Video Listening Tests: A Pilot Study

Elvis Wagner[1]
Teachers College, Columbia University

ABSTRACT

Listening has long been the neglected skill in second language acquisition research, teaching, and assessment. However, in recent years there has been an increased focus on L2 listening ability because of its perceived importance in language acquisition. The present study explored the listening process when the aural input was delivered through the use of video. Video texts were used because video allows listeners to perceive and process nonverbal information. A model of L2 listening ability was hypothesized and operationalized, and an assessment instrument was created. This video listening test was administered to 85 ESL students. The data from this test were then analyzed using reliability analyses and Exploratory Factor Analysis (EFA). The results seem to provide some evidence for the validation of a two-factor model of listening based on the ability to comprehend explicitly stated information, and the ability to comprehend implicit information in aural texts.

INTRODUCTION

For decades, tests of second language reading, writing, and speaking have garnered large amounts of attention, research, and resources in the quest to create reliable, valid, and practical assessments. Listening, however, has traditionally been the forgotten skill when it comes to testing (Douglas, 1988). Buck (1991) attributes this neglect to the lack of a widely-accepted theory of listening comprehension, and goes on to state, "It seems that in practice test constructors are obliged to follow their instincts and just do the best they can when constructing tests of listening comprehension" (p. 67). Obviously, this haphazard approach to testing listening presents serious implications for the validity of these assessments. Fortunately, in the last decade the assessment of second language listening has attracted increasing amounts of attention, and a great amount of research has been conducted on the subject.

Numerous researchers (e.g., Buck, 1991, 2001; Buck & Tatsuoka, 1998; Dunkel, Henning, & Chaudron, 1993; Richards, 1983; Rubin, 1994), have described the necessity of defining the concept of second language (L2) listening comprehension, yet an adequate definition is still elusive, and there seems to be a general consensus that there is no widely-accepted definition (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Brindley, 1998; Buck, 1994, 2001). Part of the problem lies in the fact that because so many different processes and aspects are involved in L2 listening comprehension, providing a global, comprehensive

[1] Elvis Wagner is a doctoral student in Applied Linguistics at Teachers College, Columbia University. His area of interest is language testing, with a special emphasis on the testing of second language listening. Correspondence concerning this article should be sent to Elvis Wagner, 212 W91st St., #316, New York, NY 10024. E-mail: elviswags@yahoo.com.

definition may be impossible. Richards (1983) describes how L2 listening varies according to what learners are listening for (social interaction, information, academic listening, listening for pleasure, or for some other reason). Also, the process of L2 listening varies with the level of the learner (Brown, 1986; Buck, 1994; Hale & Courtney, 1994; Shohamy & Inbar, 1991), and the context of the situation (Buck, 2001).

Also adding to the difficulty in formulating a widely-accepted definition of L2 listening ability is that a number of researchers (e.g., Bachman & Palmer, 1996; Benson, 1989; Canale, Child, Jones, Liskin-Gasparro, & Lowe, 1984) consider the attempt to deconstruct language ability into four skills, and distinguishing these skills in terms of channel and mode, as misguided and inadequate. Bachman and Palmer (1996) argue that it is much more useful to see language use being realized as learners performing specific language use tasks. They state:

> We would thus not consider language skills to be part of language ability at all, but to be the contextualized realization of the ability to use language in the performance of specific language use tasks. We would therefore argue that it is not useful to think in terms of 'skills', but to think in terms of specific activities or tasks in which language is used purposefully. (p. 75-76)

Rather than considering listening to be a "skill", they see it as a combination of language ability and task characteristics. Thus, when designing and using a test, it is necessary to define these listening language use situations in terms of their task characteristics and the language ability and topical knowledge needed to perform them (Bachman & Palmer, 1996). Bachman and Palmer refer to the importance of *authenticity* and *interactiveness* in creating tests that are construct valid. Creating test tasks that have characteristics similar to those tasks in the target language use (TLU) domain gives the tasks authenticity. Creating tasks that require the test-taker to integrate his or her topical knowledge (and affective schemata) with language ability in order to successfully complete these tasks makes the tasks interactive (Bachman & Palmer, 1996). Traditionally, many listening tests have lacked authenticity or interactiveness (Buck, 2001). Many traditional listening tests lack authenticity in that they often utilize texts that are inauthentic; the texts share few characteristics with spoken language representative of the TLU domain. Many listening tests tasks lack interactiveness in that they fail to require listeners to integrate their background knowledge with their language ability. These listening test tasks might require the involvement of a listener's ability to phonologically decode oral input, or knowledge of the sound system, but fail to include communicative aspects of a test-taker's listening ability such as listening for the global message, or interpreting a speaker's pragmatic meaning.

The purpose of the current study is to examine the construct validity of a listening test based on a model of L2 listening ability that treats listening as a complex combination of language ability and task characteristics. The listening test examined has academic listening as its TLU domain, and was specifically created to include task characteristics of the TLU domain. In addition, the listening text used in the test examined here includes both aural and non-verbal input through the use of video in delivering the spoken text. This was done in an attempt to make the characteristics of the test tasks representative of the TLU domain, to minimize sources of invalidity, and to avoid construct under-representation (Messick, 1989, 1996).

In this paper, I will first describe how researchers have defined second language listening ability. I will then briefly discuss three factors that can affect comprehension in L2 listening

tests. After that, I will describe a listening test that was developed based on the operationalization of an L2 listening model. The data from this test will then be analyzed and discussed in relation to the theorized model, and the necessary revisions to this model. Finally, I will describe some areas in which further research is needed. The study addresses the following research questions:

1. What is the nature of second language listening ability as measured by a listening comprehension test delivered through the use of video?
2. To what extent do the scores from the assessment provide evidence for the construct validity of the theoretical model?
3. Is there evidence of a test method effect caused by items designed with specific question types (multiple-choice versus limited-production)?

Review of the literature

Perhaps because of the inherent difficulty in providing a comprehensive definition of L2 listening, a number of researchers have created taxonomies of the listening comprehension skills or operations (Aitken, 1978; Buck, Tatsuoka, Kostin, & Phelps, 1997; Buck & Tatsuoka, 1998; Lund, 1991; Petersen, 1991; Richards, 1983; Weir, 1993). Richards (1983) created a taxonomy of 33 micro-skills related to conversational listening (e.g., ability to recognize stress patterns, ability to distinguish word boundaries, ability to detect sentence constituents), and 18 micro-skills related to academic listening (e.g., ability to identify purpose and scope of lecture, ability to infer relationships, ability to recognize markers of cohesion). Buck and Tatsuoka (1998) used rule-space methodology to list 15 prime attributes (e.g., ability to scan fast spoken text, ability to process large information loads, ability to understand and utilize heavy stress) and 14 interaction attributes (e.g., ability to make text-based inferences, ability to process text automatically) that explained 96% of the variance for 96% of the students involved in a listening comprehension test. While these taxonomies are important in the process of defining L2 listening comprehension, their use is somewhat limited because "few of these valuable efforts have attempted to provide clear definitions or non-redundant orderings of components in any systematic graded hierarchy" (Dunkel, Henning, & Chaudron, 1993, p. 182). Also, these taxonomies are essentially hypothetical in nature, and there has been little empirical investigation (Buck, 2001).

Buck (2001), while describing how the purpose and TLU situation should determine the appropriate construct of listening to be used in the test, gives a list of recommendations to be used when creating a listening construct, which he refers to as his "default listening construct" (p. 113). This default construct includes: focusing on the assessment of those skills that are unique to listening; testing listeners using a variety of texts on a variety of topics; using longer texts that test discourse and pragmatic knowledge, and strategic competence; going beyond literal meaning to include inferred meanings; and including aspects dependent on linguistic knowledge, while excluding aspects that are dependent on general cognitive abilities. Buck (2001) also gives a more formal definition of his default listening construct. It is the ability to (a) process extended samples of realistic spoken language, automatically and in real time, (b) understand the linguistic information that is unequivocally included in the text, and (c) make whatever inferences are unambiguously implicated by the content of the passage (p. 114). This

default listening construct is useful, in that it is broad enough to apply to most listening situations, yet flexible enough to be tailored by test creators to fit the context of the testing situation.

