12 *Evaluating lecture comprehension*

Christa Hansen and Christine Jensen

Abstract

This chapter focuses on the development of a listening test that will be used for placing students in intensive English classes or exempting them from further English language coursework. The test uses excerpts from actual university lectures and a short answer format to test more directly the listening skills students need to navigate in a U.S. university classroom. The chapter has three main sections: listening comprehension and lecture discourse theoretical considerations, a description of the format of the test based on such considerations, and statistical analysis of the performance of the test. Included in the statistics is the investigation of such issues as the effect of using both technical and non-technical lectures as listening stimuli, the effect of prior knowledge of topic on test performance, and performance of different proficiency levels on different types of questions. Recommendations are made for teaching and testing based on the findings.

Introduction

There have been two competing traditions in language testing, indirect and direct testing. Indirect tests "tap 'true' language performance obliquely or indirectly" (Henning 1987), predicting performance in language use situations. The discourse and tasks are designed to be generally accessible with a greater emphasis on skills and microskills. The tests are less natural, more contrived and are what many people call tests of general language proficiency. The emphasis is on reliability, getting the same results with different forms, different administrations, and being able to test any population. It is important to be able to say how one individual's performance compares to that of all the other individuals throughout the world who have taken the test. Examples of indirect tests of listening would be the Michigan Test of Aural Comprehension and Section A of the listening section of the TOEFL test.

Direct tests, on the other hand, measure language use in what Henning (1987) calls "real and uncontrived communicative situations." They

https://doi.org/10.1017/CBO9781139524612.019 Published online by Cambridge University Press

emphasize attaining the proficiency to perform the particular tasks needed in real world situations. These tasks are specific to the language need or performance area. Direct tests emphasize macroskills or tasks rather than the skills or microskills emphasized in indirect testing. There is a great concern about content validity and positive washback; the test reinforces the principle that both teaching and learning should focus on what students really need to know. It is important to know how a particular individual will perform in the specific situation that will follow.

The question we wanted to address with our research was whether the comprehension of academic lectures can be measured directly using authentic discourse as stimuli. In order to do this we needed to study the literature, develop a test that would directly test the comprehension of academic lectures, evaluate the performance of this test, and investigate research questions related to using this type of test.

In the first third of this chapter, we look at a number of theoretical issues, including what listening comprehension is, the roles of short-term and long-term memory in decoding information, and how scripts and schemas are used to help interpret information that listeners take in. The discussion of theoretical considerations then turns to the features of lecture discourse.

The second major portion of the chapter describes theoretical considerations which guided decisions made about the format of the test we developed. That test, called the T–LAP for the Test of Listening for Academic Purposes, is described in detail.

At this point the chapter turns to a discussion of the research conducted on the performance of the test. The first phase of the research is designed to see whether the T–LAP satisfies reliability and validity criteria critical to fair and consistent testing. In the second phase of the research, issues are raised that might concern educators looking at using this type of test. These issues include what content areas should be used for lectures, whether a single test can be used to place students with a broad range of proficiency levels, and whether prior knowledge of a lecture topic will give some test takers an unfair advantage. We will discuss the implications of our findings for teaching listening comprehension.

Listening comprehension

When we look at listening comprehension from the perspective of what listeners do, we find that listening comprehension is not *a* process but the result of a series of processes. These processes include, but are not limited to, phoneme recognition, morpheme chunking, lexical selection, and creation of a referential meaning for words. It has been a matter of

debate for those studying first language listening comprehension as to whether these processes are ordered in a serial fashion in which higher level decisions (i.e., clause or sentence level decisions) do not affect lower level decisions (phonemic or word level decisions), or whether there are interactions among the higher level and lower level processing decisions. (Carroll and Bever 1976; Fodor, Bever, and Garrett 1974; Forster 1979; Garrett 1978; Levelt 1978; Marlsen-Wilson 1976; Marlsen-Wilson and Tyler 1980) Evidence from first language research on listening comprehension (Marlsen-Wilson and Tyler 1980) and from error analysis of first and second language listening comprehension (Voss 1984) indicates that there is indeed interaction among the different processing levels. In fact, Voss concluded from the evidence he gathered in his error analysis work that the stretch governed by a single decoding decision could not be the segment, word, or even the tone group alone, nor is speech primarily processed through a sequential identification of segments and units of increasing size (1984: 119).

van Dijk and Kintsch (1983) have expanded upon the theme of interaction among the different listening processes in their development of a model of discourse comprehension in general, and listening comprehension specifically. They theorize that the stream of sound is held very shortly in the short-term memory where phoneme recognition and morpheme chunking is begun. It is there that listeners call on their knowledge of syntactic structure to organize the chunks into clauses. These clausal units are matched with information from the long-term memory to elaborate and verify the interpretation of the input.

As listeners process input, they develop a semantic representation of the text of the discourse in the long-term memory which van Dijk and Kintsch have labeled as the *textbase* (1983). The textbase is organized into semantic propositions with referential meaning. At the same time that they are creating a textbase, listeners create a unique *situational model* of the input, i.e., a cognitive representation of events, actions, and the people a situation is about, to make inferences from their knowledge structures to the referential propositions of the textbase. They do this by collecting information from the *scripts* or *schemas* from their knowledge base in the long-term memory. Scripts and schemas are knowledge organized around "predetermined stereotypical sequences of action that define well-known situations" (Schank and Abelson 1977: 41). In addition to stereotypical sequences of action, there are textual schemas, that is, knowledge of discourse-level conventions of a text; pragmatic schemas, knowledge of speech acts; contextual schemas, knowledge of the discourse situation; and rhetorical schemas, knowledge of the organizing conventions of a type of discourse. This configuration of listening comprehension takes into account the effect of a listener's real world knowledge and experience on a listening situation. Once

semantic propositions are developed, the element of the proposition that connects it with the previous propositions is held in the short-term memory to aid in connecting clauses together for an interpretation of the whole text. As the textbase is expanded, propositions are fitted into the macrostructure, or overall structure, of the discourse in the long-term memory in an approximation of a coherent whole.

Listeners use two major coherence strategies in the listening comprehension process, *global* and *local coherence strategies* (van Dijk and Kintsch 1983). Local strategies are used to connect a clause to the preceding clause and to make sense of the discourse at the sentential level. Rather than connecting one sentence to the next, global strategies are used to define the macrostructure of the discourse. Global strategies are used to recognize the discourse theme or topic, to recognize the relationships among the major ideas of the discourse, and to recognize the overall structure of the discourse. The interplay of information between the global level and the local level is used in the local strategies to predict and to verify sentence connections. However, local strategies need the support of information from the global level to be able to interpret consecutive sentences within a discourse passage. Otherwise, language users build a textbase without reference to all the information relevant to an adequate understanding of the text. Interpretations created through local strategies are used to verify the validity of the global guesses. All of this guessing, interpreting and interaction at the global and local level happens without waiting for a clause to be completely interpreted or even stored in the short-term memory space.

