Videotaped formatted test: usefulness for measuring academic lecture listening ability of learners of English as a second language

Dong-il Shin

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Bilingual, Multilingual, and Multicultural Education Commons, and the English Language and Literature Commons

Recommended Citation

https://lib.dr.iastate.edu/rtd/205

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Videotaped formatted test: Usefulness
for measuring academic lecture listening ability
of learners of English as a second language
by
Dong-il Shin

A Thesis Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF ARTS

Department: English
Major: English (Teaching English as a Second Language/Linguistics)

Approved:

________________________
In Charge of Major Work

________________________
For the Major Department

________________________
For the Graduate College

Iowa State University
Ames, Iowa
1995
TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION .................................................. 1

CHAPTER 2. TEST DEVELOPMENT ............................................ 11
   Academic Listening Tasks .................................................. 11
   Consideration of Authenticity .......................................... 18
   Consideration of Interactiveness ...................................... 28
   Level of Comprehension Difficulty ..................................... 34

CHAPTER 3. METHOD .......................................................... 41
   Subjects ........................................................................... 41
   Test Materials .................................................................... 42
   Procedures ......................................................................... 46
   Analysis ............................................................................ 47

CHAPTER 4. RESULTS AND DISCUSSION .................................. 50
   Conventional Listening Test Task ....................................... 50
   Experimental Listening Test Task ........................................ 52

CHAPTER 5. CONCLUSIONS ..................................................... 66

REFERENCES ....................................................................... 71

ACKNOWLEDGMENTS ............................................................ 74

APPENDIX A. CONVENTIONAL TEST ...................................... 75

APPENDIX B. VIDEOTAPED EXPERIMENTAL TEST .................... 82

APPENDIX C. AUDIOTAPED EXPERIMENTAL TEST ....................... 92

APPENDIX D. SCORING CRITERION IN EXPERIMENTAL TESTS ....... 99
CHAPTER 1. INTRODUCTION

ESL teachers, curriculum writers, and test developers of recent decades have been enthusiastic to reflect authentic ‘target language use’ (TLU) situations in the materials they produce. The recent trend of language testing is also towards testing language use in a wider communicative sense. In second language (L2) listening comprehension testing, in contrast, there has been very little empirical work carried out to investigate the value of the authentic factors such as kinesics (Kellerman, 1992; Hurley, 1992; Pennycook, 1985), schemata (Carrell, 1983; Byrnes, 1984) or speech modifications (Chiang and Dunkel, 1992). Instead, audiotape formatted multiple choice listening tests (e.g., in the TOEFL, 1993) have been widely used for selection and placement purposes to determine whether new international students have sufficient listening abilities to study at academic settings in the North America. The lack of authentic factors in such norm-referenced tests (NRTs) raises the question of how test performance is affected by listening tasks which lack the variety of clues that normally help L2 listeners to infer meanings (Oxford, 1993), for example, contextual information related to nonverbal and environmental clues.

Recently, videotaped materials have been used in teaching and learning settings (Secules, Herron and Tomasello, 1992; Baltova, 1994) to provide TLU situations for classroom activities. These materials raise expectations about the potential usefulness of videotaped materials for NRT to evaluate listening ability of non-native speakers of English. The videotape formatted listening test is believed to have the
capability to make it possible to reproduce many characteristics of TLU in a more communicative sense. In addition to the apparent superior authenticity of the VFT, however, much more theoretical justification and experimental research are needed to support the usefulness of videotaped formatted listening test.

While little has been written on the topic of test usefulness which can provide a basis for designing and developing a test, Bachman and Palmer's recent work (1994) clearly presented test development procedures which maximize test usefulness qualities. They argue that test usefulness, which is based on interrelated test qualities, should be very important consideration at all stages of test development. Bachman and Palmer's notion of test usefulness was described as follows (1994 I: Chapter 3, p. 2):

Usefulness = Reliability + Validity + Authenticity + Interactiveness + Impact + Practicality

The following principles of Bachman and Palmer (1994 I) state that the test qualities in combination affect the overall usefulness of a language test.

Principle 1: It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.

Principle 2: The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

Principle 3: Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation (Chapter 3, p. 2).
The test usefulness quality of authenticity is defined as "the perceived relevance of the characteristics of a given language test task to the features of a TLU task and its associated situation" (Bachman and Palmer, 1994 I: Chapter 3, p. 7). This test quality can play an essential role of ensuring that test-takers' listening comprehension ability is measured using appropriate tasks so that test scores can be interpreted relative to the domain of interest.

As for the test usefulness quality of interactiveness, Bachman and Palmer (1994 I) note:

We define interactiveness as the extent and type of involvement of the test taker in accomplishing a test task. Unlike authenticity, which pertains to the correspondence between test tasks and TLU tasks, and thus must consider the characteristics of both, interactiveness resides in the interaction between the individual (test takers and language users) and the task (test or TLU) (Chapter 3, p. 9).

To develop interactive tests, they listed three characteristics of test takers that need to be considered in test development process: topical knowledge, affective schemata, and language ability (language knowledge and metacognitive strategies). Interactiveness is defined in terms of the ways in which these test takers' characteristics are engaged by test tasks. Figure 1.1 adapted from Bachman and Palmer (1994 I) shows the involvement of the major components in the model of language use. Language use is associated with the language user's topical knowledge (called 'knowledge schemata and real-world knowledge') and affective schemata ('emotional correlates of topical knowledge'), as well as all the areas of language knowledge.
Characteristics of the Language Use or Test Task and Situation:

Environment
Input, Response

Figure 1.1 Language Use and Language Test Performance
(Bachman and Palmer, 1994 I: Chapter 2, p. 34)
Language knowledge is “a domain of information in memory that is available for use by the metacognitive strategies in creating and interpreting discourse in language use” (Bachman and Palmer, 1994 I: Chapter 2, p. 21). They emphasize the important function of metacognitive strategies in the model of language use:

What makes language use possible is the integration of all of these components as language users create and interpret discourse in situationally appropriate ways. With respect to language testing, this conceptualization of strategic competence as metacognitive components provides an essential basis both for designing and developing potentially interactive test tasks and for evaluating the interactiveness of the test tasks in use (1994 I: Chapter 2, p. 24).

Authenticity, as a testing evaluation and development concept, has been a major issue in language testing over the past decade (see papers in Language Testing, 1985, 2 (1)). Progress is made in this area by the authenticity evaluation model presented by Bachman (1990, 1991), which allows judgments to be made about the degree of authenticity of test instruments. However, later Bachman and Palmer’s well-specified test qualities, authenticity and interactiveness, provide us with even more appropriate guidelines for a refined content analysis, which is a judgmental form of construct validity evidence. They rename Bachman’s ‘situational and interactional authenticity’ (1991) as the ‘authenticity’ and ‘interactiveness,’ in order to clarify the relationship among TLU, test tasks, and test takers.

Along with reliability (consistency of measurement), validity is an essential measurement quality, which refers to justifications for inferences made on the basis of test scores, which are used to make decisions. Validity “pertains to the meaningfulness
and appropriateness of the interpretations that we make on the basis of test scores" (Bachman and Palmer, 1994 I: Chapter 3, p. 4). Authenticity and interactiveness are placed within the validity studies of other measurement specialists. For example, authenticity is strongly associated with content validity, and “provides a basis for specifying the domain to which we want our score interpretations to generalize and hence, for investigating aspects of construct validity” (Bachman and Palmer, 1994 I: Chapter 3, p. 12-3). Bachman and Palmer explain interrelated features among test qualities as follows:

Authenticity, interactiveness, and impact are three qualities that many measurement specialists consider to be part of validity. We agree that authenticity and interactiveness are related to the construct validity and that impact is part of the consequential basis of test use. However, we believe that these qualities are important enough to the development and use of language tests to warrant separate consideration, and will discuss them as separate qualities (1994 I: Chapter 3, p. 6-7).

Measurement specialists use the term “construct’ to refer to the specific definition of an ability that provides the basis for a given test or test task. The term construct validity is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability, or construct, we want to measure. Construct validity also has to do with the domain of generalization, or the TLU situation to which our score interpretations generalize. At the very least we want our interpretations about language ability to generalize beyond the testing situation itself to a particular TLU situation. In some situations we may want to determine if a test that was designed for one situation or purpose is appropriate for use in a different situation or for different purpose. For these reasons, defining the characteristics of the TLU situation is essential, and it is here that construct validity relates to situational authenticity ... (1994 I: Chapter 3, p. 5).
The relationship between interactiveness and validity is a function of the relative involvement of areas of language knowledge, topical knowledge, and metacognitive strategies. That is, the extent to which high interactiveness corresponds to construct validity will depend on how we have defined the construct and the characteristics of the test takers (1994: Chapter 3, p. 13).

In addition to the previous test usefulness qualities, impact and practicality also play important roles in designing or evaluating a test. Impact is defined "broadly in terms of the various ways in which test use affects society, an education system and the individuals within these" (Bachman and Palmer, 1994: Chapter 3, p. 21). Practicality is defined "as the relationship between the resources that will be required in the design, development and use of the test and the resources that will be available for these activities" (Bachman and Palmer, 1994: Chapter 3, p. 18).

Through an experimental study, I will examine partial, not overall, evidence for the usefulness of a videotape formatted test (VFT) as a measure of academic lecture listening ability for placement and selection purposes. Impact and practicality are not covered in this study, because of the experimental nature of VFT. To provide evidence relevant to the further use of VFT as a measure of academic lecture listening ability for placement and selection testing purposes, I will investigate four aspects of test usefulness: reliability, validity, authenticity, and interactiveness. Table 1.1 illustrates their correspondances to other concepts of test qualities and how they will be investigated in this research.

Since Bachman and Palmer (1994) didn't specify the way the validity evidence can be used to interpret the observed VFT results in this study, I will refer to Bachman's 1990 explanation of validity. He defines validity as follows:
Table 1.1 Correspondences between Bachman and Palmers test usefulness, Bachman's validity and the test qualities considered in this study

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Validity</th>
<th>Authenticity</th>
<th>Interactiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Criterion relatedness</td>
<td>Meaningfulness of construct</td>
<td>Content relevance</td>
</tr>
<tr>
<td>Interrater reliability</td>
<td>Internal structure</td>
<td>Concurrent validity</td>
<td>Empirical item manipulation</td>
</tr>
<tr>
<td></td>
<td>Basis for construct validity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Bachman and Palmer's test usefulness qualities
b Bachman's three general types of validity evidence
c The way test qualities are evaluated in this study
d Basis for construct validity in this study

Judging the extent to which an interpretation or use of a given test score is valid thus requires the collection of evidence supporting the relationship between the test score and an interpretation or use ... The evidence that we collect in support of a particular test use can be grouped into three general types: content relevance, criterion relatedness, and meaningfulness of construct [as shown on line b in Table 1.1] (1990: 243).

In this study, content relevance will be investigated through a judgmental content analysis of a developed VFT on the basis of an authenticity and interactiveness analysis.Criterion relatedness and meaningfulness of construct will be examined through data from an experimental study. First of all, I will examine correlations of test scores between a conventional type of listening test and an experimentally designed VFT and audiotape formatted test (AFT). In this study, concurrent validity is not included as a means for
investigating construct validity, because correlated criterion behavior, as seen in the
TOEFL-type of conventional listening test, cannot be validly interpreted as an indicator of
the TLU task in this study. A second way of examining construct validity will be by
comparing theory-based item difficulty to empirical item performances in AFT and VFT.
In other words, I will modify test input conditions to investigate whether that the VFT
performance behaves in concert with theories involving TLU tasks. Thirdly, I will
investigate group and item performance of AFT and VFT group, which differ in their
degree of authenticity and interactiveness, through different channel & form of input
presentation and degree of pre-listening activities. Since the input format and pre-listening
activities play significant roles in the TLU tasks, experimental test designs also focus on
this facet of test method.

All three types of evidence are used to argue that the VFT measures a targeted
ability more effectively than audiotaped listening tests and thus can increase the validity of
an ESL listening test for selection and placement testing purposes.

The authenticity and interactiveness qualities are described in the following chapter
on the basis of judgemental analysis. Empirical analysis is used to examine reliability and
validity qualities to justify the VFT method as a good evaluation instrument for measuring
a targeted ability, academic lecture listening ability, for placement and selection purposes
in educational setting. The evidence for the usefulness of VFT is targeted by posing the
following research questions:

1) To what extent is a developed VFT similar to authentic and interactive
tasks?
2) Is the VFT reliable in terms of interrater reliability? Are test takers’ performances on the whole VFT items internally consistent with each other?

3) Do the VFT scores significantly correlate with scores on a conventional listening test?

4) Does empirical item analysis provide construct validity evidence for the VFT?

5) Is there a significant mean difference in the group performances between the VFT and the audiotape formatted test? Are the test takers’ scores normally distributed for placement and selection testing purposes? Are there significant item difficulty and item discrimination differences in the item performances between the VFT and AFT?

It is expected that the presentation of test usefulness qualities will provide future test developers and users with valuable procedures for evaluating new test methods. Because there have been few studies on the use of videotaped materials in ESL testing, it is hoped that information from this study may be useful in identifying areas of weakness or strength of videotaped formatted L2 listening tests for those who are responsible for designing and administering the L2 listening tests in academic settings.
CHAPTER 2. TEST DEVELOPMENT

In this chapter, four topics will be discussed as foundation for developing and evaluating a VFT (Appendix B). First, the academic lecture listening ability will be defined as target language use (TLU) situation in this study, and construct to be measured (academic lecture listening ability) will be defined. A theoretical foundation for the analysis of authentic and interactive test will be presented within Bachman and Palmer’s (1994) framework. In the next two parts, Bachman and Palmer’s ‘authenticity’ and ‘interactiveness’ will be used to evaluate the content evidence for validity of a designed VFT. To evaluate the degree of correspondence between the characteristics of TLU tasks and the test tasks, the proposed TLU situation is then compared with the VFT task in terms of authenticity. To evaluate interactiveness, the test tasks will be analyzed as they may be perceived by test takers. Finally, based on this theory of listening comprehension authenticity and interactiveness, the levels of comprehension difficulty for particular types of items will be hypothesized.

Academic Listening Tasks

New international students have to comprehend different types of academic lectures in class in order to carry out their academic goals successfully. The lectures can be classified according to several criteria. One criteria from the lecturer’s perspective is modes of lectures: reading, informal, and rhetorical styles (Dudley-Evans and Johns, 1981). Another criteria is the students’ roles in class. In seminar-like lectures, for
example, students can actively ask for repetition and repaired messages, using both verbal and nonverbal interruptions or backchannel cues, so that the input involving the active listeners differs from that in a class with passive listeners in a large auditorium.

The TLU context in this study consists of listening comprehension tasks in an academic lecture in teacher-centered class. The TLU tasks may vary according to different contexts, but the following are the important features of this rhetorical style:

1) Turn-taking and topic-processing are controlled completely by the lecturer. It is not easy for students to take the floor away from the lecturer, which means this context does not give much opportunity for interaction (backchannel, overlapping, or interruption) between students and a teacher. Clarification requests, negotiation for meaning, and repair of communication breakdowns rarely happen between students and a teacher in class.

2) Listening tasks are not directly negotiated between a teacher and students, but the teacher frequently tries to clarify his/her message, using multiple sources of input.

3) There is not much on-line help from the teacher, and, due to the rhetorical style, ungrammatical and incomplete expressions are frequently used. So students have to develop and use their strategic competence, in order to make up any comprehension deficiencies. Following are some strategies that can be used to increase comprehension: previewing textbooks
carefully, taking notes well, transferring their L1 knowledge, retrieving background information, using effectively nonverbal behaviors, expressive intonation or lexical intensifiers.

4) In this kind of context, many types of 'macro- and micro- markers' (Chaudron and Richards, 1986) play very important role in organizing the very extended discourse.

Since the listening comprehension process required in the TLU context involves a wide range of major components in the model of language use (i.e., topical knowledge, language knowledge, and metacognitive strategies), ESL students are asked to use a wide range of their language knowledge as well as topical knowledge and metacognitive strategies together, in order to make up their limited comprehension ability. In other words, to define the construct of listening comprehension, it is necessary to include all of the following.

**Topical knowledge**

Along with 'bottom-up' processing, the TLU task “requires substantial amount of ‘top-town’ processing in which the meaning is inferred from broad contextual clues and background knowledge” (Oxford, 1993: 207). The academic lecture listening tasks are basically supported and stimulated by a generalized topical knowledge because ESL students frequently need to transfer their L1 linguistic information or their built-in scripts into the listening tasks for making up their proficiency deficiency in the classroom.
They are expected to benefit from appropriate areas or levels of topical knowledge involving (non)language use frequently in academic lecture listening tasks. So only when listening tasks narrow down the topical content toward their commonly shared TLU situations, ESL students can be relaxed and encouraged to apply their topical knowledge to the listening tasks. If they are asked to comprehend very specific content, for example, advanced research on zoology, the topical knowledge cannot be equally engaged by the test task and they are likely to find it difficult to apply their topical knowledge to the listening task.