Listening as a Two-stage Process

Traditionally, listening has been divided into a two-stage process. Many researchers (Buck, 1991; Call, 1985; Conrad, 1989; Lund, 1991; Rost, 1990; Secules, Herron, & Tomasello, 1992; Tyler & Warren, 1987; Weir 1993) have posited this idea of listening as a two-stage process, although they often use different labels for the two stages or processes. Buck (2001) describes it as: "A first stage, in which the basic linguistic information is extracted, and then a second stage in which that information is utilized for the communicative process" (p. 51). He goes on to cite a number of researchers (Carroll, 1972; Clark & Clark, 1977; Rivers, 1966) that have hypothesized this two-stage process, and states:

> …these scholars seem to have arrived at similar conceptualisations of listening comprehension, and the fact that they use different terminology suggests that they have arrived at this understanding more or less independently. This adds considerable credibility to the two-stage view of listening. (p. 52)

Brindley (1998) describes the idea of identifiable listening skills, including lower order skills that involve understanding utterances at the literal level, and higher order skills like inferencing and critical evaluation. One of the most commonly cited descriptions of listening involves the idea of both top-down and bottom-up processing. Kelly (1991) describes bottom-up processing as the process in which the listener receives the input as sound and begins to interpret the meaning. The top-down process involves "…the application of cognitive faculties in the attempt to give the sound input meaning. The mind sets up the expectations and the sound provides confirmation" (p. 135). When enough information arises from both sources, then perception occurs. Thus, both types of processing occur simultaneously (Buck, 2001), although the contribution of both types is not necessarily constant and equal over the course of an utterance. Kelly (1991) states that when the text and words are highly predictable, the listener does not need to rely much on bottom-up processing. When the listener's expectations are low, however, he or she is forced to use the sensory level bottom-up processing. Because the words and texts are rarely predictable for beginning ESL listeners, they usually have low expectations of the upcoming spoken input, and thus are forced to rely mostly on bottom-up processing.

This idea that learners with varying levels of proficiency process aural input differently is found throughout the literature. A large number of studies (Baltova, 1994; Blau, 1990; Brown, 1986; Buck, 1994; Chiang & Dunkel, 1992; Conrad, 1985; Hale & Courtney, 1994; Hansen & Jensen, 1994; O'Malley, Chamot, & Kupper, 1989; Shohamy & Inbar, 1991; Wu, 1998) have found evidence for this belief. Conrad (1985) found that as the ability of L2 learners increased, their processing showed progressively greater attention to semantic rather than to syntactic or phonological cues. With decreasing proficiency, listeners had to base their expectations of the message on cues closer to the surface of the language. Hansen and Jensen (1994) compared the results of scores (based on varying question types) between different level learners. They found that lower level learners had more difficulty in comparison to higher level learners on broad, global questions (representing the need for top-down processing), than they did with detail

questions (representing bottom-up processing). The researchers found evidence that lower level listeners relied on verbatim responses in answering questions. This worked well for detail questions, but was less effective for global questions.

Authenticity and Construct Validity

In order for the results of a test to be generalizable to non-test language situations, the tasks on the test must be sufficiently representative of the TLU domain (Messick, 1996). Creating authentic test tasks (i.e., those that are representative of the TLU domain) is important because of the role authenticity plays in contributing to construct validity (Bachman & Palmer, 1996). Bachman and Palmer define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (p. 23). If a test task (including the text used in the task) is authentic and corresponds closely to the TLU task, then it allows test users to generalize the test scores beyond the test itself, to similar non-test language uses, and "this links authenticity to construct validity, since investigating the generalizability of score interpretations is an important part of construct validation" (p. 24). Bachman and Palmer advise that when designing an authentic test task, the critical features of the TLU domain should be defined first, and then the test tasks should be designed so that they have these critical features.

Messick (1989, 1996) also describes how authenticity can contribute to construct validity. In this respect, one of the aims of creating a valid assessment should be to avoid construct under-representation. Authenticity can help a test designer meet these aims, since authentic tasks that have realistic settings or close simulations of real life language use should minimize sources of invalidity. In addition, authentic tasks should be selected that "provide representative coverage of the content and processes of the construct domain" (1996, p. 250). If authentic tasks are used that are sufficiently representative, then the score interpretation of the assessment should be generalizable to non-test language situations.

The use of authentic tasks should also serve to minimize sources of invalidity in a test (Messick, 1989, 1996). Bachman (1990) describes how a test-taker's test performance is influenced by the characteristics of the methods used to elicit the test-taker's language performance. In other words, the way in which these "test method facets" are designed and controlled has a great impact on the test-taker's performance. Bachman, (1990) and Bachman and Palmer (1996) cite numerous studies that provide evidence of the effect of test method on test performance. Bachman (1990) developed a framework to delineate the specific features or facets of test method that can affect test performance. His framework has five categories of test method facets[2], including "facets of the input", which will be focused on here.

Bachman and Palmer (1996) build on and slightly revise Bachman's (1990) framework (they use the term "task" in place of "test method", and "characteristics" in place of "facets"). Bachman and Palmer (1996) state that the task characteristics are always going to affect test scores to some extent. Since it is impossible to eliminate the effects of task characteristics, it is necessary to control them as much as possible so that the tests will be appropriate for what they are used for. The goal, then, is for test developers to understand and be aware of what characteristics can be varied, and how they can be varied to best tailor tests to make them appropriate for specific test-takers. Their framework of task characteristics has three sections,

[2] The other four categories are: the testing environment, the test rubric, the expected response, and the relationship between input and response.

including "characteristics of the input" (the other two sections are "characteristics of the setting" and "characteristics of the test rubrics").

For listening tests, some of the most important characteristics of the input that might affect test-taker performance include: the type of text, topical knowledge, the amount or level of speededness of the text, text length, the role of nonverbal communication, video (as opposed to audio-only) texts, question type, question preview, and the number of times the text is presented. Other aspects of the input might also affect test-taker performance on a listening exam, but three of the most salient features of the input, that are most relevant for the current study (type of text, the role of non-verbal communication, and question type), are briefly examined here.

Factors Affecting Comprehension in Second Language Listening Tests

Type of Text

Tannen (1982) describes how texts can be seen as ranging on a continuum from oral to literate, with one end of the continuum having texts with distinctly oral features, to texts that are planned and written and then read orally on the other end. A study which sought to address this issue of read vs. spoken texts in listening assessment was conducted by Shohamy and Inbar (1991). They compared the listening comprehension proficiency according to different types of text that varied on their "degree of orality" or "listenability". They found that the degree of orality of the text significantly affected test scores. The more listenable the text, the better the test-takers scored. The dialogue text was the easiest, then the lecturette text, and the newscast text was the most difficult. Dunkel (1988) reached a similar conclusion with her study. Selecting specific types of texts that depend on the purpose for using the tests and that are representative of the TLU domain should result in L2 listening comprehension tests that have content and construct validity.

The text of a listening comprehension assessment is a critical aspect in regards to the construct validity of that assessment. Historically, many listening comprehension assessments have used texts that were written and read aloud. While this might be representative of certain aspects of the TLU domain, such as radio and television broadcasts, it is less representative of the academic listening domain. A text that is written and read is inherently different than a text that is extemporaneously produced and simultaneously spoken. A number of researchers (Flowerdew, 1994; Hadley, 2001; Lund, 1991; Tannen, 1982; Shohamy & Inbar, 1991) have expounded on the differences between the two types of texts.