The use of global coherence strategies is referred to as *top-down processing* of information by Voss (1984), van Dijk and Kintsch (1983), and Shohamy and Inbar (1988). Top-down processing allows the language user to set up expectations about structures, meanings of sentences and the whole text. The use of local coherence strategies is characterized as the main *bottom-up processing* strategy. In contrast to top-down processing, bottom-up processing consists of interpreting the sound stream word by word to build a representation of the discourse. Proficient language users use both strategies to understand a text. They begin to process the sound stream by using bottom-up strategies to identify words and build clauses; at the same time they build a global macrostructure to be used as a top-down device to interpret subsequent sentences (van Dijk and Kintsch 1983). Voss's (1984) research demonstrates that relying solely on bottom-up processing rather than using it in combination with top-down processing is a less effective listening comprehension strategy for native speakers and second language speakers. In fact, he concluded that "successful speech perception . . . depends on the application – as a final step – of top-down procedures

assigning ultimate values to segments and other lower order units on the basis of hypotheses about longer stretches of speech" (1984: 119). Gary Buck's work on second language listening comprehension also supports van Dijk and Kintsch's model of discourse comprehension (1990). When he used introspective reporting techniques with second language listeners, he found that listeners have a different cognitive environment for listening to the same passage a second time. Cognitively, they are not listening to the same thing twice. This is certainly understandable in the light of what van Dijk and Kintsch have postulated about discourse processing. After listening to a passage for the first time, a listener's understanding of the discourse would be stored in the long-term memory in a propositional format rather than being stored verbatim in the memory. These propositions can be accessed to help create a situational model of the text for elaborating on the textbase that is being developed as they listen. This would allow listeners to use the propositions identified from the first time they listened to a passage to set up predictions and expectations of the direction of the text.

Lectures

Lectures are extended pieces of discourse that are delivered by one person to a group of people. They may range from an extemporaneous expostulation on a topic, to speaking from an outline or from detailed notes, to delivering memorized scripts or reading written scripts. Because academic lectures are rarely memorized and delivered or written and read, they contain features that have been labeled by Tannen (1982) as *oral* features, in contrast to features that predominate in written discourse. These features include the pauses, hesitancies, misspeaks, and disfluencies that reflect the spontaneity, fast pace and temporary nature of spoken discourse.

Spoken discourse is produced in spurts of language (Chafe 1979; Brown and Yule 1983), labeled as *idea units* by Chafe (1979). He defines an idea unit as having a single intonation contour followed by a pause. The idea units in lectures have a mean word count of 11 words, whereas the mean word count for idea units in conversations is 7 words (Chafe 1979). Idea units in lectures are expanded through the use of a number of different syntactic devices such as nominalizations, attributive adjectives, indirect questions, complement and restrictive relative clauses, adverbial phrases and prepositional phrases. Thus, lectures exhibit a greater degree of syntactic complexity and more literary vocabulary than is found in informal speech situations. These features are reflective of the planned nature of a lecture and the formality of the speaking occasion (Tannen 1982).

In a lecture situation, the communicative focus of speakers is on

disseminating information to the audience. To facilitate the audience's understanding, speakers present the information in a structured fashion that follows a logical sequential argument structure. They use thematic redundancies, not only to emphasize important points in the lecture, but also to help the audience deal with the pace of the flow of information and as an aid to their memories. More information is encoded in an idea unit in a lecture than in conversational discourse, but less than in such written and read discourse as news broadcasts (Shohamy and Inbar 1988). Propositional density, the amount of information encoded in an idea unit, is a feature that is affected by the degree of planning by speakers and their attention to and awareness of the audience's ability to cope with the flow of information.

In most lecture situations, speakers do not have the opportunity to negotiate meaning and verify the communicative effectiveness of the discourse with the audience. In our survey of university classes (Jensen and Hansen, in progress), we found that the number of students in the class and the format of the class directly influences the amount of listener–speaker interaction. Only in small (20 students maximum) discussion-type classes are students encouraged to interact with the speaker and the material in an active manner. In classes with up to 50 students, the speakers often make time for students to ask clarifying questions during the lecture presentation. They also tailor their presentation to maintain the attention of their audience by relating the information that they are disseminating to popular topics of the day. In very large classes (more than 100 students) speakers field questions before and after the lecture, but they rarely include a question period during their presentation. To compensate for the lack of interaction between speaker and audience, speakers often use the pronouns *we* and *you* in their presentation of information in order to develop and enhance the audience's awareness of a shared context.

Speakers also compensate for the lack of interaction with the audience by using *meta-talk* and other discourse markers to signal topic changes in a lecture (Hansen 1991). Meta-talk is defined by Schiffrin (1980) as "talk about the talk." Meta-talk has referents that point to items in the text and verbs that name acts of speech. These expressions are used to indicate something that will be done to a piece of talk, for example: "I want to say a little bit about each of these allotropes in turn . . . " (Hansen 1991: 65). Meta-talk also includes the use of expressions which have an evaluative or directional function such as *I mean*, *for example* or *in fact*. Topic change is also signaled with pauses, change in viewpoint, change in time or place, topic titles, and such discourse markers as *on the other hand*.

To recap, lectures can be characterized as planned, message-oriented discourse delivered by one person to a group of people. There is a

minimal amount of interaction between speakers and listeners. Lectures are syntactically complex and have a literary rather than a colloquial vocabulary. But they also contain the following *oral* features: redundancies, pauses, disfluencies, misspeaks and repetition of information. Gary Buck (1990) found that the normal speaking rates, pause structure and disfluencies of authentic oral discourse, i.e., Tannen's oral features, are what distinguish the listening trait from the reading trait in testing. This means that in order to separate statistically test takers' listening comprehension skill from their reading comprehension skill on a language proficiency test, the listening stimulus must have the features of oral discourse listed above. To present a lecture that has been scripted and read aloud in a listening comprehension teaching or testing situation does a disservice to the students. This is not the type of material that they will have to grapple with when they attend a lecture.

Question types

To help in the decision as to the most appropriate types of questions to use to evaluate students' comprehension of lectures, we studied the results of Powers's 1986 survey (see also Flowerdew, this volume), in which university lecturers gave their opinions as to the importance of various listening skills to students' successful academic achievement, and Richards's 1983 list of micro-skills needed for academic listening. Richards's taxonomy includes, among others, such skills as the ability to identify the purpose and scope of a lecture, the ability to recognize key lexical items related to a topic, and familiarity with different styles of lecturing. Not all of the micro-skills listed by Richards can be assessed by the use of questions. Some can only be taken into account by ensuring that students are exposed to different accents, speeds, registers, and lecture styles. It would not be appropriate to base an assessment of students' skills in lecture comprehension in the following areas even if these skills do affect their classroom performance: knowledge of classroom conventions, ability to follow different modes of lecturing and recognition of instructional/learner tasks. Other tasks like recognizing markers of cohesion and signals of discourse markers are micro-skills that listeners use to help them recognize ideas, themes, and relationships among ideas. These skills do not have to be directly assessed in proficiency testing, but may be appropriate for diagnostic testing or achievement testing.