**Language knowledge**

Language knowledge is a part of the construct to be measured in the academic listening tasks and is thought of as a main domain supported by metacognitive strategies in comprehending academic discourses. All aspects of language knowledge (grammatical, textual, functional, sociolinguistic knowledge) are required to complete the academic listening tasks. Table 2.1 shows the areas of language knowledge.

**Organizational knowledge** In academic listening tasks, students are not expected to comprehend every line of a lecture, but, at least, they are required to comprehend main ideas of each lecture and transitions from one to another. In this context, discourse markers play very important role in carrying out their academic listening tasks as indicators of topic continuation - "with the functions of addition, comparison, contrast, exemplification, explanation, restatement, result, sequence, summation, and transition" (Chaudron and Richards, 1986: 116) within a lecture.
Table 2.1 Components of language knowledge  
(Bachman and Palmer, 1994 I: Chapter 2, p. 39)

<table>
<thead>
<tr>
<th>LANGUAGE KNOWLEDGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORGANIZATIONAL KNOWLEDGE</td>
</tr>
<tr>
<td>Grammatical Knowledge</td>
</tr>
<tr>
<td>- Knowledge of vocabulary</td>
</tr>
<tr>
<td>- Knowledge of morphology and syntax</td>
</tr>
<tr>
<td>- Knowledge of phonology and graphology</td>
</tr>
</tbody>
</table>

In a study of lecture comprehension problems encountered by Chinese students at UCLA, Yuan (1982) points out the importance of textual knowledge:

In general, the subjects were rather weak at paying attention to the sequence of the lecture because of their neglect of the logical connectors of sequence... Besides, they paid more attention to decoding the speech sentence by sentence than extracting the science information from the lecture through understanding the rhetorical nature and functions of both textual and lecture discourse (p.48).
The TLU tasks require substantial textual knowledge to organize a good amount of language input, for example, through macro- and micro-discourse markers (Chaudron and Richardson, 1986).

**Pragmatic knowledge** A relevant level of sociolinguistic knowledge (of natural, idiomatic expressions, or cultural references and figures) helps in understanding academic language input. The lack of sociolinguistic knowledge, however, can be compromised by the other language knowledge and compensatory strategies. For instance, even though a language input is sequenced in very informal and humorous way, the input is typically (always involving main idea) repaired or paraphrased to clarify the intended meaning in classroom.

**Metacognitive strategies**

Bachman and Palmer (1994 I) note that the conceptualization of strategic competence as metacognitive components provides an important basis for designing potentially interactive test tasks, and that metacognitive strategies are a very important part of construct to be measured in academic lecture listening tasks:

Metacognitive strategies are always in language use, and are thus always implied in our construct definitions of language ability. Furthermore, in our approach to language testing we attempt to create tasks that are interactive, so that these tasks will always engage the strategies. For this reason, the metacognitive strategies can always be assumed to be part of the construct (1994 I: Chapter 8, p. 11).

In each TLU situation, before ESL students go to a class, they already know the content to be covered in class, the teacher's teaching style, and their role in class. Before
they listen to a lecture, they are expected, or sometimes required, to establish their particular communicative goals in given contexts, such as understanding a particular style psycholinguistics lecture which covers specific research work one week before the midterm examination.

In the TLU tasks, they are encouraged to take advantage of time and materials to prepare for each class, which means their listening tasks begin with well-prepared goal-settings and assessment strategies. During a lecture, a variety of contextual clues, aurally and visually, help them assess the characteristics of an academic listening task in assessment stage. Since they have related their topical and language knowledge to similar contexts in academic settings over and over, they can benefit from their built-in assessment strategies about the characteristics of commonly shared academic listening tasks or their own level and relevance of topical and language knowledge in each context.

So, planning is the most important strategy to be measured for successful completion of the academic lecture listening tasks. The hard part of the listening comprehension process in the TLU tasks is that students have to decide how to utilize relevant sources from language knowledge and topical knowledge together, to make up for their limited comprehension level and for successfully completing the listening tasks as effectively and quickly as they can. At that time, they need to effectively ‘bottom-up’ and ‘top-down’ process their target/native language and topical knowledge, using verbal and nonverbal input in the context.
Consideration of Authenticity

As Bachman and Palmer (1994 II) argue, characterizing authenticity needs to be the first and central consideration in language test design. In designing an authentic test task, critical aspects of TLU task should be identified and then compared to the test task. Bachman’s test method facet framework (1990) or Bachman and Palmer’s framework of task characteristics (1994) can be effectively used to analyze the degree of authenticity involving the academic lecture listening test tasks and the TLU.

One way to measure authenticity is to measure the correspondence between the characteristics of TLU task and the test task. To do this, Bachman’s framework (1990) has been useful as an evaluation tool. Hoekje and Linnell (1994) evaluated the following three spoken language tests for international teaching assistants by examining Bachman’s test method facets: SPEAK (Spoken Proficiency English Assessment Kit) test, OPI (Oral Proficiency Interview), and a performance test.

In this study, a VFT was designed to increase authenticity over the conventional listening test (Appendix B - scripts and test booklet), and this test task is expected to involve test takers’ performance in very similar ways that they are engaged in the TLU task. In order to provide a basis for construct validity of the VFT, the degree of authenticity will be analyzed using the following facets: environment, rubric, input, expected output, and the relationship between input and output. Table 2.2 shows the categories of test method facet considered in this section.
Table 2.2 Categories of test method facets (Bachman, 1990:119)*

1. Facets of the Testing Environment
   - Familiarity of the place and equipment
   - Time of testing
   - Physical conditions

2. Facets of the Testing Rubric
   - Time allocation
   - Instructions

3. Facets of the Input
   **Format**
   - Channel of text presentation (aural, visual)
   - Form of text presentation (language, nonlanguage)
   - Language of presentation (native, target, both)
   - Length of text
   - Speededness
   - Vehicle of text presentation (live, reproduced, both)
   **Nature of Language**
   - Organizational characteristics
   - Pragmatic characteristics

4. Facets of the Expected Output
   **Format**
   - Channel (aural, visual)
   - Form (language, nonlanguage, both)
   - Language (native, target, both)
   - Type (selected, limited production, extended production)

5. Relationship between Input and Output
   - Reactivity (reciprocal, nonreciprocal)
   - Scope of relationship (broad, narrow)
   - Directness of relationship (direct, indirect)

* Within this framework, I partially introduced Bachman and Palmer's terminology (1994 I:Chapter 2, p. 10-17). Their framework of task characteristics has very slightly different points from Bachman's, but I have combined the two in my study.

By modifying facets of the input format, a AFT (Appendix C - scripts and test booklet) was also experimentally designed to provide validity evidence for the VFT method. The only difference between the VFT and the AFT will be in facets of the input format.
Facets of the environment

Test takers process their comprehension tasks under very similar environmental conditions as in the TLU. Since the tests asks them to perform in their regular class during 'real' lectures, the following facets are not expected to negatively affect their listening performances across the two conditions (TLU and test): familiarity of the place and equipment, time of listening tasks asked, physical conditions.

Facets of the rubric

Even though test takers are expected to proceed through the VFT tasks in a similar way to that in the TLU tasks, the amount of time allocated for comprehension processing in the VFT may affect the test takers' performance. Students in the TLU context are typically asked to keep working on a lecture for about 50 minutes, watching and taking notes together. They have enough time and materials to make adjustments during their tasks.

Test takers in the VFT, on the other hand, have to finish a lecture listening activity within about 15 minutes: reading background information for one minute, watching one of three different lectures and taking notes for about nine minutes, and then writing down answers from open-ended questions for six minutes. Then, they quickly move on to background information of the next lecture, before they watch the next lecture. Test takers' performances in the AFT are also affected in a similar way by the amount of time.

Test instructions are emphasized in the beginning of test administration. Just as students in the TLU context already know what to do before they go to classrooms, test
takers are given instructions on the procedures to be followed, and the nature of the tasks. Also, since they are told that partial points can be given for each response, the test takers' knowledge of this can be expected to positively affect their test performances.

Facets of the input

As seen in Table 2.3, even though the VFT excludes live interaction between a teacher and students, the task includes most input format characteristics of the TLU tasks. In the input format, the channel and form of text presentation are significant differences for the VFT and the AFT.

Table 2.3 Characteristics of input format

<table>
<thead>
<tr>
<th></th>
<th>TLU TASKS</th>
<th>VFT TASKS</th>
<th>AFT TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel of text</td>
<td>Aural &amp; Visual</td>
<td>Aural &amp; Visual</td>
<td>Aural</td>
</tr>
<tr>
<td>presentation*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form of text</td>
<td>Language &amp;</td>
<td>Language &amp;</td>
<td>Language</td>
</tr>
<tr>
<td>presentation</td>
<td>Nonlanguage</td>
<td>Nonlanguage</td>
<td></td>
</tr>
<tr>
<td>Language of</td>
<td>Target language for</td>
<td>Target language for</td>
<td>Target language for</td>
</tr>
<tr>
<td>presentation</td>
<td>academic purposes</td>
<td>academic purposes</td>
<td>academic purposes</td>
</tr>
<tr>
<td>Length of text</td>
<td>Extended discourse</td>
<td>Extended discourse</td>
<td>Extended discourse</td>
</tr>
<tr>
<td>Speededness</td>
<td>Perceived as very</td>
<td>Perceived as very</td>
<td>Perceived as very</td>
</tr>
<tr>
<td></td>
<td>speeded</td>
<td>speeded</td>
<td>speeded</td>
</tr>
<tr>
<td>Vehicle of text</td>
<td>Live</td>
<td>Videotaped live</td>
<td>Audiotaped</td>
</tr>
<tr>
<td>presentation</td>
<td></td>
<td></td>
<td>reproduced</td>
</tr>
</tbody>
</table>

* In both VFT and AFT questions are visually presented in printed forms.
The TLU situation and the VFT include nonlanguage materials: notes on board, kinesic behavior including facial expression. The kinesic behavior is believed to help reduce ambiguity in language (Kelleman, 1992) and facilitate communication by increasing redundancy (Birdwhistell, 1971) in comprehending extended discourses, together with prosodic and paralinguistic features.

In the TLU task, speed of processing plays a very important role in comprehending quite a lot of (non)language input. Because students are of limited target language proficiency, they frequently must infer the meaning of difficult input from contextual clues, assess, and adjust the meanings later, in order to complete a comprehension task as quickly as possible, before another task comes next.

But the VFT task input can be perceived as much more speeded than the TLU one, because of the limited amount of test input. Also the AFT task is more speeded than the VFT one. The perception of speededness makes test takers use many types of compensatory test taking strategies in their completion of the test tasks (Bachman, 1990: 128), like in the TLU tasks.

Facets of language input

Organizational characteristics Test takers' ability to use the marked relationships among utterances and organizational development in extended discourse is emphasized in designing authentic tests. The following is a typical test task involving emphasized textual knowledge:
Test takers listen to:
...
Ah, she starts with perhaps the most obvious need for translation between different languages. Mora is from El Paso which is on the border of Mexico and the United States, so I think it's not surprising that she takes her example of translating between Spanish and English. Now we know when we from one language to another, that some problems may arise. Some words are not even translatable. Perhaps you know a word in a language and there is no equivalent in another language. Perhaps there is kind of any equivalent, but maybe the connotation is different.

Let's look at that word, 'connotations'. (writing down connotations on board) Connotations refers to the feelings and associations with a particular word. For example, let's say someone, my height, weighs 200 pounds. There are a lot of different ways we might describe such an individual. And perhaps our attitude toward that individual would determine the connotation of the word we would describe that individual with. For instance, we might say, “she's kind of stocky.” Or we might say, “she's fat.” Or we might say, “she is obese.” Or “she is a little bit overweight.”

Ah, I'll describe the same individual, the word choice, the different connotation reflects the attitudes toward that individual. And when we translate from one language to another, I think we're quite conscious of the fact that the connotations of words may not be identical...

Test takers are asked to answer following questions involving the discourse above:

Question 1. What is the first example of communication problems that Mora's poem describes?

Question 4. It is not surprising that Mora uses the example of translating between English and Spanish, why?

Question 5. Why does the lecturer give some examples of words meaning "fat"?

Pragmatic characteristics Since test takers are allowed to read specified instructions (how to do in the task) and background information (who is talking to whom about what) before listening activities, they can easily take advantage of their pragmatic knowledge in the test.
Facets of expected output format

Characteristics of output format in the TLU and VFT tasks are shown in Table 2.4. While test takers are allowed only to write down language forms, in the TLU task the expected response channel can be oral (sometimes responding by speaking) and the forms consist of language and nonlanguage materials (pictures, charts, or physical actions).

There are two types of response in the TLU and the test task: limited production (word level and phrase level) and extended production response. The following in the next page include examples of each type of response in the VFT:

<table>
<thead>
<tr>
<th>Table 2.4 Characteristics of output format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TLU TASKS</strong></td>
</tr>
<tr>
<td>Channel</td>
</tr>
<tr>
<td>Form</td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>Type</td>
</tr>
</tbody>
</table>
Example 1: Limited production response: word-level

Test takers listen to:

...And one of the first things the white masters learned to do to control these people more effectively was to strip away their cultures. They could not, however, strip away their desire to sing, their desire to communicate.

So, although their native languages, their native songs, their native religions were banned, they were not prohibited from singing...

They read and answer:

Question 23. White slave masters tried to strip away the slaves' native cultures, but what did they fail to ban in United States? (The response to this would be a word such as 'singing'.)

Example 2: Limited production response: phrase-level

Test takers listen to:

... So everytime you convert them, the grass into the steer, or the steer into the human, you lose 90% of energy (circling 90% on board). Now why is that? The reason is because, ah, most of the energy goes into just maintaining the body. You know the cow doesn't just sit there, and even just sitting there, its heart beating, and it's, gotta produce heat to maintain its body, and it's running around swatting flies and chasing other cows, it's gonna lose a whole lot of energy, right? O.K. So most of the energy then in the, in the, ah, food it eats goes into maintaining the body. So only 10% is actually stored as weight...

They read and answer:

Question 20. What is most of the lost energy used for? (The response would be a phrase such as 'to maintain body'.)

Example 3: Extended production response

Test takers listen to:

... Another aspect in the lives in which music developed, of course, was religion. As it had in their homes and most of their, their religions, most of people were prohibited from, ah, from exercising their native religions. So they had to exercise their religions that they are allowed to, which includes gospel music. So they were
learning American gospel hymns and, of course, they would eventually take those melodies and take the rhythms, and fuse them together and take the songs, as the gospel songs, and change the words, change the direction. What do gospels concern themselves with? God, religious aspects (writing down ‘God’). Well, what would happen in the development of the, of the, this early music which we eventually began calling blues (X marking ‘God’ on board)? Instead of looking at God, they looked at their daily concerns. Woman, singing about a woman, man, child, romance, troubles, problems (writing down ‘WOMAN MAN CHILD PROBLEMS’). If you were a slave in the 19th century, you wouldn’t wanna be singing about life at school, you’d wanna be singing about problems in your life, you’d be singing about your blues. These came together and we began to call it the blues (writing down ‘BLUES’ on board). B, L, if I can write this, co, correctly, U, E, S, blues...

They read and answer:

Question 25: Explain blues, the name for the 19th Century music, comparing with gospel hymns. (The response would be ‘Gospels concern themselves with religious aspects and blues are interested in their daily concerns-woman, man, child, romance, troubles, problems.’)

Relationship between input and output

Even though the VFT is designed to introduce many characteristics of the TLU context, the test task is still an artificial condition. Table 2.5 shows different relationship between input and response in TLU, VFT, and AFT tasks.
Table 2.5  Relationship between input and response

<table>
<thead>
<tr>
<th></th>
<th>TLU TASKS</th>
<th>VFT TASKS</th>
<th>AFT TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reactivity</td>
<td>Reciprocal</td>
<td>Non-reciprocal</td>
<td>Non-reciprocal</td>
</tr>
<tr>
<td>Scope of relationship</td>
<td>Broad &amp; narrow</td>
<td>Broad &amp; narrow</td>
<td>Broad &amp; Narrow</td>
</tr>
</tbody>
</table>

One of the big differences between two tasks is the extent to which input and response directly affect the following input and responses. Test takers are not allowed to interact with a teacher in the VFT context and their responses cannot affect the form of the following input. In some teacher-centered academic lectures, students benefit from feedback on their verbal or nonverbal responses. For example, when a teacher explains a part of lecture but he/she sees confused expressions on the students, this nonverbal response may result in the teacher repairing the message.

Another difference is the degree to which the response is primarily based on information in the input. The VFT task requires test takers to understand only the content of lectures explicitly asked for in the questions, while the TLU task responses sometimes ask for more than information included in scripts.