Spoken language often differs from written language because spoken language is generally produced extemporaneously, and thus contains many more pauses, fillers, and redundancies than written language (Samuels, 1984). These pauses and fillers are important in listening comprehension, because they allow more processing time for the listener to interpret the input (Rubin, 1980). Similarly, redundancies in spoken texts can also give listeners more time to process the input, and they also serve to give listeners another chance to interpret the input if they missed it the first time. Numerous studies (Blau 1990; Cervantes & Gainer, 1992; Chaudron, 1983; Chiang & Dunkel, 1992; Conrad, 1989; Parker & Chaudron, 1987; Pica, Young, & Doughty, 1987) conducted with L2 listeners found that texts with redundant language were helpful for learners in comprehending aural input. Using written texts that are read for L2 listening exams deprives test-takers of increased processing time that pauses and fillers allow.

Written texts that are read also decrease redundant structures that are helpful for listeners for comprehending the input. These pauses, fillers, and redundancies are a natural part of spoken language, and are part of the TLU domain, and to exclude them in a listening text threatens the construct validity of that test (Messick, 1996).

The Role of Nonverbal Communication

An aspect of listening comprehension that is often unacknowledged in choosing texts (and their mode of delivery) for L2 listening testing is the role of nonverbal communication (Kellerman, 1992). Tyler and Warren (1987), in their study of local and global structure in comprehending spoken language, describe an aspect of nonverbal communication, prosodic structure, that is very important in being able to understand spoken language. An utterance's prosodic structure is closely related to its syntactic and semantic properties. The utterance is structured into a sequence of intonational phrases, and each intonation phrase is marked prosodically by a closing contour. The pronunciation of certain words (due to the application of phonological rules) might change within phonological phrases, but not across their boundaries. In addition, an intonational phrase will be marked only at phonological phrase boundaries. Tyler and Warren conclude that a listener's ability to recognize the prosodic structure of a language is as important as the listener's syntactic knowledge in chunking incoming discourse appropriately.

This has very important consequences for L2 listening comprehension tests that are not always acknowledged by test designers. If prosodic structure is as important as Tyler and Warren claim, then it must be part of the construct definition of tests purporting to assess listening comprehension. But written texts that are read have very different intonational and prosodic patterns than extemporaneous oral texts. While written texts that are read orally could be seen as part of a listening TLU domain, they are not the dominant or even major part of the domain, and thus it is important that test designers include a sample of texts that are representative of the range of listening texts that learners would encounter in the TLU domain.

Similar to the importance of prosody in listening comprehension, the kinesic behavior of the speaker can be helpful for the listener to recognize the components of the incoming text and so, to chunk the input appropriately. Kellerman (1992) defines kinesic behavior as "all movements of the body, both muscular and skeletal" (p. 240). Both Kellerman (1992) and Brown (1995) describe how a speaker's body movement and stressed syllables are linked. These movements are helpful for the listener because stress often coincides with items that are semantically salient, in that they often provide new information. Even without being able to hear the words, an observer can visually see where the stressed syllables occur. In a stress-based language like English, this kinesic behavior can aid the learner's recognition, and storage in short term memory (STM), of the aural input and help the learner to chunk it appropriately (Kellerman, 1992; Pennycook, 1985; Von Raffler-Engel, 1980).

Von Raffler-Engel (1980) argues that kinesic behavior plays another important role in L2 listening comprehension. She asserts that kinesic behavior is additional way in which language is redundant, in that gestures, facial expressions, and the visible stress patterns of the speaker serve to reinforce the linguistic message. When the risk of making speaking errors (and consequently hearing misrepresentations) becomes greater, gestures and other kinesic behavior increase. Von Raffler-Engel concludes that "Communication is multi-channeled and to reduce language to the sole channel of verbalization is not communicating in full" (p. 229). Brown (1995) and Rost (1990) describe how segmentally connected speech (which includes reduction)

results in marked morphophonological changes, and these changes are paralleled by a visible change in articulation. For the L2 listener, who might not be able to recognize and understand aspects of the spoken language, the kinesic behavior and non-verbal communication of the speaker may be particularly helpful in providing clues that will be of assistance in understanding the message and chunking the input appropriately.

The importance of kinesic behavior and non-verbal communication presents a challenge to L2 listening comprehension test designers who have traditionally relied on audio recordings of listening texts. Audio recordings preclude test-takers from exploiting the speaker's kinesic behavior and non-verbal communication to aid listening comprehension. If kinesic behavior and non-verbal communication do play an important role in listening comprehension, and numerous studies (Baltova, 1994; Dunkel, Henning, & Chaudron, 1990; Kellerman, 1992; Tyler & Warren, 1987; Von Raffler-Engel, 1980) have provided evidence that they do, then it is necessary to account for these factors in the construct definition of listening comprehension, and to design tests that take this into account. The obvious answer is the use of video media in listening comprehension assessments.

Although there is a large amount of material on using video media for pedagogical purposes, there is much less research on the role and importance of video for listening assessment (Progosh, 1996). Gruba (1997) reviews the role of video media in listening assessment, and provides reasons why it is advantageous to use video in listening tests. The use of video is theory driven, in that it allows for a construct definition of listening that incorporates both visual and verbal elements. Furthermore, it is pedagogically related in that video is commonly used in teaching. Progosh (1996) used video in assessing listening comprehension, and also surveyed students on their attitudes toward the use of video in listening testing. More than 92% of the students surveyed thought that it was a good idea to use video in listening tests. Almost 92% of the students also said they preferred video to audiocassette listening assessments. Interestingly, the results to his query as to whether students thought video tests were easier than audiocassette tests were inconclusive. Baltova (1994) also found that students enjoyed a video listening assessment better than one based on an audiocassette. Because video is so commonly used in teaching listening (and other skills), it seems that learners are comfortable with, accustomed to, and in favor of its use in testing.

Another way in which the use of video may be superior to audiotape in listening comprehension assessment is that language used in video texts may better mirror realistic discourse (Baltova, 1994; Longeran, 1984; Wilkinson, 1984; Willis, 1983). These researchers argue (although empirical evidence is lacking) that language in audiotape texts has to be more verbally explicit than real life language because it has to compensate for the lack of visual cues. Kellerman (1992) maintains that this verbally explicit type of discourse is unrepresentative of the TLU domain, describing how there is a misrepresentation of address behaviors in audiotape texts because speakers need to verbally identify themselves (artificially). These address behaviors are usually realized through non-verbal behaviors in authentic discourse. This compensation for the lack of visual cues could be seen as an introduction of sources of invalidity into the assessment.

Progosh (1996) and Gruba (1997) argue that because video is so commonly used in teaching, it should also be used in testing, as this will contribute to construct and content validity (Bachman, 1990). Gruba also argues that the use of video in listening tests is feasible. In developed countries virtually all language schools and institutions have access to video players and monitors. In addition, the advent of digital video recording and transmittal by computer makes video assessment even more feasible and workable. Bejar et al. (2000) acknowledge the

potential of video in creating tests with enhanced face validity and authenticity, and seek to explore its use in the listening assessment section of the TOEFL 2000. As noted earlier, if the results of tests are to be generalizable, the test tasks must be similar to and representative of authentic tasks in the TLU domain (Messick, 1996). If, as the above researchers have found, non-verbal information is an important and integral aspect of aural communication, it is necessary to include this type of information on listening tests. It seems highly likely that the use of video in testing L2 listening assessment will continue to grow, and it is an area that deserves increased attention and research in the hopes of creating more valid tests.

Question Type

The role of question type in tasks is another important consideration in L2 listening comprehension testing. Sherman (1997) claims that comprehension questions are the commonly accepted practice in listening exams, even though they are unrepresentative of the TLU domain. Ur (1984) maintains that comprehension questions are commonly accepted and have achieved "respectability" for no better reasons than that they are similar to content-subject tests, and because students are very familiar with them. Perhaps most importantly, comprehension questions are relatively easy to create, and economical to administer in large-scale testing (Sherman, 1997). But test-taker familiarity, ease of creation, and ease of administration do not alleviate the need to examine how exactly the task questions affect the listener's comprehension of the text, or the need to examine the assessment for test method effect (Bachman & Palmer, 1983; Bachman, 1990).