The nine most important listening activities as identified by university lecturers in the Powers survey (1986) were:

1. identifying major themes or ideas of lectures
2. identifying relationships among major ideas in a lecture

3. identifying the topic of the lecture
4. retaining information through notetaking
5. retrieving information from notes
6. inferring relationships between information supplied in the lecture
7. comprehending key information presented in the lecture
8. following the spoken mode of the lecture
9. identifying supporting ideas and examples in the lecture.

The micro-skills from Richards's taxonomy that the lecturers rated as important for academic success were those that address a listener's understanding of the main points and supporting details of a lecture.

We wanted to ensure that the questions we developed would actually assess students' abilities to understand academic lectures. After studying the most important activities listed in Powers's survey and studying Richards's academic listening micro-skills, we incorporated what we know about the process of listening comprehension to make decisions about the most appropriate ways to assess comprehension of lectures. We decided that there are two major task areas, *global comprehension* and *local comprehension*, that would be appropriate for evaluating listeners' understanding of lectures. Global comprehension calls for understanding the major themes and topics of the lecture, whereas local comprehension focuses on understanding specific items within the lecture, such as identifying key terms or extracting information from key clauses.

The questions that we developed for use on this test have two objectives: to evaluate test takers' understanding of the lecture content and to assess their use of listening skills. We have developed two types of questions to assess students' abilities in these task areas, *global questions* and *detail questions*. From a lecture comprehension perspective, global questions are used to evaluate listeners' understanding of the major points in a lecture. Some of the skills listed in Powers and Richards that could be subsumed under this question type are identifying major themes or ideas, identifying purpose and scope of lecture, identifying topic of lecture and following topic development, identifying relationships among units within the discourse, and inferring relationships. The listening skills a test taker needs to answer these questions include the ability to synthesize information across clauses (or idea units as defined by Chafe 1979) in the lecture and the ability to identify the macrostructural items of the lecture.

In contrast to global questions, the comprehension focus of detail questions is on the listeners' ability to extract important details from the lecture. Listeners need to recognize key lexical items in regard to subject/topic, to deduce meaning of words from context, to identify supporting ideas and examples, and to comprehend key information. For this type

of question listeners need to be able to extract information from within a clause (i.e., idea unit); detail questions do not call for synthesizing information across clauses. The detail questions that we have created do not include numerical details and names that are not directly related to the main topic because questions of this type were found by Shohamy and Inbar "to be unstable and serve no meaningful purpose as evaluation tools" (1988: 21).

Evaluating the comprehension of academic lectures

We believed that the comprehension of academic lectures should be measured directly. To evaluate better whether this was possible and how to do this in a testing situation, we taped introductory level university lectures and compared them to the audio-taped lectures used on the TOEFL (*Listening to TOEFL* 1989). We felt this study would also tell us whether we could use a commercial test for our purposes. Although the university lecturers we taped had different lecture styles, we found certain features to be common in their lectures. Lecturers restated or paraphrased key ideas two or three different times. They used pauses to give themselves or to give listeners time to organize material. Pauses were also used to indicate topic shifts (Hansen 1991). Hesitation words, disfluencies, and misspeaks typical of natural, conversational speech were features found in all of these lectures. Speakers used restatement, paraphrasing, pauses, pacing and a decreased syntactic complexity to control the density of propositions, as work by Shohamy and Inbar (1988) and Chafe (1982) would predict. When we examined examples of lectures that were used on TOEFL tests, we found that these lectures had been scripted and recorded. They lacked repetition and paraphrase; had the syntax of written discourse; and certainly lacked the pauses, misspeaks, and disfluencies which Buck (1990) identified as distinguishing the listening trait.

After comparing university lectures with scripted lectures and finding such distinct differences in the discourse, we decided it was essential not only to use a direct test but to use authentic lectures. We audiotaped class sessions in a variety of introductory level university classes in order to identify suitable topics. The lecture segments we chose to use in the T–LAP had certain features. These lectures were no longer than 10–15 minutes and could be condensed to 5 minutes by leaving out lengthy digressions or extra examples. They did not require prior knowledge of content or vocabulary and were not dependent on visual material. These segments were coherent as segments, exhibiting clear logic structures. Finally, there was an adequate amount of testable information in the pieces selected.

Once we had chosen lecture segments, we went back to the lecturers

whose lectures we had taped and asked them to deliver the lecture in a recording session in a sound studio as they would in a class. We wanted the speaker to self-edit to retain the organization, types of examples, and discourse and delivery styles of the original lectures. In order to keep the discourse natural, we asked the lecturers not to use scripts or to memorize the material. We offered to serve as the student audience for any lecturers who seemed to want an audience to react to. We compared the original recorded material to the studio recorded material to see if they could maintain their original styles. We had to abandon recordings if they could not maintain their classroom intonation, pacing, relative level of formality, and use of examples to support explanation of new concepts.

In addition, we recognized the importance of providing context to enable listeners to activate the schemas they have available including situational, rhetorical, knowledge-based, experiential, and linguistic. We provide a situational context where the listeners are told, orally and in writing, that they are students attending a lecture in, for example, a chemistry class. We also tell them what lecture topic they will be listening to. This allows the listeners to set up expectations and to make predictions about the content and structure of the information they will hear based on their prior knowledge of the topic and experience with the structure of this type of discourse.

After listeners are given the situational context, they are given time to preview the questions that are written in the test booklets. This is done to replicate the experience students have coming into class with expectations about what will be important from what they have read in textbooks, and predictions about what they will hear based on what they have heard in earlier lectures. By setting the context and allowing listeners to preview questions, we also hope to facilitate top-down processing. This allows listeners to set up expectations about structures, meaning of sentences, and the whole lecture, a strategy that has been identified as being critical to final assignment of meaning in successful speech perception (Voss 1984).

What should listeners be asked to do to show they comprehend lectures? The TOEFL, for example, asks listeners to listen first to a mini-lecture and then hear single-sentence questions and select the correct multiple-choice answer from four written responses. However, to answer such multiple-choice items successfully, listeners must

1. recognize and store any important information from a mini-lecture as it is read
2. listen to a question
3. read four responses
4. refer back to the information stored in long-term memory to find the propositional information that answers the question

5. select the response that most closely matches the stored proposition
6. repeat the whole process for each question.