Finally, the amount of input varies. Bachman and Palmer define ‘broad scope’ input as tasks in which test takers must process a lot of input, and ‘narrow scope’ as that
involving a limited amount (1994 I: Chapter 2, p. 17). Like in the TLU context, test
takers are asked to understand a specific detail from a relatively small amount, narrow
input, or a main idea from an extended discourse, broad input.

**Consideration of Interactiveness**

The essence of interactiveness is associated with the involvement of the major
components in the model of language use: topical knowledge, language knowledge,
language function, metacognitive strategies, and affective schemata. I will examine the
extent to which each component is most likely involved in the performance of the test
tasks. The degree of interactiveness will be used to predict level of item difficulty for
empirical item analysis. Table 2.6 summarizes the points involving interactive tasks.

**Involvement of the test takers' topical knowledge**

Test tasks in the VFT require test takers to comprehend very introductory lectures
in Humanities, Biology, and Music, so they are expected to benefit, in the test task, from
their previous acquired knowledge of these topics which has been established in their
countries or in the TLU situations. Also, like in TLU tasks, they have a chance to read
background information for each topic (see example below) before beginning listening
activities, so that, with this pre-listening activity, the authentic format of the test input
(e.g., aural and visual channel) helps them assess and retrieve their appropriate topical
knowledge quickly and easily.
<table>
<thead>
<tr>
<th>EACH MAJOR COMPONENT</th>
<th>FACTORS INVOLVING THE COMPONENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topical knowledge</td>
<td>• Pre-listening activity (reading background information)</td>
</tr>
<tr>
<td></td>
<td>• Commonly shared topical content (introductory lectures)</td>
</tr>
<tr>
<td></td>
<td>• Visual information</td>
</tr>
<tr>
<td>Language knowledge</td>
<td>• Authentically designed &quot;live&quot; language input (not pre-controlled by scripts)</td>
</tr>
<tr>
<td></td>
<td>• Visual information</td>
</tr>
<tr>
<td></td>
<td>• Open-ended question format</td>
</tr>
<tr>
<td>Language function</td>
<td>• Note-taking activities for completion of the test tasks</td>
</tr>
<tr>
<td></td>
<td>• Open-ended question format</td>
</tr>
<tr>
<td></td>
<td>• Visual information</td>
</tr>
<tr>
<td>Metacognitive strategies</td>
<td>• Pre-listening activity</td>
</tr>
<tr>
<td></td>
<td>• Visual information</td>
</tr>
<tr>
<td></td>
<td>• Authentic discourse</td>
</tr>
<tr>
<td>Affective schemata</td>
<td>• Enough time and material for preparation of the test tasks</td>
</tr>
<tr>
<td></td>
<td>• Visual information</td>
</tr>
<tr>
<td></td>
<td>• Authentic discourse</td>
</tr>
</tbody>
</table>
Test takers in the VFT are asked to have an appropriate area and level of topical knowledge involving authentic academic lectures to carry out their academic goals successfully, so they are encouraged to apply their own topical knowledge to the topical content of test input to a similar degree as in their TLU situations. When the characteristics of test input are not provided in the AFT, test takers should find it difficult to apply their topical knowledge to test task.

Example: Printed background information in the VFT

Read: In a music class, a lecturer is going to talk about the development of a music which is uniquely American, and which is influenced by many cultures from around world, both in its beginning and today. That music is jazz. In the beginning of class, some important elements of jazz were listed. The four most important elements of jazz are improvisation, rhythm, lifestyle and composition involving performance. Whereas classical music is the composer’s art, jazz is the performer’s art.

Now the lecturer will talk about origins of jazz. Watch the following eight minute lecture, taking notes as consciously as you can. Then, answer the following questions by referring to your notes. You are allowed to take notes in this page.

Involvement of the test takers' language knowledge

Since the VFT consists of three ‘real’ lectures in classroom taught by ‘real’ professors, test takers are expected to use a wide range of language knowledge in processing test input and formulating responses from open-ended questions, in the almost same way as in the TLU tasks.

Visual information (nonverbal language input of lecturer, printed background information on test booklet, written information on board) can help test takers clarify aurally delivered language input and thus, effectively involve language knowledge in test tasks. For example, when lexical, syntactic, and prosodic hedged messages are added to
the lecturer's facial expressions, test takers can much more easily benefit from their pragmatic knowledge to comprehend the message than in AFT.

Involvement of language functions in the test tasks

Bachman and Palmer note:

If our objective is to design language test tasks that correspond to non-test language use, then test tasks must incorporate the goal-directed, purposed nature of language as communication, which means that they must involve the test takers in functions other than simply demonstrating his knowledge of language” (1994 II: Chapter 9, p. 8).

In a VFT context, some language functions, other than the simple demonstration of language knowledge, are strongly involved as they process the test input and formulate a response: For example, like in the TLU task, test takers are asked to take notes that are clearly organized and selective while they are watching a lecture. It is recommended that they, in their notes, include all the main points and enough specific details to use in answering questions later. After each of the lectures, they can use their notes on problem-solving tasks.

In the VFT, open-ended questions involve different types of tasks: finding supporting information or main idea in extended discourses. Moreover, the expected comprehension tasks are divided into those involving responses at word-level, phrase-level, and extended discourse-level. Since all the tasks need to be based on well-organized and goal-directed note-taking activities, the VFT task involves TLU functions of language for successfully completing their listening tasks.

Visual information helps test takers during note-taking. For example, teachers put
down words or phrases, when emphasizing them or when they want to present the way their lectures are developed. The written information in the VFT context is very useful for note-taking. Additionally, the VFT provides test takers with enough time to make notes. When teachers write on the board, they normally stop to send aural messages, so test takers can simply follow the writings at that time.

Because of limited contextual clues presented aurally, the AFT task is perceived as much more speeded than the VFT, which means many test takers may give up note-taking strategies. It is not easy to listen to what comes next, at the same time to take notes about what they heard, only from aural channel.

**Involvement of the test takers' metacognitive strategies**

In conventional listening tests (e.g., in TOEFL), goal-setting and assessment strategies are overemphasized in comprehending listening test tasks. Test takers make the most effective use of reading multiple choices and guessing a plausible context before listening to a lecture from cassette player. They need to retrieve their top-down processing from the content of multiple choices. And if they are not good at test-wiseness, they simply have to depend on bottom-up processing based only on the delivered linguistic input. Additionally, some questions in the tests ask for how well test takers can accomplish goal-setting and assessment involving listening tasks (e.g., Who is the speaker? Where does this conversation most likely take place?).

Since the experimentally designed AFT in this study doesn't allow test takers to do any pre-listening reading, and since they don't have enough environmental clues from
cassette recorder to involve a wide range of metacognitive strategies, the AFT tasks involve the metacognitive strategies to a very different degree and direction as in the TLU.

In the VFT, through reading background information before listening tasks and having a variety of contextual clues (from authentic discourse) during listening activities, they can easily set up their goals and assessment better than in the AFT. As planning is the most challenging stage for the successful completion of the TLU task, the VFT task requires that the strategies be strongly associated with planning stage. After reducing their burden of goal-setting and assessment with a pre-listening activity, the VFT task enables test takers to comprehend an academic lecture from a wide range of and a great amount of aural and visual information during the listening task. Much more opportunity for metacognitive strategies involvement is provided in the VFT.

Involvement of test takers' affective schemata

Many audiotaped listening tests evoke negative affective responses that make it difficult for test takers to perform at their best, because the input of the test task contains many negative factors to affect their listening tasks, such as irrelevant topical content, pretty dense organization, simple channel/form of input, poor make-up clues for their limited proficiency level.

In more authentic test tasks close to the TLU tasks, we would expect a positive affective response to make the test task relatively more interactive. It is expected that the test takers’ assessment of the characteristics of the VFT input will positively influence
their affective response to the listening task and their flexibility in using their language ability" (Bachman and Palmer, 1994 II: Chapter 9, p. 9), so that the positive affective response to the rich, authentic channel/form of input makes the VFT task relatively more interactive than in the AFT.

**Level of Comprehension Difficulty**

The prediction of item difficulty based on content analysis can provide a principled basis for empirical item analysis, which can be a type of construct validity evidence for a test. Also, the judgmental content analysis can be supported by the empirical item analysis (Chapelle, 1994: 171).

In this study, the ‘systematic content analysis’ of authenticity and interactiveness also has implications for the prediction of item difficulty in the test tasks. When these theoretically predicted difficulties are compared to the empirically obtained ones, we can use the results as a form of construct validity evidence for the experimental test.

When the item means within hypothesized item difficulty categories fit theoretical expectations, or when there is high correlation between individual item difficulties and predicted difficulty levels, the results can be interpreted to provide construct validity evidence. For empirical item analysis in the next chapter, all items are judgmentally placed into levels of comprehension difficulty based on analysis of the test tasks. Both degree of interactiveness and amount of the requested comprehension task explained in the previous sections are predictors of item difficulty (Table 2.7).
Test takers are expected to perform best in a highly interactive listening task when they are asked to comprehend small amount of information (e.g., word-level response). On the other hand, when they have to comprehend an extended discourse (e.g., summarizing an event) based on low interactive listening task, the observed score is predicted to be the lowest (i.e., 2).

Table 2.7 Level of Comprehension Difficulty

<table>
<thead>
<tr>
<th>Degree of Interactiveness</th>
<th>Low</th>
<th>Middle</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of response:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended comprehension</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>response:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentence-level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>limited comprehension</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>response:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phrase-level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited comprehension</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>response:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word-level</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted level: from 6 (easiest items) to 2 (most difficult items)
The TLU task in this study is also significantly affected by two factors: the amount of information which they have to comprehend and the extent (degree) to which the major components (topical knowledge, language knowledge, metacognitive strategies, and affective schemata) are involved in listening comprehension tasks. Because degree of interactiveness is the primary factor to predict task difficulty in the TLU and test context, the empirical item performances can be usefully compared to the predicted difficulty to provide evidence for the meaningfulness of interactiveness in the construct to be measured.

The VFI task is generally judged to pertain to a higher degree of interactiveness than the AFT, as discussed in the previous section. However, questions in the VFI vary with respect to how they require test takers to involve the major components in the performance of test tasks. Three examples below are presented to explain how three variations allow judgments to be made on different levels of item difficulty. Printed background information and scripts directly involving the following comprehension question are presented in each task. Difficulty level 6 is the easiest item and difficulty level 2 is the most difficult.

Example 1: one of high interactive comprehension tasks with limited comprehension response - Difficulty level: 6 (Question 13)

Background information: ... Now a lecturer will talk about food chains and how they work. In order to do that, she will give you some background information on some of the components of the ecosystems in the beginning of her lecture...

Script: ... The next group are the heterotrophs, (writing down ‘2) heterotrophs’ on board) O.K.? Now who knows what hetero means? (A student’s answer: “Variety.”) Variety, right, or mixed, and homo means same, like homogeneous crowd, ah, hetero means mixed. What we have is mixed feeders. Here we already
know that trophs (circling on board) means feeders. So we have mixed feeders and these include all the organisms that directly and indirectly are dependent on the work of the autotrophs for their food supply (pointing to ‘autotrophs’ on board). So this would include all the animals. When you think about what you eat, we’re animals, and when, when you think about what you eat, it’s either, dir, plant life directly or it’s something that was converted from, pla, plant life, right? Like if you eat a piece of chicken, the chicken came from eating corn, and the corn came from a plant, right? So everything we eat directly or indirectly comes from plant life, and that’s true with all the heterotrophs on the plant (pointing to ‘heterotrophs’ on board), which includes all of the animals and all of the fungi basically, O.K.?

... 

Question 13: Define heterotrophs, and give an example.

When they comprehend a discourse involving question 13 above, the discourse allows them to benefit from active involvement of major components in carrying out the test task. Question 13 is predicted as one of the easiest items in the VFT because it involves the major components and then a small amount is requested in the comprehension task (possible correct response: mixed feeders, animal). This information is repaired and repeated through the whole discourse, thereby reducing test takers’ burden in the comprehension task in the following ways:

1) Rich sources to retrieve topical knowledge are presented, from printed background information before listening actives (e.g., ‘... some of the components of the ecosystems ...’), and well-known content on food chain (e.g., relationship among ‘plants and animals’).

2) A wide range of language knowledge is used to comprehend, the spoken lecture, written words on board, written comprehension questions on test booklet, and informal style of speech (e.g., ‘that was converted from, pla, plant life, right?’)
3) Note-taking is allowed to process effectively the test input (very extended discourse) and formulate a response (from open-ended questions), and, as a result, the test task involves test takers in functions other than simply showing their knowledge of the language.

4) Many opportunities for metacognitive strategy involvement is provided in the test task because test takers have enough time to prepare (from pre-listening activities and extended discourse during the listening task) and authentic materials with which to plan (e.g., from repaired and repeated information).

5) The positive affective response to the rich contextual clues (e.g., both visually and aurally) is expected to make test takers feel relaxed to perform at their best.

Example 2: one of middle interactive comprehension tasks with extended comprehension response - Difficulty level: 3 (Question 10)

Background information: ... Pat Mora’s poem “Borders” is about communication problems. Now she will explain about the three aspects of the poem ...

Script: ... That doesn’t do all the communicating. O.K. We have to go beyond borders. Also borders can help define who we are. A border says, “This is me” and “This is everyone else” “This is the rest of the world.” Perhaps in some cultures, border definition is quite different. In America, we emphasize the ‘self’, the individual, “Give me my space.” But in some cultures, there’s much more emphasis on fitting in with the group, subordinating the self for the good for the community. I think being aware of that kind of personal border can help to enhance the possibility for appropriate communication ...

Question 10: When we are aware of borders in communication, what is (can be) a positive result of borders in real world?
In the performance involving question 10, less involvement of the major components is expected than question 13. For example, printed background information does not imply the discourse involving question 10, and the use of figurative terms (e.g., definition of 'border') in humanities makes test takers, most of whom majors in science, engineering, and business, find it difficult to retrieve their topical knowledge. With a response of an extended discourse (sentence-level), the question 10 is predicted as more difficult item level than the question 13.

Example 3: one of low interactive comprehension tasks with extended comprehension response - Difficulty level : 2 (Question 27)

Background information: ... Now the lecturer will talk about origins of jazz ...

Script: ... Later, after World War I, there was a massive migration of black people, north to industrialized cities to find work. As they were becoming more sophisticated and more entered into the mainstream of American society, through this move north ...

Question 27: According to the speaker, after World War I, what happened to the black people?

Because of the very limited test input in question 27, the major components in the model of language use cannot be effectively called into the test performance. Since more than one sentence has to be produced in their test output, this is predicted as one of most difficult items.

When they are asked to understand a similar amount of sentence-level information and they perform differently under different degree of interactiveness (for example, in question 10 and 27), the item difficulty index can be an useful source for a construct validity evidence involving the degree of interactiveness.
The construct to be measured in this study includes aspects of interactiveness as well as of authenticity, as discussed in this chapter. Since the features of interactive test tasks are part of the construct required for the TLU tasks, construct validity is supposed to correspond to the item performance involving the major components in the model of language use. So, empirical item analysis (using item difficulty) based on the interactiveness perspectives can be good evidence for construct validity of the test.

In the Chapter 2, it was found that the VFT task contains relatively more authentic and more interactive factors than the AFT task, relative to the TLU. The VFT task can accommodate much more characteristics of the TLU tasks than the AFT, and much more involvement of the major components is expected between the VFT task and test takers. Pre-listening activity and visual information are the key to the different degree of interactiveness between VFT and AFT.

Because the TLU task requires L2 students to make up for their limited comprehension ability through a wide range of major components of language use, the interactive tasks play an important part in construct to be measured. So it is expected that item difficulty level on the interactiveness perspectives provides a useful basis for a construct validity study of the developed VFT.
CHAPTER 3. METHOD

In addition to the judgmental content analysis in terms of authenticity and interactivity, validity and reliability studies will be discussed from experimental research in Chapters 3 and 4. Both the experimentally designed VFT and AFT, which were discussed in Chapter 2, will be used to provide additional evidence for the usefulness of VFT. This chapter discusses four aspects of the research study: The subjects, the test materials, the procedure, and the analysis.

Subjects

The subjects of this study were 83 international students in Iowa State University's seven English 101 courses (two sections of 101E: ESL listening for both undergraduates and graduates; four sections of 101C: ESL writing for undergraduates; one section of 101D: ESL writing for graduates) during the spring semester of 1995 (43 in VFT group and 40 in AFT group). The English 101 series are required ESL courses for graduate and undergraduate students who have passed the Test of English as a Foreign Language (TOEFL) with at least a score of 500, but whose English ability is not strong enough to pass ISU's ESL Placement Test. The test takers were predominantly Asian, with 24% from the People's Republic of China; 17% from Indonesia; 15% from Korea; 13% from Malaysia; 10% from Taiwan; 4% each from Japan and Pakistan; 2% from Hong Kong and Saudi Arabia. The remaining 11% of sample consisted of one or two students each from Puerto Rico, Turkey, Argentina, Russia, and Honduras. They had a wide variety of
majors, with 34% in engineering field; 24% in Business; and, 20% in Sciences. The population also included undergraduates (60%), graduates (34%), and unclassified persons (2%). All these students are enrolled with native speakers of English in other classes that are related to their fields of study. The research in this study was approved by the Iowa State University Human Subject Research Committee.