Buck (1991) examined the feasibility of writing L2 listening comprehension questions that test higher-level processing, and found that it was very difficult to write such questions. He operationalized the distinction between lower-level processing and higher-level processing, and attempted to create two distinct question types, "those which asked for information clearly stated in the text, and those which required testees to make inferences based on that clearly stated information" (p. 76). The questions did not perform as Buck had anticipated, however. He attributes much of this to the effect of the short-answer format, in that test-takers could give different answers to the same question, and thus questions meant to test lower-level processing sometimes had answers that required higher-level processing, and vice-versa. The data suggest that creating short-answer comprehension questions to test learners' higher level processing skills is a very difficult task, for a number of reasons. Still, Buck feels that with skillful item writing and test piloting, it is possible to do so.

Shohamy and Inbar (1991) also studied how the type of task question affected test-takers' scores. The task questions studied were of two types: those meant to assess overall/global comprehension, and those meant to assess specific/local comprehension. They found that specific/local comprehension questions were significantly easier for test-takers to answer correctly than overall/global questions, concluding that test-takers have more difficulty inferring and synthesizing information than finding specific information in a text. They also found that most students who answered global questions correctly were able to also answer local questions, but the opposite was not true.

Lund's (1991) study provided somewhat different results. In comparing reading comprehension with listening comprehension, he found that test-takers reading a written text were able to recall more details of the text than test-takers who listened to a read text. The listeners, meanwhile, were able to recall more main ideas than readers. However, a number of

caveats must be mentioned. Comparing Lund's and Shohamy and Inbar's study is not possible because Lund compared listening to reading comprehension, while Shohamy and Inbar focused only on listening. Also, Lund used the same text for both tests. As mentioned earlier in detail, using a written text that is read aloud for a listening comprehension presents a number of validity issues, although Lund chose a text that he deemed to be in the middle of the literate/oral continuum.

Wu (1998) studied the effect of the task question type (multiple-choice) on the scores of a listening comprehension test. He found four main effects:

1. Viewing the questions and options seemed to help the processing of information by higher level learners by helping to form anticipations of the input, and it provided foci for listening.
2. Misinterpretation of the options may have contributed to some test-takers choosing the wrong answers.
3. The multiple-choice format led to much uninformed guessing, because of too much dependence on non-linguistic knowledge, and through the "lure" of the distractors.
4. Uninformed guessing sometimes led to the test-takers choosing the correct answers, but for the wrong reasons. (p. 40)

Wu concludes that "while the MC format favours the advanced listener, it adds difficulty for the less able listener, and that, owing to its allowance for much uninformed guessing, the construct validity of the test is left open to question" (p. 38).

The studies reviewed here present an important beginning to the investigation of the test task method effect in L2 listening comprehension tests, and how this method effect influences the construct validity of the tests.

DEVELOPING A MODEL OF L2 LISTENING ABILITY

Identifying the Target Language Use Domain

As noted earlier, there is no standard definition or model of L2 listening ability, because the act of listening necessarily differs according to what listeners are listening for, the level of the learner, and the context of the situation. Therefore, in creating and operationalizing a model of listening ability, it is first necessary to identify, and select the TLU domain, and then to describe tasks representative of that TLU domain. There are a vast number of listening situations that test users could choose as the test domain for listening comprehension, including informal conversation, radio listening, television watching, telephone conversations, specific job settings, to name just a few. One of the most common domains used in the assessment of listening, however, involves listening in an academic setting, and, following the recommendations of Bachman and Palmer (1996) to focus on specific activities or tasks in which language is used purposefully, it was decided to focus on an academic listening TLU domain. Since the test-takers who participated in this study were high school students, the academic TLU domain seemed especially appropriate.

Selecting and Describing Tasks In the TLU Domain

In operationalizing a model of L2 listening comprehension with an academic listening TLU domain, many of the taxonomies and definitions of listening listed earlier (Aitken, 1978; Buck, 2001; Buck et al., 1997; Buck & Tatsuoka, 1998; Richards, 1983; Weir, 1993) were used to identify and select tasks representative of and relevant to academic listening. However, it must also be noted that there is in fact very little empirical support for these theorized listening taxonomies (Buck, 2001), and this study is exploratory in nature.

Many researchers (Buck & Tatsuoka, 1998; Richards 1983, Weir, 1993) include in their taxonomies and definitions the need for a listener to listen for specific details and facts in a text. This ability seems to be especially important when in an academic listening TLU domain. Similarly, Weir (1993) and Richards (1983) include in their taxonomies the necessity for a listener to have the ability to process longer segments of information in order to identify relationships among discourse units such as generalizations and supporting ideas. This also seems to be an aspect of listening very relevant to successful academic listening.

A listening skill commonly cited as important by researchers (Aitken, 1978; Richards 1983, Weir, 1993) is the ability to identify the purpose or main idea of the utterance, sometimes referred to as listening for gist, and its perceived importance is apparent by the seemingly automatic inclusion of items intended to assess this ability in listening exams. In addition, this skill is indicative of top-down processing (which will be discussed later), and thus it is necessary to include it in an operationalization of academic listening.

Creators of listening taxonomies have also recognized the importance of the ability to infer meaning from a spoken text. The ability to make inferences is a very wide-ranging (and consequently problematical) definition, and can include many different kinds of inferencing. Buck and Tatsuoka (1998) include low-level bridging inferencing, higher-level reasoning, or using background knowledge. Hildyard and Olson (1978) classify three different types of inferences: propositional inferences (those that follow logically from a statement in the text); enabling inferences (those related to causal relationships); and pragmatic inferences (those that rely on the non-literal interpretations of the speakers and the text). Buck, et al. (1997) and Buck and Tatsuoka (1998) describe a type of inferencing ability as "text-based." Text-based inferencing mirrors the propositional and enabling inferences described by Hildyard and Olson.

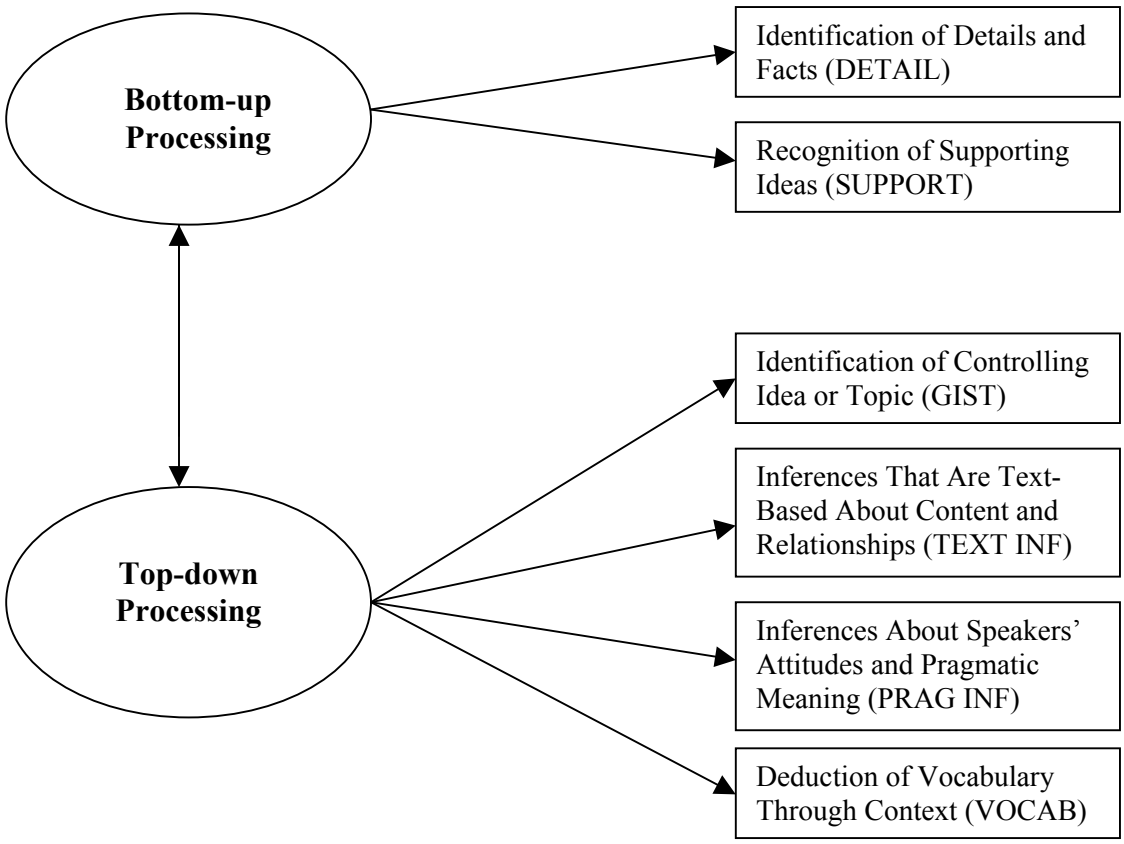
Another type of inferencing is the ability to make inferences about speakers' attitudes and pragmatic meaning. This is what Hildyard and Olson (1978) referred to as pragmatic inferences, and what Buck and Tatsuaoka (1998) refer to as inferencing based on background knowledge. This skill is cited as an important aspect of listening ability by other researchers as well (Aitken, 1978, Richards, 1983; Weir, 1993). Although superficially this might seem less indicative of an aspect of academic listening, the importance of speakers' attitudes and pragmatic meaning can never be discounted. In an academic setting, a speaker's attitude often is an indication of what he or she considers important, and thus it is necessary to include this aspect of listening in an operationalization of academic listening.