It is easy to see that the cognitive load involved in answering such questions within a matter of seconds is heavy; clearly, listeners who have good memories, are fast readers, or are strong in grammar would enjoy an advantage on this type of test. Additionally, this type of test does not allow listeners to give their understanding of the information in their own propositional formats.

Those who favor using multiple-choice style listening tests might point out that university content classes still use multiple-choice tests especially in large introductory lecture classes. However, students taking such tests have had a chance to (1) read the material, (2) hear the lecture, (3) take notes, and (4) study both text and notes. In that case, students are being tested on their ability to assimilate and store written and oral material and on their ability to retrieve information from memory. They need to rely heavily on reading skills both in preparing for the test and in taking the test itself. Such tests certainly do not test listening.

On the T–LAP, we decided to use short answer responses because this type of task would be more appropriate to their own real-life situation, a criterion set forth by Weir (1990: 24). Students answer the questions in real time, meaning they write their answers as they hear the information. Using real time prevents the test from being a memory test: listeners can answer either with the information as stated in the lecture or they can use their own propositional formats if the information has already been transferred to the long-term memory. Since the questions are answered in real time, the questions follow the chronology of the lecture.

In order to focus on the listeners' comprehension of the lecture, we use two types of questions, detail and global. These question types cover the two major task areas, local and global comprehension. We have listeners answer detail questions the first time they listen to a lecture; we reserve synthesizing questions for a second play through of the same lecture. Since work by Shohamy and Inbar (1988) shows that less proficient listeners will not be able to answer global questions unless they have extracted the answers to detail questions appropriately, it seems fairer to allow second language listeners to extract details before they are asked to answer global questions in an evaluative instrument used to test learners with a broad range of proficiency levels.

In addition, Buck's research shows that the first play through adds information to a listener's cognitive environment to set up more accurate prediction and interpretation on the second play through. However, this is not to suggest that local and global strategies are used separately. To

answer a synthesizing question, listeners might first need to use local strategies to identify words and build clauses in order to start building a global macrostructure. A continuing interplay of information from the global and local levels is used to recognize details, set up predictions about the topic, and validate those predictions. The information cannot be obtained by simply pulling out transitions and cohesive markers. Nor will using sentential markers to combine the ideas in two consecutive sentences be enough. The interplay is more complex. Recognizing the topic will make recognition of detail level information more accurate, while the addition of specific details to the textbase will make further predictions of the structure and direction and relationships of global structure more accurate.

The T–LAP test

The Test of Listening for Academic Purposes (T–LAP), which we have developed, has two parts: an academic and a non-academic part. The first part, the non-academic section, is a series of 3–4 dialogues ranging from 0.5–1.5 minutes in length; each series is based on a central theme such as buying something, renting an apartment, or planning a trip. The responses are information transfer or short answer.

The academic part of the test, which we are focusing on in this chapter, has two 3–5 minute lectures, one from a technical discipline and one from a non-technical discipline. Before the listeners hear each lecture, the context is set orally and in writing; they are told the field of study and the topic of the lecture. Then, time is allowed for them to preview the questions before they hear the lecture. The first time through the lecture, the test takers answer detail questions; on the second time through, they answer global questions. They respond with short answers written in real time. These answers are scored by trained raters using an extensive answer key, one that essentially lists the possible answers or types of answers and assigns 0, 1, or 2 points of credit.

The population

The T–LAP is being developed for a population of students preparing for university course work in the United States. The T–LAP will be part of a battery of tests – reading, composition, grammar paraphrase, and listening – used to exempt students from language study or to place them in appropriate levels of language classes in an intensive English program. This test was developed to replace the Michigan-style test currently being used since that test is really a test of oral grammar, and since it does not distinguish well among listeners at the upper end. Students who are able

TABLE 1. TEST POPULATION IN GROUPS BY MICHIGAN-STYLE TEST AND APPROXIMATE TOEFL RANGES

Group	*n*	*Michigan style test score range*	*Approximate TOEFL score range*
Group 0	12	0–100	33–43
Group 1	14	101–119	43–46
Group 2	31	120–139	45–50
Group 3	86	140–166	50–56
Group 4	92	167–200	55–60

to pass that Michigan-style test are often not able to understand classroom directions or to follow academic lectures.

This population includes students from about 65 countries. These students are undergraduates and graduates as well as students admitted for language study only. The range of listening proficiency levels represented in the population of students used for this research project is shown in Table 1. The table shows the number of students in each proficiency level, the score ranges on the Michigan-style test currently used to place students, and approximate TOEFL ranges.

The placement in the groups is based on the Michigan-style test scores. Since the Michigan and the TOEFL do not separate listeners into the same groups, the TOEFL score ranges overlap. The TOEFL ranges were provided to give some idea of what the group levels represent.

Research questions

The research we conducted had two major phases. The first phase focused on whether a listening test using authentic discourse could satisfy reliability and validity requirements.

1. Will test takers get consistent scores no matter when they take the test?
2. Will this type of test discriminate well among proficiency levels?

The second phase was designed to look at specific concerns related to using a content-based test.

3. Will there be a significant difference in the difficulty level of the technical and non-technical lectures for our population?
4. Will this test work to separate out upper level listeners? Can a single test be used to test students with such a broad range of proficiencies?
5. Will prior knowledge of a topic give some test takers an unfair advantage?

TABLE 2. DESCRIPTIVE STATISTICS FOR THE T–LAP TEST

	Mean % correct scores	*S.D.*	*Reliability*
Total test	79.83	19.13	.92
Non-academic	76.34	16.18	.88
Total academic	49.55	19.10	.87
History	52.13	23.69	.79
Chemistry	47.82	19.60	.81

Methods

We administered the T–LAP to a population of 235 students enrolled in the intensive English program at the end of the spring semester in 1991. The test was administered within two weeks of the end-of-semester proficiency test so that we could compare results to those on the Michigan-style test.

Results: Phase 1

The test was analyzed using an SPSS program. The descriptive statistics for the test are given in Table 2. Results are given for the whole test, the non-academic and the academic sections, and the non-technical (history) and technical (chemistry) lectures which make up the academic section.

Reliability

We needed to know whether this test would rank students consistently; i.e., whether students with high proficiency levels would receive high scores. The Cronbach's alpha for the whole test (63 items) was .92, well above the .80 often used to demonstrate strong reliability. This indicates that the test is performing extremely well. The alphas are lower for the individual subtests, which is not surprising since alphas decrease as the number of items decreases. The lecture portion alone (25 items) had an alpha of .87. Even the lowest coefficient, .79 for the history lecture, is more than satisfactory for a subtest with 10 items.