Test Material

Eighteen four-option, multiple-choice listening comprehension questions were used to compare two group performances in a conventional test context. Four audiotaped academic lectures (about Linguistics, Biology, History, and Food Science) made up a conventional listening test for all subjects to divide them into the two groups (VFT and AFT contexts). And dichotomous (right/wrong) scoring method was used in this test. The scripts and questions are presented in Appendix A. The lectures and test questions were from a TOEFL preparation book, "Barren’s HOW TO PREPARE FOR THE TOEFL" (1986 edition).

Reliability and validity evidence for VFT method were based on experimental test designs (VFT and AFT), as discussed in Chapter 2. In order to satisfy both contexts, differing in degree of authenticity and interactiveness, very authentic lectures were needed for VFT context. After reviewing commercially and privately produced videotaped lectures in terms of authenticity and interactiveness, three videotaped academic lectures (about Humanity, Biology, and Music) from "The Real Thing" (1991),
produced by HighLand publishing, were chosen. The authenticity level and tape quality of each lecture was judged to be higher than any other videotaped lecture materials available.

In the VFT context, test takers were asked to read printed background information before they watched the videotaped lectures in order to help them quickly retrieve their nontest knowledge and set up their assessment strategies. As they watched the lectures on video, they could benefit from the lecturer's visual information - facial expressions, gestures, notes on the board.

In the AFT context, test takers were asked to listen to these same lectures on a cassette recorder. The printed scripts in the VFT were orally presented in the beginning of the lectures. Both groups were strongly recommended to take notes on allowed pages as they were listening to and watching each lecture. After each of the lectures, test takers in both groups were asked to read the same open-ended questions about them and write down answers. Partial credit scoring method (scaling 0 to 3) was used in the experimental tests.

After transcribing the original lectures in the VFT, the scripts were modified and then recorded for AFT context. In order to minimize the effect of modification in test performances between the two groups, small segments of the language input were carefully modified. The major rules of modification were:

1) Long silences are avoided in the AFT. The relatively overlong pauses were frequent in the VFT, like in the TLU tasks, when teachers write down on the board, when they are expecting students' involvement, and so on.
However, like in conventional listening tests (e.g., TOEFL), the lectures in the AFT are tightly latched.

2) Disfluency markers such as false starts, repetitions” (Hatch, 1992) are omitted in the AFT.

3) “Preclosing and pre-opening signals” (e.g., ‘well’, ‘okay’, ‘right’) are omitted in the AFT, when they are habitually frequently repeated.

4) Ungrammatical points are corrected in the AFT. Though grammatical, some sentences are modified to make them more interpretable in the AFT context.

It might be expected that the modifications affect differences in test takers’ performances between VFT and AFT. For example, the frequent overlong pauses may provide test takers in the VFT with more involvement of metacognitive strategies and affective schemata. However, since very small amounts of content were made on the modifications, it is assumed in this study that the script modifications do not affect test performances between two groups. Both scripts are presented in Appendix A and B. One of example modifications given in the second lecture is illustrated below, and parts of original script, which were changed on the AFT script, are italicized:

**Example of original script used in the VFT**

... Now that’s probably a foreign word, but one of the things I like to do with students is, *ah, have* them translate biology because often times biology *is a lang*, is a language in and of itself. And if we take this, *this*, word apart, what we find is ‘auto’ here, *refers* to ‘self’, *right*? Like *automobile*, *automobile means* ‘self moving’, *right*? . . .
Example of modified script used in the AFT

...Now that's probably a foreign word, but one of the things I like to do with students is to have them translate biology, because often times biology is a language in and of itself. If we take this word apart, what we find is 'auto' here, referring to 'self' like automobile meaning 'self-moving'...

As explained in Chapter 2, when test takers are allowed to take notes in listening test tasks, more involvement of language functions is expected. Additionally, in answering open-ended questions, test tasks should be based on well-organized and goal-directed note-taking activities. And thus, in order to encourage test takers in their note-taking activities, two versions of the experimental test were made up of 31 open-ended listening comprehension questions, most of which require test takers to figure out the main idea from an extended discourse, like in the TLU tasks. The following discourse and a requested comprehension question in the second lecture provides a typical example of the test tasks in experimental test contexts.

In a Biology lecture, test takers in the VFT listen to:

... O.K. Then a third group are the decomposers (writing down ‘3) decomposers’ on board). Because you see in any ecosystem, autotrophs and heterotrophs are constantly manufacturing waste products. Also, they die. So what do we do with all of the waste and all of the dead bodies? And it turns out that these dead bodies and waste products are full of nutrients. You know that yourself, if you put manure in a garden it makes your plants really thrive. So how do we get the nutrients in those dead bodies back into the ecosystem in a form that can be used to generate new life? And, the answer to that question is the decomposers (pointing to ‘decomposers’ on board) do the work. And the decomposers consist primarily of two groups of organisms. These are the
bacteria and the fungi. (writing down 'bacteria' and 'fungi' on board) O.K.? And the bacteria and fungi take waste products that these organisms produce, they take the dead bodies, they pull all of the important ingredients out of those bodies, either put them back in the environment so that they can be used by plants ((pointing to 'plants' on board)) to make more food, or else put into the soils. So the plants that are existing can absorb the materials through their roots and create new plant growth. So anyway you slice it, the decomposers are providing the autotrophs with more of the things they need to continue producing food for the whole ecosystem, O.K. ...

Test takers are asked to answer the following question based on the discourse above:

**Question 15.** State one important thing decomposers do.

All parts of the conventional and experimental tests were pilot-tested with native speakers of English in order to ensure the test questions were as clear as possible.

**Procedure**

The conventional listening test was administered to determine if both groups of subjects were of equal comprehension ability at the beginning of spring semester 1995. One hundred and fourteen students were asked to take the 30-minutes audiotaped listening test in their regular English 101 classes. The test scores were also used to examine correlations among the developed experimental tests of the targeted ability.

Two versions of 50 minute experimental tests (AFT and VFT) were administered six to seven weeks later to divide into the two groups in their regular classes. The purpose of this study was explained before the test administration and test takers were encouraged to do their best in conventional and experimental tests.
Ninety-two students took the experimental test, and only 85 students took both the conventional and experimental tests. Among them two students, who have had a chance to study the conventional or experimental test materials before, were excluded in this study.

Since a partial credit scoring method (scaling 0 to 3) was used in the experimental tests, three raters participated in the ratings. A rating criteria (Appendix D) and possible expected answers were shared and studied by three raters before they independently applied the criteria to all samples. Grammar and spelling were not taken into account in the rating criteria. Majority rule was used to get an agreed score for each item performance among raters. Only when some items didn’t have majority judgement across raters, raters discussed and regraded them together to get a majority judgement.

Unlike ratings of essays or oral interviews, the raters didn’t have very different scores from “inconsistencies in the criteria used to rate and in the way in which these criteria are applied” (Bachman, 1990:180), because the length of the requested responses was, in most cases, words or phrases, and because there were no comprehension questions asking for more than information explicitly on scripts.

Analysis

The scores in both conventional and experimental tests were computed by hand, and then transferred into a data set on a UNIX computer system. Using SPSSX procedures, descriptive statistics, reliability, correlation, t-tests, and normality were computed. In the conventional test, the difference between the meanings of the two groups was tested using a t-test to estimate equivalent comprehension level of the academic
listening ability as well as descriptive statistics and reliability. KR-21 was used for investigating reliability of the conventional test.

In the experimental tests, inter-rater reliability was estimated using Cronbach's alpha. Ratings of the different raters were totaled and the reliability of these totaled ratings was estimated by computing a coefficient alpha (Bachman, 1990:181). The coefficient alpha was also used to estimate internal consistency of the whole items in both experimental tests.

For concurrent validity evidence, the Pearson product-moment correlation coefficient was calculated between the conventional test and experimental test performances within each group. In empirical item analysis, the correlation between predicted item difficulty level and observed individual item means were also computed, as well as mean scores within theoretically-divided item difficulty categories. Group difference between two experimental test groups, based on t-test, descriptive statistics, normality of distribution, was also calculated.

Item difference analysis focused on how test takers' item performance differ under different experimental tests (VFT and AFT). Two indicators were used for the item statistics (Brown, 1989:1990): Item difficulty (IF), Item discrimination (ID). IF was used to find the proportion of students who (partially) correctly answered a given item. The IF index in this study was calculated by adding up all points credited for an item and dividing the sum by the total number of possible points for the item. Item discrimination is “an index of the degree to which an item separates the ‘high’ students from the ‘low’ ones” (Brown, 1989:70). ID was calculated by contrasting the performance of the upper third of
the students on the test with that for the lower third (13 students in AFT group and 14 students in VF group). To get ID, the IF for the lower group was subtracted from the IF for the upper group on each item.
CHAPTER 4. RESULTS AND DISCUSSION

In this chapter, test results are presented and discussed as reliability and validity evidence of the developed VFT in order to provide evidence for test usefulness qualities. Reliability is estimated through interrater reliability and internal consistency. Evidence of construct validity is investigated from concurrent validity evidence, empirical item analysis, and experimental manipulation evidences. Descriptive statistics and item difference (through IF and ID) will be presented in the experimental manipulation section, identifying group/item performance differences over different test presentation mode. They can partially support the inference that variances in the VFT responses can be attributed to academic lecture listening ability to be measured, so as to provide positive evidence for usefulness of VFT in second language testing. In the beginning of this chapter, results of the conventional listening test are presented to support the interpretation of the experimental test performances.

Conventional Listening Test Tasks

In addition to the randomly assigned grouping for the two experimental tests, a comparison of means of the two independent groups (Table 4.1) showed that there was no significant difference between the two groups (n=40 & 43), t = .88, at p < .05. This showed that, on the basis of a conventional listening test, the two groups appeared to be of equivalent level of comprehension level.
Level of comprehension level. Both groups produced test scores which were moderately reliable at .74 (VFT group) and .78 (AFT group) from KR-21 estimate. Additionally, the test takers in each group were almost normally distributed in the conventional test tasks (Figure 4.1).

Table 4.1  Means and standard deviation for the conventional test

<table>
<thead>
<tr>
<th>AFT Group: N=40</th>
<th>VFT Group: N=43</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>10.95</td>
<td>2.8</td>
</tr>
</tbody>
</table>

$t= .88, \ p<.05$

Figure 4.1  Histogram in VFT group and AFT group
Experimental Listening Test tasks

Reliability

First of all, reliability was addressed by estimating the interrater reliability to determine the degree of relationship among the scores assigned by three different raters. Inter-rater reliability of the VFT and AFT was estimated using coefficient alpha (.99 in both tests), which shows a strong reliability evidence of the tests. Table 4.2 shows mean value of each rater, and Table 4.3 shows correlations among raters.

Table 4.2 Mean value of each rater for both VFT and AFT

<table>
<thead>
<tr>
<th>RATER</th>
<th>MEAN (RANGE: 0-93)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>46.79 (34.92)*</td>
<td>11.97 (14.03)</td>
</tr>
<tr>
<td>B</td>
<td>46.56 (34.70)</td>
<td>11.49 (13.61)</td>
</tr>
<tr>
<td>C</td>
<td>45.72 (34.28)</td>
<td>12.31 (14.89)</td>
</tr>
</tbody>
</table>

Table 4.3 Correlation matrix among raters

<table>
<thead>
<tr>
<th></th>
<th>RATER A</th>
<th>RATER B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Rater B</td>
<td>.98 (.98)</td>
<td>1.00</td>
</tr>
<tr>
<td>Rater C</td>
<td>.97 (.99)</td>
<td>.97 (.98)</td>
</tr>
</tbody>
</table>

Inter-rater reliability (coefficient alpha) = .99 (.99)

* AFT values are given in the parantheses
High coefficient alpha among 31 items (.81 in VFT, .86 in AFT) shows that test takers' performances on different items of the VFT and AFT are internally consistent. In other words, in the VFT, only 19% are interpreted as random variance, and 81% of variance among items as consistent variance. So we may say that all items in the VFT reliably measure a construct. Since all comprehension questions in the test are designed to measure the same construct, the items are expected to measure the construct in a very consistent manner.

Validity

Concurrent validity evidence In additional to judgmental content validity evidence from authenticity and interactiveness studies, the first empirical data I gathered in the VFT validation process is relationship between the VFT scores and criterion, performances on a TOEFL-like conventional listening test of the academic lecture listening ability, which has been traditionally believed and used to be a good indicator of the ability.

Table 4.4 shows the correlations between conventional and experimental tests of the targeted ability. In order to find out if scores on the different measures are correlated, Pearson product-moment correlations were calculated.

The VFT is slightly more correlated with the conventional test than the AFT, even though the same facets of the input format (e.g., same channel/form of text input presentation) on the AFT can be expected to provide a basis for higher correlation with the conventional test. According to the concurrent criterion relatedness, the VFT may be
Table 4.4 Person product-moment correlation between conventional test scores and experimental test scores

<table>
<thead>
<tr>
<th>Conventional Test (AFT Group)</th>
<th>EXPERIMENTAL TEST (AFT)</th>
<th>EXPERIMENTAL TEST (VFT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.65</td>
<td>.69</td>
</tr>
</tbody>
</table>

estimated to be a good measure of the test takers' academic lecture listening ability.

However, it is hard to use the satisfactory correlation between the VFT and the conventional test of the same ability as an evidential basis of validity, when we suspect that test takers' performance on the conventional TOEFL-like listening test is not a good indicator of the academic listening ability (in lecture part). Even though they have become very familiar with the TOEFL type of multiple choice questions consisting of audiotape formatted listening tests (consisting of short lectures from audiotaped materials) and there have been studies on high index of reliability of the TOEFL-like listening tests, the construct to be measured in the conventional test remains a question.

If the criterion behavior cannot be validly interpreted as an indicator of the ability measured, the concurrent validity cannot be useful information in this validation process of the VFT. Bachman notes:
Without independent evidence supporting the interpretation of the criterion as an indicator of the ability in question, there is no basis for interpreting a correlation with that criterion as evidence of validity. As Messick, Cronbach, and others have argued, only the process of construct validation can provide this evidential basis of validity (1990:249).

Only with solid construct validity evidence for the conventional test, can the compared criterion relatedness provide useful information. The next two sections, investigating empirical item analysis and experimental manipulation, more directly provide empirical evidence supporting construct validity of the VFT than concurrent validity evidence in this study.

**Empirical Item analysis** Empirical item analysis evidence involves the theoretically based difficulty prediction discussed in Chapter 2. Construct validity of the VFT can be examined by comparing the item difficulty predictions to observed item performances, within each category. Based on the predictors of item difficulty, each item in the VFT was rated from the easiest comprehension task #6, (involving high degree of interactiveness and word-level comprehension task) to the most difficult task, #2 (involving low degree of interactiveness and sentence-level comprehension task). Table 4.5 shows the relationship between the predicted difficulty levels (#2 - #6) and mean scores of individual item in the VFT and the AFT. Predicted difficulty level six includes the easiest items in the test, and level two the most difficult items. The correlation coefficient between the mean scores and theoretically predicted difficulty level is .84 in the VFT and .65 in the AFT.
<table>
<thead>
<tr>
<th>ITEM #</th>
<th>Mean Score in VFT and (AFT)</th>
<th>Predicted Difficulty Level</th>
<th>Item #</th>
<th>Mean Score</th>
<th>Predicted Difficulty Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.40</td>
<td>4</td>
<td>17</td>
<td>2.81</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(1.08)</td>
<td></td>
<td></td>
<td>(2.18)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19.90</td>
<td>4</td>
<td>18</td>
<td>1.74</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(.90)</td>
<td></td>
<td></td>
<td>(.88)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.49</td>
<td>4</td>
<td>19</td>
<td>2.95</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(.28)</td>
<td></td>
<td></td>
<td>(1.30)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.37</td>
<td>2</td>
<td>20</td>
<td>2.19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(.60)</td>
<td></td>
<td></td>
<td>(1.33)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.79</td>
<td>4</td>
<td>21</td>
<td>2.74</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(.88)</td>
<td></td>
<td></td>
<td>(2.28)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.79</td>
<td>4</td>
<td>22</td>
<td>.65</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(1.05)</td>
<td></td>
<td></td>
<td>(1.05)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.91</td>
<td>4</td>
<td>23</td>
<td>.70</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(.73)</td>
<td></td>
<td></td>
<td>(.60)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.28</td>
<td>3</td>
<td>24</td>
<td>.56</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(.20)</td>
<td></td>
<td></td>
<td>(.60)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.44</td>
<td>4</td>
<td>25</td>
<td>1.56</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(.38)</td>
<td></td>
<td></td>
<td>(1.38)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.49</td>
<td>3</td>
<td>26</td>
<td>.79</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(.28)</td>
<td></td>
<td></td>
<td>(1.53)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2.33</td>
<td>6</td>
<td>27</td>
<td>.47</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(2.10)</td>
<td></td>
<td></td>
<td>(.50)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2.14</td>
<td>6</td>
<td>28</td>
<td>.30</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(1.85)</td>
<td></td>
<td></td>
<td>(.65)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.23</td>
<td>6</td>
<td>29</td>
<td>1.81</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(1.93)</td>
<td></td>
<td></td>
<td>(1.88)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.81</td>
<td>4</td>
<td>30</td>
<td>.65</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(1.30)</td>
<td></td>
<td></td>
<td>(.30)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.77</td>
<td>3</td>
<td>31</td>
<td>2.79</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(1.40)</td>
<td></td>
<td></td>
<td>(2.15)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.00</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Since correlation coefficient (r) squared is defined as the proportion of variance in one of the variables that can be explained by variation in the other variable, in the VFT, the r squared, .71, indicates that 71% of the variation of the item difficulty can be explained by the variation in the theoretically predicted difficulty levels.