Finally, the ability to deduce meaning of unknown vocabulary through the context of the utterance is another skill that researchers have described as important in defining listening ability. Numerous researchers (Aitken, 1978; Richards, 1983, Weir, 1993) have included this skill in their taxonomies of listening ability. This would also seem to be an important ability in

an academic listening TLU domain. However, this ability is difficult to assess accurately. Most obviously problematical is the difficulty in assuring that one is testing the ability to infer meaning of unknown lexical items through the context of the passage, rather than assessing vocabulary knowledge. Nevertheless, because it is commonly cited in the literature as an important component of listening ability, and because it can be seen as an important aspect of an academic listening TLU domain, it is necessary to include this skill in the operationalization.

These six skills were used in the operationalization of an L2 listening model based on an academic TLU domain. However, because of the prevalence in the literature of the view that listening comprehension is a dual or two-stage process, it is also necessary to include this idea in an operationalization of academic listening. As noted earlier, the two processes are theorized to be occurring simultaneously, and thus they are interrelated. The interrelation of these two simultaneous processes is sometimes referred to as parallel processing (Rubin, 1994). Because these two processes are not directly observable, they must be measured through the observable skills that were theorized as constituting the academic TLU domain. Skills such as identifying details, facts, supporting ideas, and more local points of information are observable skills that seemingly constitute the latent ability to perform bottom-up processing. Global skills such as listening for gist, making inferences, and deducing vocabulary through the context of the text are generally considered the observable skills that constitute the latent ability to perform top-down processing. A graphic representation of the operationalization for this assessment is shown in Figure 1.

Figure 1
Operationalization of a Model of Second Language Listening Comprehension



However, aspects of this operationalization are problematic. If indeed listening requires simultaneous or parallel processing, it is inherently difficult to differentiate the levels of processing, or to attribute the responses on the test to any one skill or construct (Brindley, 1998; Buck, 1991, 2001). However, since this study is exploratory in nature, and because some construct definition of listening is necessary, this two-factor, six-skill model is hypothesized here.

MATERIALS AND METHOD

Study Design

An ex-post-facto correlational research design with exploratory factor analysis (EFA) was used as a primary analytical tool in this exploratory study to examine the validity of the L2 listening model. A model of L2 listening ability was hypothesized and operationalized. Changes were then made to the model based on these statistics and substantive rationale.

Study Participants

A total of 85 ESL students in three different public high schools in the Bronx, New York, participated in the study. The students ranged in age from fourteen to eighteen, and all were non-native speakers of English, living in the United States. The vast majority of the subjects had Spanish as their first language. Their level of proficiency was determined to be intermediate to advanced learners of English, based on their placement in their ESL classes.

The Assessment Instrument

The operationalization of the L2 listening model with an academic listening TLU domain that was shown in Figure 1 consisted of two latent factors (top-down and bottom-up processing). The skills "Identification of Details and Facts" (DETAIL) and "Recognition of Supporting Ideas" (SUPPORT) constituted bottom-up processing, while the skills "Identification of Controlling Idea or Topic" (GIST), "Inferences that are Text-based" (TEXT INF), "Inferences about Speakers' Attitudes and Pragmatic Meaning" (PRAG INF), and "Deduction of Vocabulary Through Context" (VOCAB) constituted top-down processing. For the assessment, 20 items were created to measure these different skills, with at least three items measuring each of the skills. An attempt was made to use both multiple-choice (MC) and limited-production (LP) items to measure each scale (as recommended by Berne, 1995, and Brindley, 1998), but this was not always feasible. The break down of the test items by skill and question type is shown in Table 1.

TABLE 1
Hypothesized Factors and Scales of Second Language Listening Ability (20 Items)

Hypothesized Factor	*Scale*	*Items*
Bottom-up Processing	**DETAIL**	10 (MC), 13 (MC), 14 (MC), 15 (MC)
Bottom-up Processing	**SUPPORT**	4 (LP), 7 (MC), 18 (LP)
Top-down Processing	**GIST**	1 (LP), 16 (MC), 17 (LP)
Top-down Processing	**TEXT INF**	2 (LP), 3 (LP), 8 (MC), 19 (LP)
Top-down Processing	**PRAG INF**	5 (LP), 11 (MC), 20 (LP)
Top-down Processing	**VOCAB**	6 (MC), 9 (MC), 12 (LP)

The assessment consisted of three separate tasks. Task One employed a 90-second video text, and had five limited-production items. Task Two employed a three-minute video text, and had 11 multiple-choice items. Task Three employed a 110-second video text, and had four limited-production items. The test task specifications, organized according to Bachman and Palmer's (1996) framework, can be seen in Appendices A, B, and C. The 20-item assessment is presented in Appendix D, and a transcript of the video listening texts is presented in Appendix E. In addition, a five-item questionnaire asking test-takers their opinions of the test was administered.

Procedures

The assessment was administered by the classroom teacher during regular class time. The teacher explained to the class that they were going to do a listening exercise. The teacher then gave out the written tests, and began the video. All of the instructions for the test were given on the video, and the teacher only had to monitor the assessment, and did not have to give instructions. The entire test took about 25 minutes (the test with the questionnaire included took 30 minutes).

The assessment consisted of three separate tasks. Task One had five limited-production items. Test-takers were given one minute to read over the five items, and then a video text was played. The video text was a 90-second, two-person dialogue. The speakers discussed the grade one of them had received in his class. The test-takers then had three minutes to answer the five items. They were instructed to answer the questions as completely as they could in 25 words or less. Task Two consisted of 11 multiple-choice questions. Test-takers were given one minute to read the task questions, and then a video text was presented. The text was a monologue about Wild Bill Hickok. Test-takers had one minute to work on the questions, and then the video was played again. They then had two minutes to complete the 11 items. Task Three was very similar to the first task. Test-takers were given one minute to read over the four limited-production items, and then a video text was played. The video text was a two-person dialogue that lasted for approximately 110 seconds. In the dialogue, one speaker told the other speaker the story of a person falling asleep in his biology class. The test-takers then had three minutes to answer the four items. Again, they were instructed to answer the questions as completely as they could in

25 words or less. After the video ended, they had five minutes to complete the five-item questionnaire.