Validity

The use of authentic lectures with response tasks that match those that successful students need to use gives the T–LAP strong content validity (i.e., it tests directly what students need to be able to do). However, tests must also satisfy the requirement of construct validity. The test should

TABLE 3. STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS FOR ACADEMIC AND NON-ACADEMIC SUBTESTS

Subtest	*Function coefficient*
Non-academic	.51971
Academic:	
History	.39248
Chemistry	.43573

Note: One significant discriminant function $U = 42$, x^2 (18) = 223.23, $p = .0000$ was obtained.

TABLE 4. STRUCTURE COEFFICIENTS OF NON-ACADEMIC AND ACADEMIC SUBTESTS

Subtest	*Structure coefficient*
Non-academic	.78165
Academic:	
History	.72149
Chemistry	.71282

assign scores consistent with language proficiency, separating out the different language levels.

We ran a discriminant analysis to look at how the non-academic subtest, the history lecture, and the chemistry lecture separated out listeners. To do this, the listening proficiency groups were defined as described in Table 1. Discriminant analysis was used to determine whether the T–LAP subtests predict group membership; in other words, would listeners from group 1 be assigned to that group based on scores on the new test. The discriminant function coefficients showed that each of the subtests contributes significantly to separating out the groups (Table 3). This is indicated by the positive signs and the fact that the coefficients are relatively closely grouped.

Structure coefficients were run (Table 4). These coefficients are meaningful if they are positive and greater than or equal to .30, showing that there is a strong correlation between the scores on the individual subtests and the composite which is the score which best assigns membership to the correct level. The non-academic subtest contributes the most to separating out the groups while the chemistry lecture

TABLE 5. MEAN DISCRIMINANT FUNCTION SCORES BY LANGUAGE PROFICIENCY GROUP

Group	*Mean discriminant function score*
0	−2.92
1	−1.97
2	−1.22
3	−0.09
4	1.18

contributes the least as the descending order of structure coefficients indicates. There is not a great difference in their relative contributions; these magnitudes may reflect no more than the difference in the number of items in the subtests.

Finally, when group centroids were run, they showed that groups were being separated out magnificently. The group centroids should increase as the proficiency level goes up, and there should be approximately the same distance between each of the neighboring centroids. Not only is the directionality correct (Table 5), but there is maximum separation between levels.

The T–LAP satisfied reliability and validity requirements at a very high level. We are satisfied that we can measure listening comprehension directly using authentic discourse as the stimulus.

Results: Phase 2

There are other questions raised by having a content-based test with authentic discourse. In general terms, these questions address these areas: What content areas should be used? Will this type of test better separate out upper level listeners? Will a test with detail and global questions be useable for a population with such a broad range of abilities or will low level listeners be disadvantaged? Will the specific content of lectures give students with prior knowledge an unfair advantage on the test.

Technical vs. non-technical lectures

Since our students represent around 50 different majors every semester, it was, of course, not possible to develop tests for each discipline. Instead, each T–LAP test form will have two lectures. We elected to use a technical and a non-technical lecture in each T–LAP as a result of a survey of our population which showed that of the students with

TABLE 6. MEAN PERCENTAGE CORRECT SCORES ON HISTORY AND CHEMISTRY LECTURES FOR PROFICIENCY GROUPS

	History		*Chemistry*	
	Mean	*S.D.*	*Mean*	*S.D.*
Low proficiency[1] $n = 119$	38	21	38	17
High proficiency[1] $n = 114$	66	16	58	16
Total group $n = 233$[2]	52	24	48	20

Note:
[1]Groups divided by Michigan-style test scores. Low proficiency = below 50th percentile, High proficiency = above 50th percentile.
[2]2 cases rejected because of missing data.

declared majors, 53% were in technical fields and 47% in non-technical fields.

As we piloted the tests, we wanted to see whether there would be a significant difference in the way our population would score on the non-technical and technical sections. This particular test form had a history and a chemistry lecture. In addition, we wanted to see whether there would be a difference in the way high and low proficiency students would perform on the two lectures since some believe that technical material will be less accessible. To compare the two sections, we translated the mean scores for the subtests to percentages. When the mean scores for the total group are translated to percentages, the total group scored 52% on the history lecture and only 48% on the chemistry lecture, a difference of 4% (Table 6). This was a significant difference (.001). The history lecture was easier than the chemistry lecture for the group as a whole.

After looking at how the whole group performed on the two lectures, we divided the test takers into high and low proficiency groups by their scores on the Michigan-style test. As Table 6 shows, the low group had a mean percentage correct of 38 for both lectures while the high proficiency group went from a mean of 66 on the history lecture to one of 58 on the chemistry lecture. The low group finds the two lectures equally difficult while the upper group finds the history lecture easier. The sharp contrast in the performance of the two groups can be seen graphically in Figure 1.

To test whether this difference in the change for the two proficiency groups was significant, we ran a MANOVA (Table 7). The results

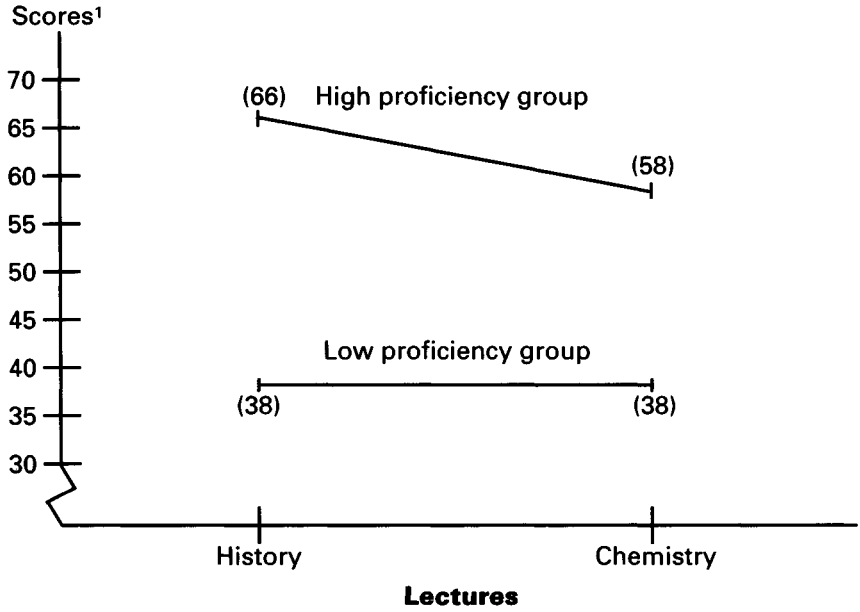

Figure 1 Graphic representation of change in performance on history and chemistry lectures for low and high proficiency groups

(p = .002) show that there is less than .2% probability that this difference would happen by chance.