But, in the AFT only 42% of the variation of the item performances can be responsible for the variation in the difficulty levels, and 58% of the variance of the performances is due to factors other than the two predictors of item difficulty. These results support the hypothesis that when test takers in the AFT didn’t have a chance to infer from background information and visual information, the degree of interactiveness becomes less responsible for the variance of test scores.

Table 4.6 shows mean scores of the items within each category of predicted difficulty level. The theory-based item difficulty prediction is strongly associated with test takers’ performances in the VFT. They performed best in comprehending word-level of information under high degree of interactiveness of listening tasks (OM = .83). It is seen that the test takers on the VFT had more success correctly answering items at PL = 5 than ones at PL = 4. They had more success correctly answering items at PL = 4 than at PL = 3. The VFT results show clear borders among the categories. In AFT, difficulty borders are not as clear as in the VFT.

Thus, it is found that the variance of item performances in VFT clearly corresponds to the hypothesized theory of the TLU task, which are mainly based on the degree of interactiveness. Since the construct of the academic listening ability highly
Table 4.6  Mean score of the items within predicted difficulty level of category

<table>
<thead>
<tr>
<th>Predicted difficulty level (PL)</th>
<th>Low Degree of Interactiveness</th>
<th>Middle Degree of Interactiveness</th>
<th>High Degree of Interactiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-Level (Extended Discourse)</td>
<td>PL = 2  OM = .22 (.27)  n = 6</td>
<td>PL = 3  OM = .44 (.28)  n = 3</td>
<td>PL = 4  OM = .59 (.59)  n = 5</td>
</tr>
<tr>
<td>Comprehension Task</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phrase-Level Comprehension Task</td>
<td>PL = 3  OM = .09 (.07)  n = 1</td>
<td>PL = 4  OM = .53 (.25)  n = 3</td>
<td>PL = 5  OM = .73 (.44)  n = 1</td>
</tr>
<tr>
<td>Word-Level Comprehension Task</td>
<td>PL = 4  OM = .30 (.26)  n = 4</td>
<td></td>
<td>PL = 6  OM = .83 (.62)  n = 8</td>
</tr>
</tbody>
</table>

Predicted difficulty level (PL), Observed mean score (OM)
( ) shows item difficulty index in the AFT

involves the interactiveness, the findings from empirical item analysis can provide useful evidence for construct validity of the VFT involving the targeted ability.

**Experimental Manipulation Evidence**  Two types of evidence are based on experimentally designed tests to support construct validity of the VFT: descriptive statistics and distribution, and item difference.

**Descriptive statistics and distribution**  In the experimental tests, observations were made to examine group differences between the VFT and the AFT (Table 4.7). A comparison of experimental test means for the two groups revealed a significant difference between the VFT group and AFT group, t = 4.24, n = 84 at p < .05. Test takers on the VFT (mean: 46.86) performed much better than on AFT (mean: 37.70).
Table 4.7 Means (M) and standard deviations (SD) for the post-test

<table>
<thead>
<tr>
<th>VFT GROUP: N=43</th>
<th>AFT GROUP: N=40</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>46.86</td>
<td>34.70</td>
</tr>
<tr>
<td>SD</td>
<td>SD</td>
</tr>
<tr>
<td>11.78</td>
<td>14.31</td>
</tr>
</tbody>
</table>

This implies that test takers of VFT, who benefited from many types of contextual clues, were able to use their strategies to improve their comprehension, as they use many contextual clues to make up for their communication deficiency in the TLU situations.

We may also conclude that poor performances on the AFT may not be largely due to their low level of comprehension skills. It might be due to a gap between their test task strategies and the TLU task ones. They couldn’t find alternative ways to make up for their deficiency in the 8 - 10 minute audiotaped lectures, unlike their 3 - 4 minute conventional audiotaped test tasks where they are already familiar with their own test taking strategies from their experience. In other words, poor performances on the AFT may be attributed to a lack of authentic and interactive factors involving TLU situations.

Because the observations in the previously taken conventional test showed that both randomly assigned groups are equivalent with respect to listening comprehension ability, we can attribute the differences in experimental test performances to different facets in the two experimental tests: channel/form of test input and printed background information, which are strongly related to the degree of authenticity and interactiveness, as I discussed in Chapter 2.
Since authenticity and interactiveness provide a basis for specifying the ability on which test score interpretations are made, and for examining aspects of construct validity (Table 1.1), the performance differences between the two tests, which are significantly affected by different degrees of authenticity and interactiveness, can be an indirect evidence to support the construct validity of the VFT.

Degree of normality of each test score distribution provides positive evidence for use of VFT within placement and selection testing purposes. Normal distribution of test scores is one of the important requirements of NRTs. Although test takers in both groups were normally distributed in the conventional test (Figure 4.1), and although the VFT has a nature of normality, the AFT has a positively skewed distribution (from Table 4.8 and Figure 4.2), which shows that the test cannot be used effectively for NRT purposes. Table 4.8 shows all skewness and kurtosis index in each distribution of test results, and Figure 4.2 shows histogram of scores in VFT and AFT.

**Item difference** At the item performance level, the experimental design can also be used to investigate the effect of different degrees of authenticity and interactiveness in test tasks. The effect provides an important basis for construct validity evidence of the VFT.

Item differences (by IF and ID combined) can provide evidence for the importance of test input facets involving validity of ESL listening test. Table 4.9 shows how IF and ID differ under two tests with different test input facets. Item performances with an asterisk before them are considered to be considerably different between two groups, and the items need to be examined by qualitative response analysis.
When selecting items for placement or selection purposes, the items within a range of .30 to .70 in IF and within .30 and up in ID (bolded in Table 4.10) are considered to be good ones. They can be assumed to be well-centered and discriminating well between the low and the high students (Brown, 1989; 1990). When we set the criteria for selecting items, the VFT has 11 acceptable items (item # 1, 3, 4, 5, 6, 7, 14, 15, 16, 18, 25) and the AFT 9 acceptable items (# 2, 11, 12, 13, 15, 16, 19, 20, 25, 29). There are only two items (# 15, 25) which are within the specified IF and ID ranges in both tests.

In the VFT context, for example, item 1, 3 and 4 can be considered to be good items with reasonable difficulty and discriminating power. They, however, fell to the problematic item categories in the AFT context. As for item 2 and 19, the AFT context provided more desirable item difficulty and discrimination power for NRT purposes than the VFT one. These findings show how inauthentic and low interactive test tasks affect the AFT takers' item performances to a different degree than the VFT tasks.
Table 4.8  Skewness and Kurtosis showing normality evidences

<table>
<thead>
<tr>
<th></th>
<th>AFT GROUP</th>
<th></th>
<th></th>
<th>VFT GROUP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conventional Test</td>
<td>Experimental Test</td>
<td>Conventional Test</td>
<td>Experimental Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>.08</td>
<td>.37*</td>
<td>-.06</td>
<td>.055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.98</td>
<td>-.66</td>
<td>-.88</td>
<td>-.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* evidence for positive skewedness of distribution

Figure 4.2  Histogram of Scores in the AFT and VFT
When I expanded the acceptable item selection criteria to a range of .30 to .80 in IF and .28 and up in ID, additional acceptable items in the VFT were #11, 12, 13, 20, 29 and in the AFT, #11, 17, 21. This means that the VFT context produced more good items for the targeted purposes than the AFT. We can say that high degree of authenticity and interactiveness, whose effects play a very important role in the TLU context, helped the VFT accept more items for the placement and selection testing purposes. For example, according to the item difficulty prediction based mainly on degree of interactiveness (Table 2.6), the items which were accepted only in the VFT require test takers to involve high interactive tasks (# 5, 6, 7, 14) or at least a middle level of interactiveness (#1, 3).

The differences in item performances between the two groups can be attributed to the different degree of authenticity and interactiveness found in the AFT and VFT tasks. So, according to the indirect evidence from the differences in group and item performances, the VFT may be judged to be a valid measure of the TLU tasks, which are also significantly affected by degrees of authenticity and interactiveness.
Table 4.9  Item difficulty and item discrimination in each item

<table>
<thead>
<tr>
<th>ITEM</th>
<th>VFT</th>
<th>AFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>.4651</td>
<td>.3333</td>
</tr>
<tr>
<td>2*</td>
<td>.6357</td>
<td>.4286</td>
</tr>
<tr>
<td>3*</td>
<td>.4961</td>
<td>.3114</td>
</tr>
<tr>
<td>4*</td>
<td>.3721</td>
<td>.3517</td>
</tr>
<tr>
<td>5*</td>
<td>.5969</td>
<td>.5000</td>
</tr>
<tr>
<td>6*</td>
<td>.5969</td>
<td>.5000</td>
</tr>
<tr>
<td>7*</td>
<td>.6357</td>
<td>.5000</td>
</tr>
<tr>
<td>8</td>
<td>.0930</td>
<td>.0952</td>
</tr>
<tr>
<td>9</td>
<td>.1473</td>
<td>.0714</td>
</tr>
<tr>
<td>10</td>
<td>.1628</td>
<td>.1428</td>
</tr>
<tr>
<td>11*</td>
<td>.7752</td>
<td>.3095</td>
</tr>
<tr>
<td>12</td>
<td>.7132</td>
<td>.4048</td>
</tr>
<tr>
<td>13</td>
<td>.7442</td>
<td>.3810</td>
</tr>
<tr>
<td>14*</td>
<td>.6047</td>
<td>.6190</td>
</tr>
<tr>
<td>15</td>
<td>.6047</td>
<td>.6190</td>
</tr>
<tr>
<td>16</td>
<td>.6667</td>
<td>.3571</td>
</tr>
<tr>
<td>17*</td>
<td>.9380</td>
<td>.1666</td>
</tr>
<tr>
<td>18*</td>
<td>.5814</td>
<td>.5238</td>
</tr>
<tr>
<td>19*</td>
<td>.9845</td>
<td>.0476</td>
</tr>
<tr>
<td>20*</td>
<td>.7287</td>
<td>.5952</td>
</tr>
<tr>
<td>21*</td>
<td>.9147</td>
<td>.0714</td>
</tr>
<tr>
<td>22</td>
<td>.2171</td>
<td>.1666</td>
</tr>
<tr>
<td>23</td>
<td>.2325</td>
<td>.3810</td>
</tr>
<tr>
<td>24</td>
<td>.1860</td>
<td>.3334</td>
</tr>
<tr>
<td>25</td>
<td>.5194</td>
<td>.3333</td>
</tr>
<tr>
<td>26*</td>
<td>.2636</td>
<td>.0476</td>
</tr>
<tr>
<td>27</td>
<td>.1550</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>.1008</td>
<td>.0714</td>
</tr>
<tr>
<td>29*</td>
<td>.6047</td>
<td>.2857</td>
</tr>
<tr>
<td>30*</td>
<td>.2171</td>
<td>.3571</td>
</tr>
<tr>
<td>31</td>
<td>.9302</td>
<td>.1190</td>
</tr>
</tbody>
</table>

* Items considerably different in IF or (and) ID
In summary, our analysis of the test results provides convincing support for the reliability and validity evidence of the VFT. In addition to the judgmental evidence from authenticity and interactiveness in Chapter 2, the test qualities provide positive evidences for the usefulness of the VFT.
CHAPTER 5. CONCLUSIONS

This chapter will summarize the answers to the research questions in the introduction. For convenience, they are restated here.

1) To what extent is a developed VFT similar to authentic and interactive tasks?

2) Is the VFT reliable in terms of interrater reliability? Are test takers' performances on the whole VFT items internally consistent with each other?

3) Do the VFT scores significantly correlate with scores on a conventional listening test?

4) Does empirical item analysis provide construct validity evidence for the VFT?

5) Is there a significant mean difference in the group performances between the VFT and the audiotape formatted test? Are the test takers' scores normally distributed for placement and selection testing purposes? Are there significant item difficulty and item discrimination differences in the item performances between the VFT and AFT?

Question one was answered by the systematic analysis from Bachman and Palmer (1994) and Bachman (1990) frameworks. There is evidence that, in contrast to audiotaped test tasks which overlook many important correspondance between test tasks and the TLU context as well as between the test tasks and test takers, VFT can successfully introduce the necessary characteristics to make the test tasks more authentic and interactive.

In answer to question two, based on the interrater reliability in the VFT, it can be concluded that VFT can be reliable when it is designed to be an authentic test task. Since
it has been believed that it is not possible to design test tasks that are authentic and at the same time reliable, the relationship between authenticity and reliability needs to be reconsidered in the test development and evaluation process. Internal consistency among items, from coefficient alpha, provides another evidence for the usefulness of the VFT.

Although not strong, the satisfactory correlation index between the conventional and the VFT provides concurrent validity evidence for question three. In order to conclude that the VFT is a good measure of the TLU tasks, construct validity of the related criterion needs to be approved.

With regard to the answer to question four, the item difficulty prediction based mainly on degree of interactiveness was found to function as construct validity evidence, showing high correlation with empirical item performances. Even though the factors affecting item difficulty are oversimplified by two predictors, the empirical item analysis can support the hypothesis that interactive tasks are important aspects of the construct involving the academic lecture listening ability, and that they provide a useful basis for examining construct validity of VFT.

Question five was answered by the results of group and item performance differences between the VFT and the AFT. The performance differences are considered to be indirect construct validity evidence, because like in the TLU situation, they can be attributed to the different degree of authenticity and interactiveness from different facets of two tests. Since normal distribution of test performances is an important requirement for selection and placement testing purposes, the AFT scores of which were positively skewed, were judged to be inappropriate for the purposes.
The main conclusion of this study is that the VFT method can be a useful assessment tool to measure academic listening ability for placement and selection testing purposes, and that the strong relationship between test performances and aspects of authentic and interactive tasks in the experimental tests provides a justification for revision of test method facets in conventional ESL listening tests. However, this research only investigated some of the qualities of test usefulness, leaving additional issues for future research.

First of all, impact and practicality, which can be investigated in supplemental usefulness studies, need to be considered to justify the usefulness of the VFT. To provide additional evidence for justifications of test results, "the relevance and utility, value implications, social consequences of a test" (Messick, 1989) need to be examined within the context of test use. Chapelle (1994) stated:

Construct validity is central to all facets of validity inquiry, as most researchers have agreed for some time. In addition to construct validity, however, researchers should also consider justifications pertaining to test utility and the consequences of testing. In other words, according to the current conception of validity, researchers are obligated to use construct validity evidence as a basis for considering how their tests impact the contexts in which they are used (161).

Secondly, in order to support authenticity and interactiveness studies, more qualitative research needs to be done to investigate the language users' and the test takers' listening process, because the tasks can be "differently perceived by each group of test takers, users, or developers" (Bachman and Palmer, 1994 I: Chapter 3, p. 8). As evidence for construct validity, the test taking process is expected to hold "perhaps the most promise for providing new insights into the factors that affect test performance"
Bachman and Palmer note:

In either designing new tests or analyzing existing tests, our estimates of authenticity and interactiveness are only guesses. We can do our best to design test tasks that we believe will be authentic and interactive for a given group of test takers, but we need to realize that different test takers may process the same test task in different ways, often in ways we may not anticipate (Bachman and Palmer, 1994 I: Chapter 3, p. 10).

A more critical limitation to correlational and experimental approaches to construct validation ... is that these examine only the products of the test taking process - test scores - and provide no means for investigating the process of the test taking themselves (Bachman, 1990: 269).

In Buck's introspective studies on listening process (1991, 1994), many useful topics in need of research are presented. He points out a "lack of knowledge of how listening comprehension works" in tests of second language listening ability, saying:

... a review of the literature on listening, both L1 and L2, suggests that there is no generally accepted, explanatory theory of listening comprehension on which to base these tests (Buck, 1990). It seems that in practice test constructors are obliged to follow their instincts and just do their best they can when constructing tests of listening comprehension (1991:67).