All 20 items were scored dichotomously. The 11 multiple-choice items were scored using a key, and were rechecked for errors by a second scorer. The nine limited-production items were also scored dichotomously. Because the nature of these questions were very limited in scope, and test-takers were instructed to answer the questions in 25 words or less, dichotomous scoring seemed more appropriate than assigning partial credit to answers. An answer key for these items was created, and then revised slightly after a number of the tests had been scored. This new answer key was then used to score the limited-production items for all the tests.

An identification number was assigned to each participant, and the scores were input into an SPSS (version 10.0 for Windows) data spread sheet, and the data set was examined for missing values. There were a number of missing values, but for 81 of the 85 tests, the missing values did not appear systematic, and I treated these missing values as wrong. Four of the 85 tests scored, however, had at least 15 of the 20 items left blank, with the last two sections left totally blank. These four papers came from three different classes, and there did not seem to be any systematic basis for these unanswered tests. Because of this, I discarded these four scores.

Analyses

Statistical analyses were performed using SPSS for Windows, version 10.0, and Mplus for Windows, version 2.02.

First, the mean, median, skewness, kurtosis, and standard deviation for each of the assessment items were calculated, in order to examine the central tendencies and variability of the responses. This was done so that the appropriateness of each item in the assessment could be considered. Items with extreme means would indicate that these items might be too easy or too difficult for this population of students, and might not be suitable for the analysis.

I then performed a series of internal consistency reliability analyses. First, the reliability of each of the six scales was analyzed. The standard error of measurement for each scale was also computed, as well as the corrected item-total correlation for each item in order to determine how each item related to the other items in the scale. The reliability of each of the two hypothesized factors was then analyzed, as well as the standard error of measurement, and the corrected item-total correlation for each item. I then performed a reliability analysis to examine the overall reliability of the assessment instrument as a whole. From these analyses, I then determined which items performed poorly, and which items should be rejected. The reliability analyses for each scale, factor, and overall assessment with those items deleted, as well as the corrected-item total correlation for each of the remaining items, was then computed.

I then performed a number of exploratory factor analyses (EFAs) in an attempt to examine the patterns of correlations on the assessment to explore the basic underlying factors on the assessment. The data from this exam were based on answers scored dichotomously, and thus the variables were treated as categorical. As a result, tetrachoric correlations were required in performing EFAs. I performed these EFAs using Mplus for Windows, version 2.02, which computes tetrachoric correlations. I first prepared the correlation matrix, and examined the determinant of the matrix to determine the appropriateness of the data for factor analysis.

The EFA was then performed, using unweighted least squares analysis (which is an appropriate analysis to use with categorical data) to extract the initial factors. I examined the eigenvalues and the scree plot as indicators of the number of factor represented by the data, and then used this information in conjunction with the theoretical design of the assessment in an attempt to determine the number of underlying factors represented by the data. Another EFA was then performed, using unweighted least squares analysis with a Varimax rotation to obtain an orthogonal solution, and a Promax rotation to obtain an oblique solution. To determine which rotation procedure was most appropriate for these data, I examined the interfactor correlation matrices, and used meaningful interpretations as the final criteria for deciding the best number of factors to extract.

Finally, I examined the five-item questionnaire that the majority of the test-takers completed in an attempt to investigate test-taker attitudes about the exam.

FINDINGS

Qualitative Evidence of Construct Validity of the Assessment Instrument

Bachman and Palmer (1996) hypothesize that test tasks that are perceived as more authentic by test-takers will have a more positive impact on the test-takers. This positive impact causes test-takers to be more highly motivated, and might also lead the test-takers to perform better on the assessment, which should serve to increase the reliability of the assessment. Therefore, a short questionnaire (five questions) was given to the test-takers, in the hopes of determining their attitudes towards the assessment. The results of this questionnaire indicated that most test-takers felt that the test was neither too difficult nor too easy, and that they generally had enough time to answer the questions. Virtually all of the test-takers reported that the video was helpful in understanding the spoken text, and many also indicated that the videos were interesting to watch. Progosh (1996) found similar results when he reported that virtually all of the test-takers in his study reported that they found the video helpful in comprehending the text. Baltova (1994) also reported that test-takers thought the use of video was helpful in understanding the listening texts. That the assessment was at the appropriate level, that test-takers had enough time to answer the questions, and that they felt that the video was helpful and interesting, indicate that the impact of the test was positive for the test-takers, which in turn indicates that the test did not appear to introduce construct irrelevant sources of variance (Messick, 1989, 1996).

Tasks One and Three, which required the test-takers to write limited-production items to comprehension questions, also provided qualitative information about the role of video in the assessment. Test-takers indicated on the questionnaire that the video was helpful for them in comprehending the aural text. Their responses on the tasks requiring limited-production also indicated that they were attending to the video, and that the non-verbal aspects of the input assisted in comprehension. In these limited-production items, a number of test-takers described different actions of the speakers. They described how a speaker nodded his head, or made a face, or looked unhappy. These descriptions illustrated how listeners attended to the speakers' non-verbal communication, and these acts of non-verbal communication usually served to reinforce

the speakers' utterances. This might be seen as indicating the importance of including the non-verbal communication in listening tests (and might also serve to validate the use of video in listening assessments), because to not do so would introduce a source of invalidity into the assessment, because of construct underrepresentation.

Descriptive Statistics

First, I inspected the item means and standard deviations to ensure that each item had sufficient variability. In this study the means ranged from .31 to .98. Of the 20 items on the assessment, 17 performed well, with means ranging from .31 to .80. Three items had somewhat extreme means. Item 4 mean had a mean of .93; item 8 had a mean of .90; and item 17 had a mean of .98, indicating that these items were very easy for the test-takers.

To examine the dispersal of the scores for each item, I then examine the standard deviations of the items. These ranged from .16 to .50. The three items with very high means also had very low standard deviations. The standard deviation for item 4 was .26, for item 8 the standard deviation was .30; and for item 17 the standard deviation was .16. The other 17 items had standard deviations ranging from .40 to .50.

In an attempt to determine if the score distributions were approximately normal, I also examined the skewness and kurtosis of each items. Again, items 4, 8, and 17 had extreme values, all with a skewness absolute value more than 2. Item 4 had a skewness of –3.31. Item 8 had a skewness of –2.74. Item 17 had a skewness of –6.24. These items also had a kurtosis absolute value more than 3. Item 4 had a kurtosis of 9.21, item 8 had a kurtosis of 5.65, and item 17 had a kurtosis of 37.90. The absolute values of the skewness and kurtosis for the other 17 items were less than 2.1, all within the acceptable range.

Because of the extreme mean, standard deviation, skewness, and kurtosis values for items 4, 8, and 17, it was necessary to examine these items closely for their suitability in the assessment. All of the items were scored dichotomously, and therefore the mean values for each items corresponded to the difficulty level of the item. These three items were all very easy for the test-takers: 93% answered item 4 correctly; 90% answered item 8 correctly, and 97% answered item 17 correctly. Because the vast majority of the test-takers answered these items correctly, there was little variation on these items (and thus the standard deviations were expected to be low). Similarly, because so many test-takers answered these items correctly, the three items were, from a univariate perspective, negatively skewed and leptokurtic. These extreme kurtosis and skewness values indicated a non-normal distribution for these items, which is problematic with some of the statistical procedures utilized in this study. Although it is not inappropriate to have a small number of easy items on an assessment, for statistical purposes, I decided to remove items 4 and 17, since they both had extremely high kurtosis values, and both had skewness values more than 3. While item 8 also had high skewness and kurtosis values, these values did not necessitate automatic rejection, but must be examined for their effect when statistical procedures requiring normal distributions are conducted. These values can be seen in Table 2.