Since the means for the two lectures are the same for the low group, it is important to ask whether a floor effect is involved. In other words, since the same test is being administered to persons with a broad range of proficiency levels, are the questions on the academic lectures so difficult that the low group could not answer them? One way to determine whether this is the case is to look at the range of scores, represented in the standard deviations shown in Table 6, for the low and high proficiency groups. The standard deviations for the low group for the two lectures (21 and 17) are larger than those for the high group (16 and 16). These large standard deviations for the low groups for both lectures indicate that these lectures do elicit a range of performance across the low group; it is safe to assume that there was not a floor effect.

The drop in scores on the chemistry test by the high proficiency group needs to be investigated further since other factors could have been involved. The chemistry subtest is the last on the test, and while the whole test takes less than 35 minutes, fatigue could have been a factor.

[1] Mean percentage scores from Table 6.

https://doi.org/10.1017/CBO9781139524612.019 Published online by Cambridge University Press

TABLE 7. TEST OF SIGNIFICANCE ON CHANGE IN PERFORMANCE ON HISTORY AND CHEMISTRY LECTURES FOR LOW AND HIGH PROFICIENCY GROUPS

	SS	*DF*	*MS*	*F*	*p*
Subtest	2007.71	1	2007.71	11.26	.001
Proficiency × subtest	1708.29	1	1708.29	9.58	.002
Error	41190.97	231	178.32		

TABLE 8. SYNTACTIC ANALYSIS OF CHEMISTRY AND HISTORY LECTURES (HANSEN, IN PROGRESS)

	History lecture	*Chemistry lecture*
Words	377	400
Syllables	580	598
T-units	26	23
Sentence nodes	55	49
Prepositional phrases	51	50
Modals	7	1
Present tense	15	31
Present progressive tense	5	2
Past tense	25	2
Future tense	1	0
Present passive tense	1	0

However, if that were the case it would be logical to assume that the lower proficiency group would also be affected by fatigue. One of the areas we are looking at more closely is that of relative syntactical complexity of the text for the two lectures. As Table 8 shows, the lectures are relatively equivalent in overall length – 377 words to 400 words, T-units, sentence nodes and prepositional phrases, indicating that syntactic complexity of the text is not a factor in the difference in performance for the two groups. Interestingly enough, the main area of deviation in regard to syntax shows up in the verb tenses; the history lecture has a more sophisticated verb tense system. This certainly does not explain why the chemistry lecture is more difficult for the proficient listeners.

Once the element of syntactical differences had been eliminated from consideration, we looked more closely at the vocabulary content of the two lecture texts and found that the chemistry lecture featured a more field-specific vocabulary than the history lecture. This may be the feature that distinguishes the two lectures.

TABLE 9. MEAN PERCENTAGE CORRECT SCORES ON DETAIL AND GLOBAL QUESTIONS FOR PROFICIENCY GROUPS

	Detail questions		*Global questions*	
	Mean	*S.D.*	*Mean*	*S.D.*
Low proficiency[1] $n = 119$	41	17	33	21
High proficiency[1] $n = 114$	63	15	60	18
Total group $n = 233$[2]	52	19	46	24

Note:
[1]Groups divided by Michigan-style test scores. Low proficiency = below 50th percentile, High proficiency = above 50th percentile.
[2]2 cases rejected because of missing data.

TABLE 10. TEST OF SIGNIFICANCE ON CHANGE IN PERFORMANCE ON DETAIL AND GLOBAL QUESTIONS FOR LOW AND HIGH PROFICIENCY GROUPS

	SS	*DF*	*MS*	*F*	*p*
Subtest	2949.05	1	2949.05	17.01	.000
Proficiency × subtest	721.15	1	721.15	4.16	.043
Error	40039.10	231	173.33		

Detail vs. global questions

In light of Voss's (1984) and van Dijk and Kintsch's (1983) work showing that proficient language users employ top-down processing skills, it would be reasonable to expect more proficient language users to perform better than less proficient language users on global questions, which would mean listeners with high proficiency would be separated out better.

To look at whether our data bore this out, we again divided the test takers into high and low proficiency groups by scores on the Michigan-style test. We compared the performance of the two proficiency groups on the detail and global questions by examining mean percentage correct scores. Not surprisingly, the high group outscored the low group by 60 to 33 on global questions (Table 9).

But did this really mean that global questions are functioning to separate out the high level listeners? Or, will there be a parallel difference on detail questions? The performance of the high group is quite

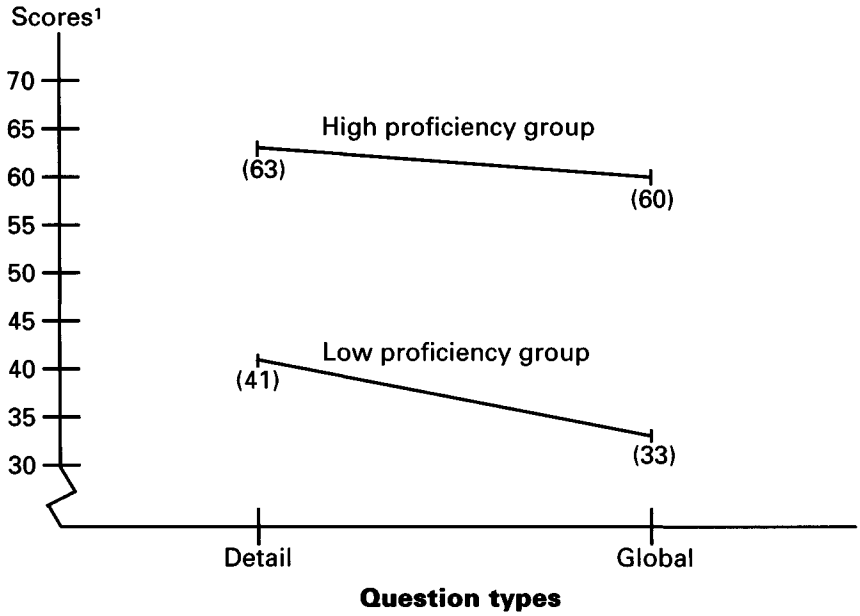

Figure 2 Graphic representation of change in performance on detail and global questions for low and high proficiency groups

stable on detail and global questions, decreasing only 3 percentage points, from 63 to 60. This comparison is displayed graphically in Figure 2. The change in performance for the low proficiency group is much sharper, dropping from 41 on detail questions to 33 on global.

The graphic representation makes it clear that the performance of the two groups does not remain parallel from detail to global questions. It appears that the global questions do in fact separate out the high group; that difference in the change of performance was significant at the 0.43 level (Table 10).

In summary, high proficiency listeners do better than low proficiency listeners on global questions. This is also true for detail questions, but the relative difference is greater on global questions. This would suggest that there is a difference in the processing strategies for high and low proficiency listeners.