In this study for example, like the incorrect options from multiple choice questions, much more extended discourse, in itself, may function as 'distracters' which interfere with the test takers' comprehension process. If we find a relationship between the 'distracters' hidden in authentic lectures and the test taking process, it will yield useful information about characteristics of text difficulty in designing a test.

Finally, as we continue to understand VFT, computer based listening tests can be thought of for future studies. Like with videotaped materials, a computer-based listening test with video is expected to be an useful measurement tool because it can also provide rich contextual information in test tasks reflecting TLU situations. Much research has
been already conducted to implement, for example, computer adaptive listening tests (CALT), which show a lot of advantages over paper-and-pencil tests (Dunkel, 1991). CALT can make it possible to introduce both characteristics of adaptability and of reciprocacy between comprehension input and output, so that test tasks can meet TLU tasks more closely than any test, and thus potentially increase test usefulness.

Differences in the test takers' performance on the VFT and AFT in this study suggest further research on test usefulness of the CALT. The performance differences, for example, could throw doubt on the reliability of a calibrated level of item difficulty and construct validity of CALT, because the item difficulty levels used in CALT are still based on audiotaped pre-tests and CALT has similar test method facets to AFT. Since listening ability cannot be measured "in a linear progression from low ability to high ability" (Buck, 1994:148), the computer-based listening test projects need to consider "multidimensional measurement models, cognitive information-processing models in language testing" (Buck, 1994:164) based on authentic and interactive tasks in TLU or test contexts.

Because of the preliminary nature of this study, the results of this research cannot be conclusive concerning the usefulness of VFT. Given the significance of this topic for the future of listening comprehension testing, additional research is needed using different sampling and test materials. Each test usefulness quality may function differently from one testing situation to another. Based on various evidence for the usefulness of the VFT in this study, however, positive expectations are justified on the basis of this study for achieving the appropriate balance among test usefulness qualities in L2 listening test.
REFERENCES


ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. Carol A. Chapelle for serving as my major professor and providing me with patient guidance throughout this project. Her ability to make sense out of all the data when I was unable to see the forest through the trees, was most appreciated. I also would like to express my gratitude to Dr. Barbara Schwarte and Dr. Ann Thomson, the other members of my committee, for their time and thoughtful suggestions. I cannot possibly list all students and teachers who participated in my study. Thanks to all those who have aided me in some way to complete this thesis.

I would like to express a very special thanks to my wife, Eunsook, for her support and encouragement during my graduate studies. She is a wonderful blessing to me.
APPENDIX A. CONVENTIONAL TEST

Directions: In this test, you will hear short lectures. After each of them, you will read the questions about them.

After you read a question, read the four possible answers and decide which one is the best answer to the question you read. Then, mark the letter of the answer you have chosen.

1. What central theme does the lecture examine?
   (A) The relationship between language and culture
   (B) The culture of Hopi society
   (C) American Indian cultures
   (D) The life of Benjamin Lee Whorf

2. Who was Whorf's teacher?
   (A) Boas   (B) Sapir   (C) Franz   (D) Yale

3. What was Sapir's most well-known book?
   (A) A Handbook of American Indian Languages
   (B) The Technology Review
   (C) Language
   (D) Linguistic Patterns

4. What is another name for linguistic relativity?
   (A) The Sapir Hypothesis
   (B) The Sapir-Whorf Hypothesis
   (C) The Sapir-Whorf-Boas Hypothesis
   (D) The American Indian Model of the Universe

5. According to the lecturer, what is linguistic relativity?
   (A) All languages are related
   (B) All American Indian languages are related
   (C) Language influences the manner in which an individual understands reality
   (D) Language and culture are not related
6. What was the contribution made to medicine by William Harvey?

(A) The theory of germs and bacteria  
(B) The discovery of a vaccine against smallpox  
(C) The discovery of a mechanism for the circulation of the blood  
(D) The Materia Medica

7. Who is known as the father of biology?

(A) Hippocrates  (B) Aristotle  (C) Dioscorides  (D) Edward Jenner

8. What was Hippocrates' greatest work in the field of biology as it relates to medicine?

(A) The classification of plants on the basis of body structure  
(B) The sterilization of surgical instruments  
(C) The scientific recording of symptoms and treatments  
(D) The theory that disease was caused by the gods

9. According to this lecturer, the greatest single contribution of biology to medicine was made by which man?

(A) Sir Joseph Lister  (B) Louis Pasteur  
(C) Edward Jenner  (D) William Harvey

10. This discussion is concerned with a literary movement that was active during which century?

(A) Seventeenth century  (B) Nineteenth century  
(C) Eighteenth century  (D) Twentieth century

11. According to the lecture, what did the Puritans do?

(A) They stressed the importance of the individual  
(B) They supported the ideals of the Transcendental Club  
(C) They believed that society was more important than the individual  
(D) They established a commune at Brook Farm

12. Which of the following is the title of one of Emerson's essays?

(A) "Judge Yourself"  (B) "Self-Reliance"  
(C) "The Puritans"  (D) "Society and Individual"
13. What is Walden?

(A) A book by Emerson  (B) A history of Puritanism
(C) A novel by Nathaniel Hawthorne  (D) A book by Thoreau

14. Why did Thoreau go to jail?

(A) He wrote an essay criticizing the government
(B) He refused to pay taxes
(C) He built a cabin on someone else's land
(D) He refused to pay rent for his cabin

15. What was the main idea of this talk?

(A) Health food  (B) Organic gardens
(C) The processing of bread  (D) Poisons

16. Which term is used to distinguish between types of the same food?

(A) Refined foods  (B) Organic foods
(C) Natural foods  (D) Unprocessing foods

17. Which did all of the addictives in bread have in common?

(A) They are all used to keep the bread from getting moldy
(B) They are all poisons
(C) They are all organic
(D) They have all killed laboratory animals

18. What happens to food which it is processed?

(A) The ultimate content remains the same
(B) Vitamin information is not available after processing
(C) Vitamins are added to the food
(D) The vitamin content is reduced
Scripts:

Lecture 1

In the class time remaining, I would like to outline the development of the Sapir-Whorf Hypothesis concerning the relationship between language and culture.

Prior to the twentieth century, most linguists had been concerned with historical linguistics; that is, the comparison of European languages. In contrast, Franz Boas, a specialist in American Indian languages, asserted that historical studies were inappropriate to the material that he was investigating. He further suggested that the inner logic of each language excluded the application of any general principle or method for describing it.

One of Boas' greatest achievements was the publication of the monumental work, A Handbook of American Indian Languages. This is not only a collection of admirably assembled and classified material on American Indian languages, but also a fundamental contribution to the theoretical problem of establishing a relationship between language and culture.

A student of Boas at Columbia University, Edward Sapir became the foremost authority on the science of linguistics, especially in the area of American Indian languages.

In one of his most well-known books entitled Language, he advanced the idea of linguistic patterns. Sapir believed that each man carried within himself the basic patterns in the organization of his language, and that, in order to understand the patterns, a very thorough knowledge of the cultural environment of the language was necessary.

Unlike Boas, who had only suggested language as a measure of culture, Sapir proposed a mutual relationship between culture and language.

When Sapir was teaching at Yale, Benjamin Lee Whorf enrolled in his class. A chemical engineer, Whorf neither sought nor obtained a higher degree in linguistics. His contribution to the science of language is nonetheless significant.

Whorf was recognized for his investigations of the Hopi language, including his authorship of a grammar and a dictionary. Even in his early publications, it is clear that he was developing the theory that the very different grammar of Hopi might indicate a different manner of conceiving and perceiving the world on the part of the native speaker of Hopi.

In 1936, he wrote "An American Indian Model of the Universe," which explored the implications of the Hopi verb system with regard to the Hopi conception of space and time.

Whorf is probably best known for his article, "The Relation of Habitual Thought and Behavior to Language," and for the three articles which appeared in 1941 in the Technology Review.

In these articles, he proposed what he called the principle of "linguistic relativity," which states, at least as a hypothesis, that the grammar of a man's language influences the manner in which he understands reality and believes with respect to it.
Since the theory did not emerge until after Whorf had begun to study with Sapir, and since Sapir had most certainly shared in the development of the idea, it came to be called this Sapir-Whorf Hypothesis.

Lecture 2

Today's lecture will include the most outstanding achievements in biology as it relates to the medical sciences.

Early in Greek history, Hippocrates, who lived from 460 to 370 B.C., began to study the human body and to apply scientific method to the problems of diagnosis and the treatment of diseases.

Unlike other physicians of his time, he discarded the theory that disease was caused by the gods. Instead, he kept careful records of symptoms and treatments, indicating the success or failure of the patient's cure. He had been recognized as the father of modern medicine.

About a century later, Aristotle began a scientific study of plants and animals, classifying more than five hundred types on the basis of body structure. Because of his great contribution to the field, Aristotle has been called the father of biology.

By the first century A.D., Dioscorides had collected a vast amount of information on plants which he recorded in the now famous Materia Medica, a book which remained an authoritative reference among physicians for fifteen hundred years.

During the Middle Ages, scientific method was scorned in favor of alchemy. Some scientists were even imprisoned for carrying out their investigations.

Thus, medicine and biology had advanced very little from the time of the ancients until the seventeenth century when the English physician and anatomist William Harvey discovered a mechanism for the circulation of the blood in the body. Harvey's "Essay on the Motion of the Heart and the Blood," published in sixteen twenty-eight, made possible a clear understanding of the respiratory functions of the blood.

By the end of the eighteenth century, Edward Jenner had discovered a vaccine against smallpox. His contribution not only controlled the disease itself, but also established the science of immunization.

Louis Pasteur's theories about germs and bacteria advanced in the nineteenth century are considered by many to be the greatest single contribution of biology to medicine. Within a few decades, the causes were isolated for such ancient diseases as leprosy, plague, diphtheria, and tuberculosis.

But the advance of the twentieth century in curative and preventive medicine and biology are far more numerous than all other periods combined. Consider that the sterilization of surgical instruments, largely the idea of Sir Joseph Lister, is a process only seventy-five years old. Sulfa drugs, antibiotics, and X-rays are discoveries of the past fifty years.
Lecture 3

Today we will discuss Transcendentalism, which is a philosophical and literary movement that developed in New England in the early nineteenth century.

Transcendentalism began with the formation in 1836 of the Transcendental Club in Boston, Massachusetts, by a group of artists and writers. This group was the advance guard of a reaction against the rigid Puritanism of the period, especially insofar as it emphasized society at the expense of the individual.

The Transcendental Club published a literary magazine, the Dial, and some of its members participated in an experiment in communal living at Brook Farm.

One of the most distinguished members of the club was Ralph Waldo Emerson who served as editor of the Dial. His essays stressed the importance of the individual. In one of his most well-known essays, "Self-Reliance," he appealed to intuition as a source of ethics, asserting that each person should be the judge of his own actions, without the rigid restrictions of society.

From 1841-1843, Emerson entertained in his home the naturalist and author Henry David Thoreau. Partly as a result of their friendship, Thoreau became a member of the Transcendental Club.

Probably more than any other member, he demonstrated by his life-style the ideas which the group advanced. He preferred to go to jail rather than to pay taxes to the federal government, and affirmed that the best government was that which governed least.

Upon leaving Emerson's home, Thoreau built a small cabin along the shores of Walden Pond near Concord, Massachusetts, where he lived alone for two years. Devoting himself to the study of nature and to writing, he published an account of his experiences in Walden, a work which is generally acknowledged as the most original and sincere contribution to literature by the Transcendentalists.

Lecture 4

Health food is a general term applied to all kinds of foods that are considered more healthful than the types of foods widely sold in supermarkets. For example, whole grains, dried beans, and corn oil are health foods. A narrower classification of health food is natural food. This term is used to distinguish between types of the same food. Raw honey is a natural sweetener, whereas refined sugar is not. Fresh fruit is a natural food, but canned fruit, with sugars and other additives, is not. The most precise term of all and the narrowest classification within health foods is organic food, used to describe food that has been grown on a particular kind of farm. Fruits are vegetables that are grown in gardens that are treated only with organic fertilizers, that are not sprayed with poisonous insecticides, and that are not refined after harvest, are organic foods. Meats, fish, dairy and poultry products from animals that are fed only organically grown feed and that are not injected with hormones are organic foods.

In choosing the type of food you eat, then, you have basically two choices:
inorganic, processed foods, or organic, unprocessed foods. A wise decision should include investigation of the allegations that processed foods contain chemicals, some of which are proven to be toxic, and that vitamin content is greatly reduced in processed foods.

Bread is typically used by health food advocates as an example of a processed food. First, the seeds from which the grain is grown are treated with bichloride of mercury, an extremely toxic poison. Later, the grain is sprayed with a number of very toxic insecticides and pesticides. After the grain has been made into flour, it is bleached with nitrogen trichloride or chlorine dioxide, both toxic. Next, a dough conditioner, usually ammonium chloride, is added along with a softener, polyoxylethene. The conditioner and softener are poisons, and in fact, the softener has sickened and killed experimental animals.

A very toxic antioxidant is now added, along with coal tar, a butter-like yellow dye. Finally calcium propionate, an anti-fungal compound, is added to keep the bread from getting moldy.

Other food from the supermarket would show a similar pattern of processing and preserving. You see, we buy our food on the basis of smell, color, and texture, instead of vitamin content, and manufacturers give us what we want—even if it is poisonous.

The alternative? Eat health foods, preferably the organic variety.
APPENDIX B. VIDEOTAPED EXPERIMENTAL TEST

Directions: In this test, you will watch 7-11 minute videotaped lectures. After each of them, you will read open-ended questions about them. After you read a question, write an answer to the question you read.

When an incomplete sentence is given, you are asked to write words that can complete the sentence.

While you are watching each lecture, take notes on the allowed page. Make your notes clearly organized and selective. Try to include all the main points and enough specific details to support and explain them. That should be really helpful when you answer the open-ended questions.

Read: Humanities is the interdisciplinary study and appreciation of creative human endeavor. In a Humanities class, a lecturer introduced a poem, "Borders," written by Pat Mora in the beginning of class. Pat Mora's poem "Borders" is about communication problems. Now she will explain about the three aspects of the poem. Watch the following seven minute lecture, taking notes as conscientiously as you can. Then, answer the following questions about meaning of "borders" in Pat Mora's poem, by referring to your notes. You are allowed to take notes in this page.

1. What is the first example of communication problems that Mora's poem describes?

2. What is the second example of communication problems that Mora's poem describes?

3. What is the third problem the poem describes? The third problem is the climax of her poem.

4. It is not surprising that Mora uses the example of translating between English and Spanish, why?

5. Why does the lecturer give some examples of words meaning "fat"?

6. Why does the lecturer gives her personal experience about her little daughter?
7. Why does the lecturer talk about the meanings of success, happiness, and honor?

8. What is the internal dialog Mora's poem describes?

9. Words change in the way people use them. If speakers are categorized, the changes are going to be in different manners. Give an example of ways speakers can be isolated?

10. When we are aware of borders in communication, what is (can be) a positive result of borders in real word?

Read: Ecosystems are giant assemblages of plants and animals all interacting together, such as a redwood forest or tropical rain forest. Ecologists study interactions between organisms in an ecosystem, and what happens if any part of the ecosystem is altered. Now, a lecturer will talk about food chains and how they work. In order to do that, she will give you some background information on some of the components of the ecosystems in the beginning of her lecture. Watch the following 11 minute lecture, taking notes as conscientiously as you can. Then, answer the following questions by referring to your notes. You are allowed to take notes in this page.

11. Define autotrophs, and give an example.

12. Autotrophs' chlorophyll transforms _______ into _________ by the process of photosynthesis.

13. Define heterotrophs, and give an example.

14. Explain how the heterotrophs are totally dependent on the autotrophs.

15. State one important thing decomposers do?

16. Define abiotic, or give an example of abiotic features affecting the ecosystem.

17. Who are the primary producers in the food chain?

18. Why is our food chain pyramid shaped, instead of box shaped in which there is an even number organism in each level?

19. How much energy is stored when one organism converts it to another?

20. What is most of the lost energy used for?
Read: In a music class, a lecturer is going to talk about the development of a music which is uniquely American, and which is influenced by many cultures from around the world, both in its beginning and today. That music is jazz. In the beginning of class, some important elements of jazz were listed. The four most important elements of jazz are improvisation, rhythm, lifestyle and composition involving performance. Whereas classical music is the composer's art, jazz is the performer's art.

Now the lecturer will talk about origins of jazz. Watch the following eight minute lecture, taking notes as conscientiously as you can. Then, answer the following questions by referring to your notes. You are allowed to take notes in this page.