TABLE 2
Descriptive Statistics of the 20-item Assessment

Item	*Mean*	*SD*	*Skewness*	*Kurtosis*	*Changes Made*
1	**.75**	**.43**	**-1.20**	**-.58**	
2	**.80**	**.40**	**-1.55**	**.41**	
3	**.31**	**.46**	**.84**	**-1.32**	
4	**.93**	**.26**	**-3.31**	**9.21**	**Item dropped from assessment**
5	**.69**	**.46**	**-.84**	**-1.32**	
6	**.56**	**.50**	**-.23**	**-2.00**	
7	**.47**	**.50**	**.13**	**-2.04**	
8	**.90**	**.30**	**-2.74**	**5.65**	
9	**.33**	**.47**	**.72**	**-1.52**	
10	**.70**	**.45**	**-.91**	**-1.20**	
11	**.33**	**.47**	**.72**	**-1.52**	
12	**.65**	**.48**	**-.66**	**-1.60**	
13	**.60**	**.49**	**-.44**	**-1.86**	
14	**.65**	**.48**	**-.66**	**-1.60**	
15	**.65**	**.48**	**-.66**	**-1.60**	
16	**.73**	**.45**	**-1.05**	**-.93**	
17	**.98**	**.16**	**-6.24**	**37.90**	**Item dropped from assessment**
18	**.62**	**.49**	**-.49**	**-1.80**	
19	**.68**	**.47**	**-.78**	**-1.43**	
20	**.51**	**.50**	**-.03**	**-2.05**	

Reliability Analyses

I first performed reliability analyses for each of the six scales as measured by the 18 remaining items. The internal consistency reliability for these scales ranged from -.256 to .443. That two of the scales (GIST and VOCAB) had negative reliability indicated that a number of items were not performing well, and that revisions to the assessment instrument would be necessary. I also calculated the item-total correlation for each item in each scale, in an attempt to determine which items were performing well.

I then formed composite variables, and performed reliability analyses based on the two-factor processing model. The bottom-up processing factor consisted of the six DETAIL and SUPPORT items, and the top-down processing factor consisted of the 12 GIST, TEXT INF, PRAG INF, and VOCAB items. The bottom-up processing factor had an internal consistency reliability of .502, while the top-down processing factor had a reliability of .531. I then examined the corrected item-total correlation for each item, in order to investigate which items were performing well within the two theorized factors. Of the six items in the bottom-up processing factor, five performed adequately, with item-total correlations ranging from .236 to .411. Only item 7 performed poorly, with a corrected item total correlation of .019. Because of this very low item-total correlation, as well as the low item-total correlation that item 7 exhibited in the SUPPORT scale, I decided to reject this item. I then recalculated the item-total correlations for each item (without item 7). These results can be seen in Table 3.

Of the 12 items in the top-down processing factor, eight performed adequately, with corrected item-total correlations ranging from .195 to .500. Four items performed poorly, with

item-total correlations ranging from -.162 to .092. Of these four items, three of them (items 6, 9, and 12) were the items making up the VOCAB scale. Because these items performed poorly in relation to the VOCAB scale and in relation to the top-down processing factor, I decided to remove these three items from the instrument. Similarly, I decided to remove item 11, because as part of the top-down processing factor it had an item-total correlation of only .092, and as part of the PRAG INF scale, it had an item-total correlation of -.011.

I then performed another reliability analysis of each scale, with the five items (6, 7, 9, 11, 12) deleted. These revised values, along with the revised item-total correlation of each item when grouped by factor, can be seen in Table 3.

TABLE 3
Revised Reliability Analysis of Each of the Scales and Factors (13 Items)

Item	*Type*	*Scale*	*Processing Factor*	*Item-Total Correlation When Grouped By Factor*	*Revised Scale alpha*
			Bottom-up		
10	MC	DETAIL		.250	.443
13	MC	DETAIL		.245	
14	MC	DETAIL		.276	
15	MC	DETAIL		.448	
18	LP	SUPPORT		.463	n.a.
			Top-down		
1	LP	GIST		.424	.071
16	MC	GIST		.236	
2	LP	TEXT INF		.393	.431
3	LP	TEXT INF		.298	
8	MC	TEXT INF		.313	
19	LP	TEXT INF		.354	
5	LP	PRAG INF		.535	.579
20	LP	PRAG INF		.406	

Items Rejected: 4, 6, 7, 9, 11, 12, 17

Because three of the five rejected items were devised to measure the VOCAB scale (items 6, 9, and 12), it was necessary to drop the VOCAB scale from the instrument. The three VOCAB items had very low item-total correlations in relation to the VOCAB scale, the top-down processing factor, and the overall assessment. Why these types of questions performed so poorly is a matter of speculation. It could be that some of the test-takers already knew the meaning of the vocabulary words, and thus these items were testing vocabulary knowledge rather than the ability to infer an unknown word's meaning through the context of the surrounding passage of the text. It may also be that this skill is regularly included in construct definitions of reading ability, and because of this (and the apparent similarities between listening and reading), the skill has also been included in defining listening ability. This may be a skill that is relevant for reading, but is less salient (or at least less testable) in listening. A reader can focus on, study, and try to infer the meaning of an unknown word in a written text, but because of the immediacy

of a spoken text, it may be impossible for L2 listeners to perform this elaborate process. This idea obviously warrants further research, but for this assessment the VOCAB scale was dropped.

I then re-estimated the internal consistency reliability of each of the factors in the two-factor processing model, as well as the reliability for the entire assessment (the 13-item revised version). The internal consistency reliability for the two hypothesized factors were: bottom-up processing scale (α = .576) with five items; and the top-down processing scale (α = .677) with eight items. The reliability for the overall, 13-item (revised) assessment was .774. The revised version of the assessment is seen in Table 4.

TABLE 4
Revised Listening Assessment (13 Items)

	Reliability α =	*Items Kept for Each Factor Ordered From Best to Worst Indicator*
Overall Assessment	.774	
Bottom-up Processing Factor	.576	18, 15, 14, 10, 13
Top-down Processing Factor	.677	5, 1, 20, 2, 19, 8, 3, 16

Factorial Structure for the Assessment Instrument

Because the reliability analyses provided limited evidence for the construct validity of the assessment, I then performed a series of EFAs in an attempt to examine the factor structure of the exam. I first examined the appropriateness of the data for performing EFAs. While the determinant of the correlation matrix was not positive, it was very close to zero, and the estimates given were still valid (Mplus, version 2.02), indicating that the data were acceptable for factor analyses. However, the fact that the determinant is not positive may lead to inflated factor loadings (Kline, 1998).

In the (slightly revised) operationalization of the theoretical model for listening, I hypothesized that academic listening was composed of five skills. However, a five-factor solution did not fit the data well. This may in part be due to the small number of items (13) used in the final assessment, which would make a five-factor solution unlikely. Although Mplus extracted four eigenvalues greater than 1.0, an inspection of the scree plot suggested that two factors best represented the data. I therefore compared solutions with one, two, three, four, and five factors. The two-factor promax solution seemed to maximize parsimony, as shown in table 5.

Factor 1 includes items designed to measure all five of the scales. Factor 2 includes items designed to measure three of the five scales. Factor 1 contains a GIST item, a TEXT INF item, two PRAG INF items, and a SUPPORT and a DETAIL item. Factor 2 contains a GIST item, three DETAIL items, and two TEXT INF items. Neither the five-scale model, nor the two-factor model based on bottom-up versus top-down processing, initially seems to be validated by the EFA. This two-factor PROMAX rotation solution can be seen in Table 5

TABLE 5
Pattern Matrix and Interfactor Correlations for the 13-Item Assessment

PROMAX Rotation

Item	*Scale*	*Item Type*	*F1*	*F2*
1	GIST	LP	1.045	-.287
2	TEXT INF	LP	.906	-.141
5	PRAG INF	LP	.662	.202
18	SUPPORT	LP	.612	.263
20	PRAG INF	LP	.566	.097
13	DETAIL	MC	.392	.135
8	TEXT INF	MC	-.235	.954
16	GIST	MC	-.181	.803
19	TEXT INF	LP	.071	.550
14	DETAIL	MC	-.009	.522
15	DETAIL	MC	.389	.513
3	TEXT INF	LP	.172	.491
10	DETAIL	MC	.043	.453

Interfactor Correlation Matrix

	F1	*F2*
Factor 1	1.000	.515
Factor 2	.515	1.000

Internal Consistency Reliability of Each Factor

	Reliability α =	*Items for Each Factor Ordered From Highest to Lowest Loading*
Factor 1	.738	1, 2, 5, 18, 20, 13
Factor 2	.663	8, 16, 19, 14, 15, 3, 10

The items designed to measure the five different scales that are part of second language listening ability were fairly evenly scattered between the two factors. This would seem to indicate that the theoretical rationale was inadequate, or that the items designed for those constructs were not measuring what they were designed to measure. A closer analysis of the factor analysis might indicate that it was a combination of inadequate theoretical rationale and improperly coded items. Because the two-factor model created by the exploratory factor analysis correlated inadequately with the hypothesized model, it was necessary to reanalyze the original items and use substantive rationale in order to construct some sort of meaningful interpretation of the factor structure.

