "talking head" delivery, rating.

C. Research

1. Understand the nature of language proficiency in an academic context
2. Understand the nature of spoken discourse in language testing
3. Understand the conditions that influence test performance
4. Should test sections, items, or tasks require test takers to display ability in each knowledge component, or should such differentiation be a part of rating?
5. Confirm the Speech Production Model, nature of language knowledge, interaction among components.
6. Strategic styles/preferences
7. Relationship between method facets and engagement of discourse domains
8. Psychometric effects of fuzzy rating categories
9. Use of fuzzy logic computers to rate speaking performances
10. Use of computer delivery technology to provide speaking test input
11. Effects of integration of oral/aural skills in testing
12. Score user needs for diagnostic/profile scores

References

Adamson, H. D. (1993). *Academic competence.* London: Longman.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing 2*: 192-204.

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. (1991). What does language testing have to offer? *TESOL Quarterly, 25.4:* 671-704.

Bardovi-Harlig, K., & Hartford, B. (1990). Congruence in native and non-native conversations: status balance in the academic advising session. *Language Learning, 40*: 467-501.

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

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

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students.* (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

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

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

Chapelle, C., & Douglas, D. (1993, March). *Interpreting second language performance data.* Paper presented at Second Language Research Forum, Pittsburgh, PA.

Chapelle, C., & Green, P. (1992). Field independence/dependence in second language acquisition research. *Language Learning, 42*: 47-83.

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

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

Clapham, C. (1993). Can ESP testing be justified? In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 257-271). Alexandria, VA: TESOL Publications.

Cohen, A. (1993, April). *The language used to perform cognitive operations during full-immersion math tasks.* Paper presented at American Association of Applied Linguistics, Atlanta, GA.

Cohen, A., & Olshtain, E. (1993). The production of speech acts by EFL learners. *TESOL Quarterly, 27*: 33-58.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14*: 175-187.

Cummins, J., Swain, M. (1986). *Bilingualism in Education.* London: Longman.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics, 13*: 1-24.

Dorr-Bremme, D. (1990). Contextualization cues in the classroom: Discourse regulation and social control functions. *Language in Society, 19*: 379-402.

Douglas, D. (1992, March). *Testing methods in context-based second language research.* Paper presented at American Association of Applied Linguistics, Seattle, WA.

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. (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., and 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.* (TOEFL Research Report No. 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1992). *TOEFL 2000 purpose and domain.* Working draft. Princeton, NJ: Educational Testing Service.

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.

Erickson, F., & Schultz, J. (1981). When is a context? Some issues in the analysis of social competence. In J. Green & C. Wallat (Eds.), *Ethnography and language in educational settings* (pp. 147-160). Vol. 5, *Advances in discourse processes,* R. Freedle (Ed.). Norwood, NJ: Ablex.

Ervin-Tripp, S. (1993). Conversational discourse. In J. Berko Gleason & N. Bernstein Ratner (Eds.), *Psycholinguistics* (pp. 238-270). Orlando, FL: Harcourt, Brace, Jovanovich.

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

Fromkin, V. (1990). Speech Production. In J. Berko Gleason & N. Bernstein Ratner (Eds.), *Psycholinguistics* (pp. 272-300). Orlando, FL: Harcourt, Brace, Jovanovich.

Garman, M. (1990). *Psycholinguistics.* Cambridge: Cambridge University Press.

Goffman, E. (1976). Replies and responses. *Language in Society, 5*: 254-313.

Gold, R. (1991). Answering machine talk. *Discourse Processes, 14*: 243-260.

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

Gumperz, J. (1976). Language, communication, and public negotiation. In P. R. Sanday (Ed.), *Anthropology and the public interest* (pp. 273-292). New York: Academic Press.

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

Hale, G. (1988). Student major field and text content: Interactive effects on readding comprehension in the Test of English as a Foreign Language. *Language Testing, 5.1:* 49-61.

Hall, J. 1993). The role of oral practices in the accomplishment of our everyday lives: The sociocultural dimension of interaction with implications for the learning of another language. *Applied Linguistics, 14*: 145-166.

Harris, R. (Ed.). (1992). *Cognitive processing in bilinguals.* Amsterdam: North-Holland.

Hartford, B., & Bardovi-Harlig, K. (1992). Closing the conversation: Evidence from the academic advising session. *Discourse Processes, 15*: 93-116.

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

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

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*: 243-270.

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.

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.

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.

Kramsch, C. (1993). *Context and culture in language teaching.* Oxford: Oxford University Press.

Lambert, W., & Tucker, G. R. (1972). *Bilingual Education of Children: The St. Lambert Experiment.* Rowley, MA: Newbury House.

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

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: The MIT Press.

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

Molholt, G., & Presler, A. (1986). Correlation between human and machine ratings of *Test of Spoken English* reading passages. In C. Stansfield (Ed.), *Technology and language testing* (pp. 111-128). Washington, DC: TESOL Publications.

Murphy, J. (1991). Oral communication in TESOL: Integrating speaking, listening, and pronunciation. *TESOL Quarterly, 25*: 51-76.

Olaofe, I. (1992). A communicative model for assessing second language performance. *IRAL, 30*:207-222.

Oller, J. (1979). *Language tests at school.* London: Longman.

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 of Foreign Student Advisors.

Powers, D. (1985). *A survey of academic demands related to listening skills.* (TOEFL Research Report No. 20). Princeton, NJ: Educational Testing Service.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for specific purposes, 9*: 109-122.

Rekart, D., & Dunkel, P. (1992). The utility of objective (computer) measures of the fluency of speakers of English as a second language. *Applied Language Learning, 3*: 65-85.

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, PA: Center for Curriculum Development.

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

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14*: 357-386.

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

Selinker, L., & Douglas, D. (1985). Wrestling with "context" in interlanguage theory. *Applied Linguistics, 6:*190-204.

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

Snow, R., & Lohman, D. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (pp. 263-331). New York: Macmillan.

Stansfield, C. (Ed.). (1986). *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference.* (TOEFL Research Report No. 21). Princeton, NJ: Educational Testing Service.

Sternberg, R. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tedick, D. (1990). ESL writing assessment: Subject-matter and its impact on performance. *English for Specific Purposes, 9:* 123-144.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*: 713-730.

Tyler, A., & Bro, J. (1992). Discourse structure in nonnative English discourse: The effect of ordering and interpretive cues on perceptions of comprehensibility. *Studies in Second Language Acquisition, 14*: 71-86.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques.* Cambridge: Cambridge University Press.

Walker, C. (1990). Large-scale oral testing. *Applied Linguistics, 11*: 200-219.

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

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*: 693-712.

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

58701-14589 • Y37M.75 • 253706 • Printed in U.S.A.