In the next phase of our research, we are looking at the processing strategies used by the five different proficiency levels to extract information from lectures. Preliminary evidence from that research shows that low level listeners rely much more heavily on verbatim response

[1] Mean percentage scores from Table 9.

https://doi.org/10.1017/CBO9781139524612.019 Published online by Cambridge University Press

patterns, meaning they answer with the exact words used in the lecture. Global questions, however, call for synthesizing information across clauses, and, therefore, verbatim response patterns are less successful on those questions. In order to synthesize information successfully, listeners need to give propositional answers; in other words, these answers are in the listener's own words. This is a task that the high group is much more successful at. Interestingly enough, high proficiency listeners are also able to use propositional answers successfully in response to detail questions.

A second concern we had was whether this type of test could be used for the broad range of listening proficiencies represented in our population. Since it separated out the upper level listeners, would that mean it was not accessible to the low level listeners? A review of the group centroids given in Table 5 shows that, in fact, the separation of the lowest two groups is exactly what we would hope for: it is similar to the separation between other groups and certainly not a larger separation. The test was not relatively more difficult for these low groups.

Prior knowledge

One of the chief concerns that people have about content-based tests is whether the test will be biased in favor of listeners with prior knowledge. To test the effect of prior knowledge of a lecture topic on the performance of a test taker, we asked all the test takers to indicate whether or not they had studied the lecture topics before. Of the 235 test takers, 30 reported prior knowledge of the history lecture. In contrast, only 8 reported they had studied the chemistry topic.

To test the effect of prior knowledge of the history topic (Coronado's exploration of the New World), we ran multiple linear regressions which used the part of the test not being examined as a measure of listening comprehension proficiency. To determine whether prior knowledge helped to predict scores on the history subtest, the first step was to covary or control for the amount of variance that is accounted for by the performance on the remainder of the test. The performance on the rest of the test accounts for 48% of the variance (see R^2 in Table 11). When prior knowledge is added in as a predictor of performance on the history subtest, the R^2 does not change, indicating that prior knowledge does not have a predictive value. The finding that prior knowledge of the history topic did not improve listening scores was reassuring especially since we have a fairly large representation of Spanish speakers.

The same procedure was followed for the chemistry lecture. Again the measure of listening comprehension predicted the chemistry subtest scores. However, for the eight listeners reporting prior knowledge of

TABLE 11. MULTIPLE REGRESSION ANALYSIS OF PRIOR KNOWLEDGE ON HISTORY LECTURE

Step	*Variable(s) entered*	*b*	*Beta*	*R*	R^2	R^2 *change*
I	Total test minus history	.1992**	.6945**	.69	.48**	
II	Total test minus history	.1996**	.6960**			
	Prior knowledge of history	−0.0488	−0.0052	.69	.48**	.00

*$p \leq .05$, **$p \leq .01$

TABLE 12. MULTIPLE REGRESSION ANALYSIS OF PRIOR KNOWLEDGE ON CHEMISTRY LECTURE

Step	*Variable(s) entered*	*b*	*Beta*	*R*	R^2	R^2 *change*
I	Total test minus chemistry	.2443**	.6544**	.65	.43	
II	Total test minus chemistry	.2343**	.6283**			
	Prior knowledge of chemistry	2.5087**	.1557**	.67	.45**	.02

*$p \leq .05$, **$p \leq .01$

the chemistry lecture (allotropes of carbon), prior knowledge was a significant factor in their performance of the section. The R^2 change of .02 indicates that prior knowledge added 2% to the prediction of the performance on that subtest (Table 12). Somewhat surprisingly, that amount of change is significant.

However, the fact that there were only eight cases, and of those eight listeners seven were speakers of Arabic, makes the finding difficult to interpret. To know whether this finding is significant, we will need to collect a larger number of cases with prior knowledge of that topic. This will allow us to see if the particular language grouping is a factor and to find out why in this case there were so few self-reported cases of prior knowledge of a relatively basic chemistry topic. We should also find out whether test takers have studied the topic in English.

Our strongest evidence on the influence of prior knowledge, a group of 30 out of the total of 235 test takers with prior knowledge of the history topic, suggests that prior knowledge is not a factor in performance on the subtest involved. However, the fact that the evidence is mixed at least suggests to us that the exact underlying factors haven't been pinpointed.

Conclusions

The T–LAP was developed for the purpose of placing non-native speakers of English into listening courses in a college preparatory intensive English program. The final outcome of the listening coursework is that students will be operating in English-only university classrooms. The design problem we grappled with as test constructors was creating an instrument that would directly measure listening comprehension in replicated real world situations and would also reliably place students across the spectrum of language proficiency in the appropriate intensive English coursework. In the secondary phase of our research on this project we looked at such issues as performance on technical and non-technical lectures, performance across proficiency levels on different question types and the effect of prior knowledge of topic on test performance.

In order to measure listening comprehension directly in lecture situations, we used segments from actual university lectures for the listening stimuli, we set the situational context for the students before they listened to the lectures to replicate the classroom experience, and we used a short answer format to allow students to provide answers in their own words rather than recognizing answers developed by the test makers. Detail and global questions were used to assess students' understanding of the lecture. Detail questions focus on the important supporting information in a lecture and require listeners to extract information from within clauses; global questions focus on the main ideas and relationship of ideas in the lecture and require listeners to synthesize information across clauses. All of these considerations and decisions indicate that the test has high *content validity*; it directly measures students' listening comprehension proficiency in regard to academic lectures.

However, as Henning pointed out, content validity, which requires a diversity of items and comprehensive coverage of the content, often conflicts with the reliability of internal consistency of an instrument because this form of reliability depends upon the homogeneity of items (1987). The first question we addressed in our assessment of the test is whether the T–LAP is a reliable testing instrument. Does it rank students with high listening proficiency skills high and students with low skills as low? The statistical evidence (Tables 2, 3 and 5) has borne this out. It is reliable, parts and whole. This indicates that we have found a balance between content validity and internal reliability.

After looking at the reliability of the instrument as a whole, we investigated a number of questions about the test content. We had elected to use lectures from technical disciplines and non-technical disciplines on the T–LAP because our population almost equally divided

between these two general areas. This would ensure that the test would not give preference to students in non-technical fields over students in technical fields, or vice versa. The different statistics that we ran on the lecture sections of the test gave us a mixed picture as to the performance of the technical (chemistry) lecture and the non-technical (history) lecture. The Cronbach alphas for the subsections of the test show the chemistry lecture (.81) as more reliable than the history lecture (.79), but the structure coefficients identify the history lecture (.72) as more important for identifying group membership than the chemistry lecture (.71). Nevertheless, the lack of magnitude of these differences indicates that the two lectures are equivalent. Interestingly enough, the chemistry lecture proved to be harder than the history lecture with the performance of the high proficiency group marking that difference between the two lectures. The source of the difference in performance across the high proficiency group has not been identified yet. Because the two lectures have comparable syntactic complexity except for the amount of variety of verb tenses, syntax does not seem to be the source of difference between the two lectures. Although the lectures have similar syntactic complexity, the chemistry lecture features more field-specific vocabulary than the history lecture. This may even be a factor in the effect of prior knowledge of topic on performance on the lectures.