21. In Africa, what were the 2 main functions of music?

22. Communication at the speed of sound is very important beginning of the development of rhythm in Africa, why?

23. White slave masters tried to strip away the slaves' native cultures, but what did they fail to ban in Africa?

24. Explain the field holler that slaves used in the fields.

25. Explain blues, the name for the 19th Century music, comparing with gospel hymns.

26. According to the speaker, as the slavery period ended, what happened to black people?

27. According to the speaker, after World War I, what happened to the black people?

28. What is Dixieland?

29. As jazz grew more sophisticated, the banjo used for rhythm was replaced by the ________.

30. What is the element of classical music that some educated jazz musicians added to their music?

31. After World War I, which city became the new center of jazz?
I think the narrator is asking these questions of her husband, and I think this is a very healthy beginning, for the best way to translate is to try to communicate. And I think the fact that she has written this poem is a very good beginning.

Ah, she starts with perhaps the most obvious need for translation between different languages. Mora is from El Paso which is on the border of Mexico and the United States, so I think it's not surprising that she takes her example of translating between Spanish and English. Now we know when we go from one language to another, that some problems may arise. Some words are not even translatable. Perhaps you know a word in a language and there is no equivalent in another language. Perhaps there is kind of any equivalent, but may be the connotations are different.

Let's look at that word "connotations." Connotations refers to the feelings and associations with a particular word. For example, let's say someone, my height, weighs 200 pounds. There are a lot of different ways we might describe such an individual. And perhaps our attitude toward that individual would determine the connotation of the word we would describe that individual with. For instance, we might say, "she's kind of stocky." Or we might say, "she's fat." Or we might say, "she is obese." Or "she is a little bit overweight."

Ah, I'll describe the same individual, but the word choice, the different connotation reflects the attitudes toward that individual. And when we translate from one language to another, I think we're quite conscious of the fact that the connotations of words may not be identical. But since we are conscious of that, we can be careful with it. We're less likely to run into difficulties because we're aware of it, ironically, even though it's two totally different languages.

Now, her second example is a difference in languages we might not be quite so aware of between child talk and adult talk. I'll give you an example of my personal experience. My little daughter sometimes comes to me when I am working in my office at home and I'm working on my computer. And I hear her coming down the hall and sure enough she comes in and she says, "Mom, Mom, will you give me a drink of juice?" So that means I walk into the kitchen and get her a drink and I put it on the counter for her. She doesn't even take a sip. What she really wanted was my time and attention.

She didn't care about getting a drink of juice at all. So I have learned to translate "I want a drink of juice" into really "please give me a little time and attention."

I think that Mora's example also points out an internal dialogue that each one of us has, regardless of our culture, ah, between that part of ourselves which is the child, Freud would say that "id" in dialog with the adult, the "superego," as Freud would say. How are we reconciled the demands of one against the attitude of the other? For example, the child is, even each of us might say, "I want another piece of cake," or "I want that vacation," or "I want that new car," or "I want your love." But adult part of ourselves says, "Do I have to work for these things?" So we have to have this ongoing dialogue and the ego
resolves it, or the, ah, the adult may resolve it with the parent and child within each one of us, ar, arguing back and forth. So this is a border even within ourselves. I think good poem like this one prompts additional examples; it's thought provoking.

Her third example is the most powerful. It's the climax of her poem, communication between intimates. Talk, talk, talk, but they are not communicating. They are using the same words. What could the problem be? The problem is that words don't mean the same thing to everyone.

She uses two examples. Success. What does success mean? Well, it depends on who's using the word. Or happiness. I wonder if you can think for a moment of any words that might possibly mean something different, perhaps if a man uses it or if a woman uses it. Now I don't want to stereotype here. But if you think of any general things that might cause some differing definitions and some problems in communication. Do you think of any examples? (pointing to a student) Yeah. (A student's answer: "Honor.") Honor, that's good one. What is honor for a man? What is honor for a woman? Ah, what is honor for a country? I think words change meaning over time. That's good example of the word that changes over time.

Let's look at, ah, how all language changes. Spoken language is constantly evolving. Just as fashions change, hairstyles change, words change in the way people use them, constantly changes. And if you isolate speakers, the changes are going to be in different ways. Speakers might be isolated due to ages, due to geographical locations, due to social class or educational level, or due to experience growing up as a man or a woman. Words are very slippery things. A word is not a word, but a word is something used by people and we try to communicate with them. We have to be aware, though.

There are many borders in communication and Mora's poem, I think, brings us to our attention a word is not all that simple. That doesn't do all the communicating. O.K. We have to go beyond borders. Also borders can help to define who we are. A border says, "This is me" and "This is everyone else." "This is the rest of the world." Perhaps in some cultures, border definition is quite different. In America, we emphasize the self, the individual, "Give me my space." But in some cultures, there's much more emphasis on fitting in with the group, subordinating the self for the good of the community. I think being aware of that kind of personal border can help to enhance the possibility for appropriate communication.

You might think about times in your life when borders have made a big difference. Perhaps you recognized the border that you had not even realized was there. Or perhaps you created a border. Maybe you eliminated a border. Maybe you crossed over a border. Now talking in the figurative sense, not the literal sense.

Well, in this humanities course as we study more literature, I would like you to keep this in the back of your mind. And as we read works by a variety of people from all over the world, I want you to see, if you can recognize, that particular characters have bumped into a border, or overcome a border, or perhaps need to redefine their own
borders. I like her metaphor and I think we'll find it applicable not only in this literature, but in our lives as well. If you have any questions, please do come see me during my office hours. And, ah, continue the reading, and I look forward to seeing you next class if not in my office before then. Thanks.

Lecture 2 "How Food Chains Work"

So let's look at some of the components of the ecosystems, what sorts of things are interacting in these assemblages and we can start out with the non living components.

So when we're looking at the components, we can start out with the non living components and start out with the most important organism in the ecosystem which turn out to be the autotrophs. Now that's probably a foreign word, but one of the things I like to do with students is, ah, have them translate biology because often times biology is a, lang, is a language in and of itself. And if we take this, this word apart, what we find is auto here, refers to self, right? Like automobile, automobile means self moving, right? So this has to do with self, and troph always refers to feeding. So what we have is self-feeding organisms, or self feeders. Now, who can feed themselves? I mean, what organism can feed itself? Well, it turns out green plants can do that, O.K.? So, if you go into an ecosystem like a redwood forest, everything green you're seeing in there that's plant ( ) can actually feed itself, O.K.? So this includes the plants.

And now we have to ask ourselves, how do they do that? How do they feed themselves? Well, they contain a unique pigment called chlorophyll, chlorophyll pigment. And this Chlorophyll pigment is able to transform sunlight energy into food directly. In fact, our very existence depends on this process, O.K.? So to give you an idea, a visual idea of what that would look like, try to imagine a huge leaf in this room here and imagine radiant sunlight energy landing on that leaf and sugar cubes popping out. As the sunlight energy flows through the leaf, you'll see sugar cubes coming out. So these leaves are actually food making factories. They transform radiant sunlight energy into food and all of the organisms in the entire biosphere in this planet are dependent on that process. Now, what's that process called? That process in biology is called photosynthesis. Now we can again take the word apart. And synthesis means to make, right? Synthesize, to make, and photo means, refers to from sunlight, photo refers to light. So we're making food from sunlight. That's what photosynthesis refers to. So autotrophs, they are the key components in the ecosystem because, without them, there'd be no food.

The next group are the heterotrophs, O.K.? Now who knows what hetero means? (A student's answer: "Variety.") Variety, right, or mixed and homo means same, like homogeneous crowd, ah, hetero means mixed. What we have is mixed feeders. Here we already know that trophs means feeders. So we have mixed feeders and these include all the organisms that directly or indirectly are dependent on the work of the autotrophs for their food supply. So this would include all the animals. When you think about what you eat, we're animals, and when, when you think about what you eat, it's either, dir, plant
life directly or it's something that was converted from, pla, plant life, right? Like if you eat a piece of chicken, the chicken came from eating corn, and the corn came from a plant, right? So everything we eat directly or indirectly comes from plant life, and that's true with all the heterotrophs on the plant, which includes all of the animals and all of the fungi basically. O.K.? So, the heterotrophs are totally dependent on the autotrophs. Ahh, when you think about it, for instance, in the human population, you have almost 6 billion people on the planet. Most of those people are dependent on 3 crops: rice, wheat, and corn. All of which are autotrophs, autotrophically based, ah, food supplies.

O.K. Then a third group are the decomposers. Because you see in any ecosystem, autotrophs and heterotrophs are constantly manufacturing waste products. Also, they die. So what do we do with all of these waste and all of these dead bodies? And it turns out that these dead bodies and waste products are full of nutrients. You know that yourself if you put manure in a garden it makes your plants really thrive. So how do we get the nutrients in those dead bodies back into the ecosystem in a form that can be used to generate new life? And the answer to that question is the decomposers do the work. And the decomposers consist primarily of 2 groups of organisms. These are the bacteria and the fungi, O.K.? And the bacteria and fungi take waste products that these organisms produce, they take the dead bodies, they pull all of the important ingredients out of those bodies, either put them back in the environment, so that they can be used by plants to make more food or else put into the soils. So the plants that are existing can absorb the materials through their roots and create new plant growth. So anyway you slice it, the decomposers are providing the autotrophs with more of the things they need to continue producing food for the whole ecosystem, O.K.? So these are the key living players: the autotrophs, heterotrophs, and decomposers.

Now, obviously, these organisms don't live in a vacuum. They're actually connected with the whole physical environment, and in that physical environment that we can look at a set of what's called abiotic, or non living factors. O.K.? And again, we can take this word, people, and take it apart to understand it. Bio means living. This part here means living or life, bio. A means without. So what we're talking about is, a stra, substances or, ah, features in the environment that are non living. Now what are some non living features that an organism would interact with? Well, we can feel that right here we're interacting with temperature, the amount of humidity in the room, the amount of sunlight in the environment, ah, the, how much snow falls, how much precipitation, what kind of soil there is, all of these are types or, or features of the environment. They are not living organisms.

So what we have then in ecosystems is these key components all interacting together to form a very integrated whole, O.K.? So how do they feed upon each other? How do food chains work? We're down now to our last point here. And to look at how food chains work, where do you think we'd start? With the autotrophs, because they produce the food everyday, right? O.K. So we'll start with the autotrophs and the autotrophs, because they produce the food for everybody, are often times referred to as primary producers. Because they produce the food for everybody else, so we can start a food chain here with our autotrophs at the base. Now let's just make one up--a real simple one. Ah, let's say that
you're going to have steak for dinner. So what we're gonna do is start out with where the steak came from, and where does the steak come from? (A student's response: "A cow.") A cow, where does the cow get its food? (A student's response: "Grass.") From the grass, O.K.? an, and which are primary producers, or autotrophs, right? So we'll have a little simple food chain with grass here. This is our primary producer level.

And then we're gonna have a cow come along and I guess it's a steer, it's that right? A steer come along, and we're gonna have the steer eat the grass and then you are up here, and you're gonna have steak dinner, O.K.? Now if you notice, I made this food chain here, kind of pyramid shaped rather than box shaped. And what we are gonna look at is why it is that, ah, there's not as many, as much people on the, on the planet basically as many people as there are autotrophs or grassy areas. I mean why is it that our food chain is pyramid shaped instead of box shaped in which we'd have an even number organism in each level. And answer that question is that everytime you convert one form of organism into another, up a food chain, you lose 90% of the potential energy that's stored in that, that, particular type of organism, O.K.?

Now I'll give you a real good analogy here 'cause sometimes I say things or, you know, some of, the con, the concepts in biology are sometimes complex. But an analogy will make it real clear. So let's do this. Let's pretend like we have a little cow here and we want it to gain one pound in a day, O.K.? So what we do is we weigh our little cow in the morning. We, we send our cow out in the meadow. We say, “Nice little cow, I want you to gain one pound, pound." And we keep track of every single thing that cow eats, all of the weight it puts into its body, O.K.? So we send the cow out, the cow's running around. At the end of the day, we bring it back, we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weigh it, and we weight it and it's all digested, it's gained one pound, everyone follows?

So everytime you convert them, the grass into the steer, or the steer into the human, you lose 90% of the energy. Now why is that? The reason is because, ah, most of the energy goes into just maintaining the body. You know the cow doesn't just sit there, and even just sitting there, its hearts beating, and it's, gotta produce heat to maintain its body, and it's running around swatting flies and chasing other cows, it's gonna lose a whole lot of energy, right?

O.K. So most of the energy then in the, in the, ah, food it eats goes into maintaining the body. So only 10% is actually stored as weight. And for that reason then, ah, food chains are pyramid shaped rather than box shaped and they are the bases for understanding the complex relationships between organisms in ecosystems, simply because you have to have a plant base that's large enough to support the organism at the top of the food chain, O.K.?
Lecture 3 "Origins of Jazz"

Now, Jazz music is primarily, some people think, only a music of the black voice. The black, the music, Jazz sprang from the experiences and the lifestyle of black people through their voices. Now before the blacks that came over here were taken from their home in Africa, they were already playing music mostly in two different parts of their lives. One was communication. Africa, a very difficult place to live in many parts of the continent with vast distances between places. It's just a huge continent. One of the ways people could communicate was through sound. If you can't see somebody because they're two miles away, it doesn't mean you can't communicate. If you have a drum made out of a tree trunk that's been made hollow or some ceramics or something, and you beat that drum loudly enough, somebody's gonna hear it several miles away. So you have communication at the speed of sound. That was a very important, ahm, beginning of the development of rhythm in Africa, because these various rhythms could signify different emotions or different facts over its large distances.

The other of the two important ways in which, or two important fields in life, the music took in Africa was ceremony. So if you look at communications in ceremony, dances, funerals, births, places like that, people would sing. Now, ah, of course, when the blacks were brought over here, were they brought as teachers? Were they brought as instructors? Were they brought as musicians? What were they brought as? As slaves. And one of the first things the white masters learned to do to control these people more effectively was to strip away their cultures. They could not, however, strip away their desire to sing, their desire to communicate. So although their native languages, their native songs, their native religions were banned, they were not prohibited from singing. Now most of them were working in the fields, tobacco, cotton, other agricultural pursuits. One of the earliest form of music that is uniquely American, was the field holler. There'd be a crew of blacks working in a field and a leader, and they would begin to work in rhythm. You know how, if you go jogging, how you, it's nice to have a rhythm, it makes the work easier. If you are working with a shovel or broom, if you have a certain rhythm, it makes it easier. They began to sing and generate a new form of music in the fields. These were called field hollers, and the boss or the leader may yell out, "Morning time." And they'd all yell back, "Morning time." "Time for supper." "Time for supper." and they'd call back and they'd have variations, improvisations and rhythm.

Another aspect in the lives in which music developed, of course, was religion. As it had in their homes and most of their, their religions, most of the people were prohibited from, ah, from exercising their native religions. So they had to exercise their religions that they were allowed to, which includes gospel music. So they were learning American gospel hymns and, of course, they would eventually take those melodies and take the rhythms, and fuse them together and take the songs, as the gospel songs, and change the words, change the direction. What do gospels concern themselves with? God, religious aspects. Well, what would happen in the development of the, of the, this early music which we eventually began calling Blues? Instead of looking at God, they looked at their daily concerns. Woman, singing about a woman, man, child, romance, troubles,
problems. If you were a slave in the 19th century, you wouldn’t wanna be singing about life at school, you’d wanna be singing about the problems in your life, you’d be singing about your blues. These came together and we began to call it the blues. B, L, if I can write this, co., correctly, U, E, S, blues.

I have to move very quickly in order to give you an overview of the things we’ll be talking about in the next few days. As, as slavery ended and as the black people began to move into the cities, they’re, be, they’re ( ) and began to be educated and go to school. They began to learn some musical instruments and began to learn the instruments of white people. They began to learn the orchestral instruments. And one of the most important centers for the development of black music was New Orleans which was a major center of commerce, of world commerce and culture. New Orleans, where the Mississippi River which before the days of airplane and ocean and, large ocean going, ah, liners, it was the place where the whole central part of United States and the east coast, ah, ah, eastern, ah, part of countries on, on the inland side of the mountains, all their commerce went there and interfaced with the world. They had influence also because of the communication with the Caribbean and south and central America.

Now in New Orleans, the music began to be called Dixieland music. The musicians now, they were going to school and they’re learning to play and they were using in their bands instruments, such as the tuba, the trombone, the trumpet, the clarinet, and for rhythm, banjo. Louis Armstrong was perhaps the most important and influential musician from the New Orleans era. This music was called Dixieland. Later, after World War I, there was a massive migration of black people, north to industrialized cities to find work. As they were becoming more sophisticated and more entered into the mainstream of American society, through this move north.