The other major area that we looked at in regard to the test is students' performance on detail and global questions. This area of research directly addresses the theories behind the construction of this type of test. Our assumption, based on previous work by Voss (1984) and Shohamy and Inbar (1988), was that high proficiency students would perform better on the global questions than low proficiency students, and this assumption was validated. High proficiency students performed well on both types of questions, while low proficiency students' performance dropped dramatically from the detail to the global questions. They did not maintain the same level of performance in relation to the high group on the two types of questions. We submit this as evidence that global questions are effective in spreading out test takers at the high end of the proficiency scale, a design problem that we needed to address with the new test. We also feel that it is indirect evidence that low proficiency students rely heavily on bottom-up processing skills and do not yet know how to process across clauses. We found that low proficiency students generally rely on extracting information from the text and recording it verbatim rather than formulating a response in their own words. This appears to be a factor in their performance on the test. When we looked at students' responses, we found that they can successfully use propositional responses for both detail and global questions, but verbatim responses are generally not a successful strategy for answering global questions, that is, questions that

call for synthesizing information across clauses. In contrast to low proficiency students, high proficiency students can synthesize information across clauses, indirect evidence that high proficiency students use top-down processing strategies in listening.

What are the implications of our findings for testing and teaching listening comprehension, specifically of lecture material? The first thing we can conclude is that it is possible and even efficacious to use authentic lecture material as the listening stimuli. The information we have about the performance of the test, and therefore the performance of the test takers, gives credence to the validity constructs and the psychological theories that we used in the construction of the test. Lecture comprehension requires much more from a listener than just to recognize phonemes and to understand information at the essential level. Listeners also need to be able to recognize the macrostructural items in a text, synthesize information across clauses and be able to put lecture information into their own words. Teachers and testers of listening comprehension must be willing to expand the scope of the amount of information and the type of information that students will be exposed to and tested on. Buck's 1990 finding about natural speech, making listening a trait clearly separable from reading, reinforces the need to provide students with material that features natural speech, not contrived written and read discourse. Test takers' performances on this test indicate that they can handle natural speech, extended discourse and technical and non-technical lectures, even at the lowest level of proficiency. What this means for teaching listening comprehension is that students from all proficiency levels should be exposed to natural speech and to extended discourse as a regular part of their listening curriculum. The curriculum should work on developing listeners' strategies to comprehend extended discourse. In order for listeners to comprehend extended discourse effectively, they need to use global and local coherence strategies, and both top-down and bottom-up processing strategies. Teaching students to integrate information culled from global strategies with the detail information from local strategies should be an integral part of a listening comprehension curriculum.

Acknowledgements

We would like to acknowledge the efforts of Dr. Glasnapp's EPR 921 class from Fall, 1991 in helping us with the statistical analyses for our research. We are grateful for the extra work of Monica Castator, Sandy Gahn and Jeff Townsend. We would especially like to express our appreciation to Jeff Townsend for his work on this project. His help on the interpretation of the results of the analyses was invaluable.

References

Brown, G., and G. Yule. 1983. *Discourse Analysis*. New York: Cambridge University Press.

Buck, G. 1990. The testing of second language listening comprehension. Ph.D. dissertation, University of Lancaster, England.

Carroll, J. M., and T. G. Bever. 1976. Sentence comprehension: a study in the relation of knowledge to perception. In *The Handbook of Perception, Vol. 5, Language and Speech*, E. C. Carterette and M. P. Friedman (Eds.). New York: Academic Press.

Chafe, W. L. 1979. The flow of thought and the flow of language. In *Syntax and Semantics* 12, T. Given (Ed.), 159–83. New York: Academic Press.

1982. Integration and involvement in speaking, writing and oral literature. In *Spoken and Written Language: Exploring Orality and Literacy*, D. Tannen (Ed.), 35–53. Norwood, N.J.: Ablex.

Dijk, T. A. van, and W. Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.

Fodor, J. A., T. G. Bever, and M. F. Garrett. 1974. *The Psychology of Language*. New York: McGraw-Hill.

Forster, K. 1979. Levels of processing and the structure of the language processor. In *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*, W. E. Cooper, and E. C. T. Walker (Eds.). Hillsdale, N.J.: LEA.

Garrett, M. F. 1978. Word and sentence perception. In *Handbook of Sensory Physiology, Vol. VIII, Perception*, R. Held, H. W. Leibowity, and H.-L. Teuber (Eds.). Berlin: Springer Verlag.

Hansen, C. 1991. Topics in a lecture: how does a linguistic analysis compare to the professor's and students' notes? Unpublished Master's thesis, University of Kansas, Lawrence, Kansas.

In progress. Syntactic analysis of T–LAP lectures.

Henning, G. 1987. *A Guide to Language Testing*. Cambridge: Newbury House.

Jensen C., and C. Hansen. In progress. Survey of university classes.

Levelt, W. J. M. 1978. A survey of studies in sentence perception: 1970–1976. In *Studies in the Perception of Language*, W. J. M. Levelt and G. B. Flores D'Arcais (Eds.). New York: Wiley.

Listening to TOEFL. 1989. 45–60. Princeton, N.J.: Educational Testing Service.

Marlsen-Wilson, W. D. 1976. Linguistic descriptions and psychological assumptions in the study of sentence perception. In *New Approaches to the Study of Language*, R. J. Wales, and E. C. T. Walker (Eds.). Amsterdam: North-Holland.

Marlsen-Wilson, W. D., and L. K. Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 8: 1–71.

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

Richards, J. C. 1983. Listening comprehension: approach, design, procedure. *TESOL Quarterly* 17 (2): 219–240.

Schank, R. C., and R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, N.J.: Erlbaum.

Schiffrin, D. 1980. Meta-talk: organizational and evaluative brackets in discourse. *Sociological Inquiry* 50: 199–236.

Shohamy, E., and O. Inbar. 1988. Construct validation of listening comprehension tests: the effect of text and question type. ERIC Doc. No. ED296594.

Tannen, D. 1982. The oral literate continuum in discourse. In *Spoken and Written Language: Exploring Orality*, E. Tannen (Ed.), 1–16.

Voss, B. 1984. *Slips of the Ear. Investigations into the Speech Perception Behavior of German Speakers of English*. Tubingen: Narr.

Weir, C. 1990. *Communicative Language Testing*. Great Britain: Prentice Hall International.