Their instruments changed and because of the different lifestyles, their instruments changed as well. In Chicago, they weren’t playing outdoors in the street like they were in New Orleans. It was too cold. They were playing indoors. Banjo was replaced by the piano. The tuba was replaced by the string bass. That’s much more of an orchestral instrument. The music became more sophisticated. Instead of being largely a music of improvisation, and vocal, and singing, it became also a compositional music. Because as musicians were graduating from music schools, learning techniques and composition, they would apply these techniques to their music, to the jazz. So Chicago jazz, much more sophisticated than the south New Orleans jazz, it’s, it’s perhaps best remembered by the big band era, the large groups, the orch, like jazz orchestra. We’ll go on next week and talk about development of bebop and, mo, modern jazz, and post-modern jazz. Thanks for your attention and I’ll see you next week.
Modified scripts:

Lecture 1 "Borders"

Humanities is the interdisciplinary study and appreciation of creative human endeavors. In this Humanities class, I’ll talk about a poem written by Pat Mora. Pat Mora’s poem, "Borders," is about communication problems.

She starts with perhaps the most obvious need for translation between different languages. Mora is from El Paso, which is on the border of Mexico and the United States, so I think it’s not surprising that she chooses an example of translating between Spanish and English. When we go from one language to another, some problems may arise. Some words are not even translatable. Perhaps you know a word in a language and there is no equivalent in another language. Perhaps there is kind of an equivalent, but maybe the connotations are different.

Let’s look at that word "connotations." Connotations refer to the feelings and associations of a particular word. For example, let’s say someone, my height, weighs 200 pounds. There are a lot of different ways we might describe such an individual. Perhaps our attitude toward that individual would determine the connotations of the word we would use to describe that individual. For instance, we might say, "she’s kind of stocky" or "she’s fat" or "she is obese" or "she is a little bit overweight."

I’ll describe the same individual, but the different connotations reflect the attitudes toward that individual. When we translate from one language to another, we’re quite conscious of the fact that the connotations of words may not be identical. But since we are conscious of that, we can be careful with it. We’re less likely to run into difficulties because we’re aware of it, ironically, even though it’s in two totally different languages.

Her second example is a difference in languages we might not be quite so aware of between child talk and adult talk. I’ll give you an example from my personal experience. My little daughter sometimes comes to me when I am working on my computer at home and I hear her coming down the hall. She comes in and says, "Mom, Mom, will you give me a drink of juice?" So I walk into the kitchen, get her a drink and put it on the counter for her. She doesn’t even take a sip. What she really wanted was my time and attention. She didn’t care about getting a drink of juice at all. So I have learned to translate "I want a drink of juice" into "please give me a little time and attention."

I think that Mora’s example also points out an internal dialogue that each one of us has, regardless of our culture, between that part of ourselves which is the child, Freud would say that "id," in dialog with the adult, the "superego," as Freud would say. How have we reconciled the demands of one against the attitude of the other? For example, the child might say, "I want another piece of cake" or "I want that vacation" or "I want that new car" or "I want your love." But adult part of ourselves says, "Do I have to work for these things?" So we have to have this ongoing dialogue and the ego resolves it, or the
adult may resolve it with the parent and child within each one of us, arguing back and forth. So this is a border even within ourselves. I think a good poem like this one prompts additional examples. It's thought provoking.

Her third example is the most powerful. It's the climax of her poem, communication between intimates. Talk, talk, talk, but they are not communicating. They are using the same words. The problem is that words don't mean the same thing to everyone.

She uses two examples. Success. What does success mean? It depends on who's using the word. Or happiness. Let us think for a moment of any words that might possibly mean something different, perhaps if a man uses it or if a woman uses it. Now I don't want to stereotype here. Think of any general things that might cause some differing definitions and some problems in communication. For example, honor. What is honor for a man? What is honor for a woman? What is honor for a country? I think words change meaning over time. That's good example of a word that changes over time. Let's look at how all language changes. Spoken language is constantly evolving. Just as fashions change and hairstyles change, words change in the way people use them. And if you isolate speakers, the changes are going to be in different manners. Speakers might be isolated due to ages, due to geographical locations, due to social classes or educational levels, or due to experiences growing up as a man or a woman. Words are very slippery things. A word is not a word, but a word is something used by people and we try to communicate with them. We have to be aware, though.

There are many borders in communication and Mora's poem brings us to our attention that a word is not all that simple. It doesn't do all the communicating. We have to go beyond borders. Also borders can help to define who we are. A border says, "This is me" and "This is everyone else" and "This is the rest of the world." Perhaps in some cultures, border definition is quite different. In America, we emphasize the self, the individual, "Give me my space." But in some cultures, there's much more emphasis on fitting in with the group, subordinating the self for the good of the community. I think being aware of that kind of personal border can help to enhance the possibility for appropriate communication.

You might think about times in your life when borders have made a big difference. Perhaps you recognized a border that you had not even realized was there. Or perhaps you created a border. Maybe you eliminated a border. Maybe you crossed over a border. Now I am talking in the figurative sense, not the literal sense.

In this humanities course, as we study more literature, I would like you to keep this in the back of your mind. And as we read works by a variety of people from all over the world, I want you to see, if you can, that particular characters have bumped into a border, or overcome a border, or perhaps need to redefine their own borders. I like her metaphor and I think we'll find it applicable not only in this literature, but in our lives as well. If you have any questions, please do come see me during my office hours. And continue the reading, and I look forward to seeing you in the next class if not in my office before then. Thanks.
Lecture 2 "How Food Chains Work"

During the last class, I talked about ecosystems which are giant assemblages of plants and animals all interacting together, such as a redwood forest or tropical rain forest. Ecologists study interaction between organisms in an ecosystem and what happens if any part of the ecosystem is altered.

Today, I'd like to talk to you about food chains and how they work. In order to do that, I need to give you some background information on some of the components of the ecosystems. What sorts of things are interacting in the assemblages?

We start out with the most important organism in the ecosystem which turn out to be the autotrophs. Now that's probably a foreign word, but one of the things I like to do with students is to have them translate biology because often times biology is a language in and of itself. If we take this word apart, what we find is auto here, referring to self, like automobile meaning self moving. And troph always refers to feeding. So we have self-feeding organisms, or self-feeders. Now, who can feed themselves? What organism can feed itself? It turns out green plants can do that. If you go into an ecosystem like a redwood forest, everything you see there that is green can actually feed itself. This includes the plants.

Now we have to ask ourselves, how do they feed themselves? They contain a unique pigment called chlorophyll, chlorophyll pigment. This pigment is able to transform sunlight energy directly into food. Our very existence depends on this process. To give you a visual idea of what that would look like, try to imagine a huge leaf in this room here and imagine radiant sunlight energy landing on that leaf and sugar cubes popping out. As the sunlight energy flows through the leaf, you'll see sugar cubes coming out. These leaves are actually food making factories. They transform radiant sunlight energy into food and all of the organisms in the entire biosphere of this planet are dependent on that process. That process in biology is called photosynthesis. Now we can again take the word apart. Synthesis means to make, and photo refers to from sunlight or light. So we're making food from sunlight. That's what photosynthesis refers to. Autotrophs are the key components in the ecosystem because, without them, there'd be no food.

The next group are the heterotrophs. While homo means the same, like a homogeneous crowd, hetero means variety or mixed. And trophs means feeders. So heterotrophs mean mixed feeders and these include all the organisms that directly or indirectly are dependent on the work of the autotrophs for their food supply. This would include all the animals. When you think about how you eat, we're animals, and when you think about what you eat, it's either plant life directly or something that was converted from plant life. For example, if you eat a piece of chicken, the chicken came from eating corn, and the corn came from a plant. Everything we eat directly or indirectly comes from plant life, and that's true with all the heterotrophs on the plant, which include basically all of the animals and all of the fungi. So, the heterotrophs are totally dependent
on the autotrophs. In the human population, we have almost 6 billion people on the planet and most of those people are dependent on 3 crops: rice, wheat, and corn, all of which are autotrophically based food supplies.

Then a third group are the decomposers. In any ecosystem, autotrophs and heterotrophs are constantly manufacturing waste products. Also, they die. What do we do with all of this waste and all of these dead bodies? It turns out that these dead bodies and waste products are full of nutrients. You know that if you put manure in a garden it makes your plants really thrive. So how do we get the nutrients in those dead bodies back into the ecosystem in a form that can be used to generate new life? The answer to that question is that the decomposers do the work. The decomposers consist primarily of 2 groups of organisms. These are bacteria and fungi. Bacteria and fungi take the waste products that these organisms produce, the dead bodies, pull all of the important ingredients out of those bodies, and put them back into the environment so that they can be used by plants to make more food, or else they are put into the soil. The existing plants can then absorb the materials through their roots and create new plant growth. The decomposers are providing the autotrophs with more of the things they need to continue producing food for the whole ecosystem. So these are the key living players: autotrophs, heterotrophs, and decomposers.

Now, obviously, these organisms don't live in a vacuum. They're actually connected with the whole physical environment. In the environment, we can look at a set of what's called abiotic, or non living factors. And again, we can take this word apart to understand it. Bio means living or life. A means without. So what we're talking about are substances or features in the environment that are non living. What are some non living features that an organism would interact with? We can feel that right now we're interacting with temperature, the amount of humidity in the room, the amount of sunlight in the environment, how much snow falls, how much precipitation and what kind of soil there is. All of these are types or features of the environment. They are not living organisms.

In ecosystems, we have these key components which are all interacting together to form a very integrated whole. So how do they feed upon each other? How do food chains work? We're now to our last point here. To look at how food chains work, we should start with the autotrophs because they produce food every day. Because they produce food for everybody, autotrophs are often times referred to as primary producers. So, we can start a food chain with our autotrophs at the base. Now let's just make one step up—a real simple one. Let's say that you're going to have steak for dinner. What we're gonna do is start out with where the steak came from. The steak come from a cow. The cow got its food from the grass which is a primary producers. So we'll have a little simple food chain with grass, which is at the primary producer level.

And then we're gonna have a cow come along and a steer come along, and we're gonna have the steer eat the grass and then, above them, you're gonna have steak dinner. Now, I made this food chain kind of pyramid shaped rather than box shaped. And what we are gonna look at is why is it that there's not as many people on the planet as there are autotrophs or grassy areas. Why is it that our food chain is pyramid shaped instead of box shaped where we'd have an even number organism in each level? An answer to that
question is that every time you convert one form of organism into another, you lose 90% of the potential energy that's stored in that particular type of organism.

Sometimes, some of the concepts in biology are complex, so now I'll give you a good analogy which will make it clear. Let's pretend like we have a little cow here and we want it to gain one pound in a day. So we weigh our little cow in the morning and send the cow out into the meadow. The cow's running around. At the end of the day, we bring it back and keep track of every single thing that the cow eats and all of the weight it puts onto its body. If it weighs one pound more, we look and check on how much food that cow actually ate and it turns out that the cow ate 10 pounds of grass to gain the one pound of weight on its body. So that's essentially saying that if we have a little cow here, and we want it to gain one pound, we have to have 10 one pound sacks of grass. We open the cow's mouth, shove one after another into the cow, and then after a period of time for digestion, we weigh the cow and it's gained one pound.

So every time you convert them, the grass into the steer, or the steer into the human, you lose 90% of the energy, because most of the energy goes into just maintaining the body. The cow doesn't just sit there. Its heart is beating and the cow has to produce heat to maintain its body temperature. If it's running around swatting flies and chasing other cows, then it's gonna lose a whole lot of energy.

So, most of the energy in the food it eats goes into maintaining the body. Only 10% is actually stored as weight. For that reason, food chains are pyramid shaped rather than box shaped. They are the basis for understanding the complex relationships between organisms in ecosystems, simply because you have to have a plant base that's large enough to support the organism at the top of the food chain.

Lecture 3 "Origins of Jazz"

Today, we are going to talk about the development of a music which is uniquely American, and which is influenced by many cultures from around the world, both in its beginnings and today. That music is jazz.

The four most important elements of jazz are improvisation, rhythm, lifestyle and composition involving performance. Whereas classical music is the composer's art, jazz is the performer's art.

Jazz music is primarily only a music of the black voice. Jazz sprang from the experiences and the lifestyles of black people through their voices.

Now before the blacks that came over here were taken from their homes in Africa, they were already playing music mostly in two different parts of their lives. One was communication. Africa is a very difficult place to live. In many parts of the huge continent, there are vast distances between places. One of the ways people could communicate was through sound. If you can't see somebody because they're two miles away, it doesn't mean you can't communicate. If you have a drum made out of a tree trunk that's been made hollow or some ceramics or something, and you beat that drum loudly enough, somebody's gonna hear it several miles away. So you have communication
at the speed of sound. That was a very important beginning of the development of rhythm in Africa. These various rhythms could signify different emotions or different facts over its large distances.

The other of the two important ways that the music was used in Africa was ceremony. If you look at communications in ceremonies, dances, funerals, births and things like that, people would sing.

When the blacks were brought over here, were they brought as teachers, or as instructors, or as musicians? No, they were brought as slaves. One of the first things the white masters learned to control these people more effectively was to strip away their cultures. They could not, however, strip away their desire to sing, their desire to communicate. So although their native languages, their native songs and their native religions were banned, they were not prohibited from singing.

Since most of them were working in the fields, tobacco, cotton or other agricultural pursuits, one of the earliest forms of music that is uniquely American was the field holler. There’d be a crew of blacks working in a field and a leader, and they would begin to work in rhythm. As you know, if you go jogging, it’s nice to have a rhythm and it makes the work easier. If you are working with a shovel or broom, and you have a certain rhythm, it makes it easier. They began to sing and generate a new form of music in the fields. These were called field hollers. For example, the boss or the leader may yell out "Morning time" and they’d all yell back "Morning time." "Time for supper." "Time for supper." They’d call back and forth and they’d have variations, improvisations and rhythm.

Another aspect of their lives in which music developed was religion. As it had been in their homes and most of their religions, most of the people were prohibited from exercising their native religions. Instead, they were allowed to learn gospel music. Gospels concern themselves with God, religious aspects. They were learning American gospel hymns. But they would eventually take those melodies and rhythms, fuse them together, take the songs, as the gospel songs, and change the words and the direction.

What would happen in the development of this early music which we eventually began calling Blues? Instead of looking at God, they looked at their daily concerns. Woman, singing about a woman, man, child, romance, troubles and problems. If you were a slave in the 19th century, you wouldn’t wanna be singing about life at school, you’d wanna be singing about the problems in your life. You’d be singing about your blues. These came together and we began to call it the blues.

I have to move very quickly in order to give you an overview of the things we’ll be talking about in the next few days. As slavery ended and as the black people began to move into the cities, they began to be educated and to go to school. They began to learn some of the musical instruments of white people. They also began to learn the orchestral instruments. One of the most important centers for the development of black music was New Orleans, which was a major center of world commerce and culture. Before the days of airplanes and large ocean going ships, New Orleans, near the Mississippi River, was the place where the whole central part of United States and the eastern parts of some countries on the inland side of the mountains sent all of their commerce and interfaced with the world. They also had influence because of the communication with the Caribbean
and South and Central America.

In New Orleans, the music began to be called Dixieland music. The musicians were going to school and they were learning to play and they were using instruments in their bands such as the tuba, the trombone, the trumpet, the clarinet, and for rhythm, the banjo. Louis Armstrong was perhaps the most important and influential musician from the New Orleans era. This music was called Dixieland. Later, after World War I, there was a massive migration of black people, north to industrialized cities to find work. They were becoming more sophisticated and more integrated into the mainstream of American society through this move north.

Because of the different lifestyles, their instruments changed as well. In Chicago, they weren't playing outdoors in the street like they were in New Orleans. Because it was too cold, they were playing indoors. The banjo was replaced by the piano. The tuba was replaced by the string bass. That's much more of an orchestral instrument. The music became more sophisticated. Instead of being largely a music of improvisation and singing, it became also a compositional music because many musicians were graduating from music schools, learning techniques and composition and they applied these techniques to their music, to jazz. Chicago jazz, much more sophisticated than the south New Orleans jazz, is perhaps best remembered by the big band era and the large groups, like the jazz orchestras. Next week we'll go on and talk about the development of bebop, modern jazz and post-modern jazz. Thanks for your attention and I'll see you next week.
APPENDIX D. SCORING CRITERION IN EXPERIMENTAL TESTS

Listening comprehension in academic lecture

Construct definition: Ability to comprehend extended academic lectures through authentic and interactive listening tasks

<table>
<thead>
<tr>
<th>Levels of ability</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No evidence of comprehension</td>
</tr>
<tr>
<td></td>
<td>No response, irrelevant information</td>
</tr>
<tr>
<td>1</td>
<td>Limited evidence of comprehension</td>
</tr>
<tr>
<td></td>
<td>small piece of information involving</td>
</tr>
<tr>
<td></td>
<td>requested item performance</td>
</tr>
<tr>
<td>2</td>
<td>Extensive evidence of comprehension</td>
</tr>
<tr>
<td></td>
<td>not complete, but satisfactory information</td>
</tr>
<tr>
<td></td>
<td>for a given task</td>
</tr>
<tr>
<td>3</td>
<td>Evidence of complete comprehension</td>
</tr>
<tr>
<td></td>
<td>well-organized requested information for a</td>
</tr>
<tr>
<td></td>
<td>given task</td>
</tr>
</tbody>
</table>