It was hypothesized that the five different scales were part of a two-factor model representing two different types of aural processing (top-down and bottom-up). Since a two-factor model was in fact found by the factor analysis, this is an attractive interpretation.

DETAIL items were designed to measure fairly minute and local pieces of information in the text. Similarly, SUPPORT items were designed to measure slightly broader (yet still local) details and supporting ideas. These items were broader than DETAIL items, but still were meant to test listeners' ability to comprehend fairly focused information—local information that helped create meaning in the text, but it was not necessary for the listener to comprehend this specific information in order to understand the larger meaning of the overall text. DETAIL and SUPPORT items were thus designed to measure bottom-up processing.

The GIST, TEXT INF, and PRAG INF items were designed to assess listeners' comprehension of the broader and global meaning of the text, and their ability to process aural information in a top-down manner. The GIST items were very broad, and asked listeners the overall theme or mood of the texts. The TEXT INF and PRAG INF items were somewhat narrower in scope, but still were designed so that in order for listeners to correctly answer these questions, they had to be able to comprehend and process large pieces of information, and process the information in a top-down manner in order to create specific meaning from the global information. However, a closer analysis of the coding of the items is helpful here, and may help explain the factor structure.

According to the two-factor (bottom-up and top-down) processing model, one of the two factors would be represented by DETAIL and SUPPORT items, while the second factor would be represented by GIST, TEXT INF, and PRAG INF items. The initial results of the factor analysis performed, however, do not seem to provide much evidence for the validity of this model, because factor 1 is represented by items from all five of the skills, and factor 2, while represented by GIST and TEXT INF items, also contains DETAIL (rather than PRAG INF) items. But a more detailed analysis (and slightly revised conceptual framework) might account for the two-factor model suggested by the data. A possible interpretation for the two-factor model would include the idea of information stated explicitly in the text. One factor would include items in which the answers were explicitly stated in the text, and the second factor would include items in which the answers were not explicitly stated.

Items 1, 2, 5, 18, 20, and 13 all loaded on factor 1. The answers for all of these items (except, perhaps, for item 5) are all explicitly stated in the text. Item 1, which loaded heavily on factor 1 (1.045)[3], asks the listener to describe Bob's (one of the speakers in the text) mood. This item was originally coded a GIST item, because the whole conversation is centered on the fact that Bob is unhappy. This item was intended to measure a listener's ability to process (top-down) a large amount of information, and synthesize meaning from it. But in the text, in response to the question "Hey Bob, how's it going?", Bob responds "I'm a bit upset, because…." Virtually all of the correct responses to this item include "Bob is upset because…." While this was designed to be a GIST item, listeners were able to take (a very limited and specific piece of) information that was explicitly stated, and answer the question correctly.

Similarly, item 2 (TEXT INF), that also loaded heavily on factor 1 (.906), was designed to measure a listener's ability to inference information given in the text. Listeners were asked, "Do you think Bob had a good reason for missing class? Why or why not?" The information was explicitly stated in the text about why he missed the class (he had to go to California for his sister's wedding), but listeners had to give their opinion on his reason for missing class. Test-takers could argue for or against his missing class, but they had to provide some rationale for

[3] That this item loaded above 1.000 may be indicative of inflated loadings due to the non-positive determinant of the correlation matrix.

their argument, and this rationale was that he had to go to California for his sister's wedding. The rationale was explicitly stated in the text.

Item 18 was a SUPPORT item that loaded on factor 1 with a .612 value. The answer was explicitly stated in the text. Item 20, with a .566 loading on factor 1, was a PRAG INF item, designed to measure a listeners' ability to make inferences based on pragmatic aspects of the speaker's aural output. The question asks "What is Amy's attitude toward David when he tells the story?" Rather than describing Amy's attitude, virtually all of the test-takers (not incorrectly) described what Amy said that indicated her attitude. Test-takers used the explicit information given by the speaker to answer the question. Item 13, which also loaded on factor 1 (.392), was designed to be a DETAIL question, and the answer is explicitly stated in the text.

Of the six items to load on factor 1, only item 5 seems problematic. The question asks, "Will Bob do what Julie advises? Why or why not?" Julie told Bob that he should go and talk to his teacher about his grade, and Bob responds, "Yeah, you're probably right." There is no guarantee that Bob will actually do as Julie advises, so test-takers were required to give a rationale for their answers. Since there is not one, correct answer given explicitly in the text, this item loading on factor 1 is difficult to interpret. It could be argued that since Bob said in the text "Yeah…", the answer was explicitly given. And the vast majority of correct answers were in the affirmative. Also, this may be an item in which the visual aspects of the text played a role. Bob not only says, "Yeah, you're probably right", but he also nods his head and his body language indicates that he agrees with her. Still, it is questionable how "explicit" this body language is, and this item is somewhat problematic for the explicit/implicit two-factor model.

Factor 2, which might be seen as corresponding to items that do not have answers explicitly stated in the text, has one GIST item, three TEXT INF items, and three DETAIL items. Item 8 is a TEXT INF multiple-choice item that does not have the correct answer explicitly stated. In fact, the question includes the statement, "From the video, we might conclude that…", which indicates to the test-takers that the answer was not specifically stated. Item 8 loaded heavily (.954) on Factor 2. Item 16, the GIST item, is a multiple-choice item asking for the best title for the passage. This answer obviously is not explicitly stated, and this loads the highest on factor 2 of all the items (.803).

Items 19 and 3 both loaded on factor 2 (at .550 and .491 respectively). These were both TEXT INF items. Item 19 asked "What is Amy's attitude towards Tina when she hears the story?" This is very similar to item 18 (that loaded on factor 1) that asked Amy's attitude towards David. The difference between the two items, however, is that the answer to item 18 is explicitly stated, while it is not for item 19.

Items 14 and 15 are DETAIL items. These were coded DETAIL because they relate to very local and focused points in the text. Still, the answers are not explicitly stated, and some inferencing is necessary. Items 14 and 15 load on factor 2 at .522 and .513 respectively. However, it should also be noted that item 15, while loading on factor 2 at .444, also loaded fairly heavily on factor 1, at .389.

Item 3 asks, "How did Bob do on his final exam?" This item was also somewhat problematic. The correct answer was that he was not sure how he had done. He thought that he did well on the exam, but he hadn't gotten his grade back yet. Many test-takers answered this answer incorrectly (it had the lowest mean (.31) of any of the retained items), answering that he had gotten a "C" on the exam. Earlier in the text, Bob stated that he had gotten a "C" for the class, and many test-takers apparently heard this explicit information, and used it (incorrectly) for this answer. It should also be noted that this item had the second lowest item-total correlation

of the retained items (.28), which might indicate that test-takers had difficulty differentiating between explicitly stated information that was an incorrect answer, and the implicit and unstated correct answer.

Item 10 is a DETAIL item that loaded at .453 on factor 2. This question asks, "How many people were killed at Rock Creek Station?", and the four possible answers are two, three, five, and ten. The correct answer was explicitly stated in the text (three people were killed), which makes it somewhat problematical that it loaded on factor 2. However, this question could be seen as somewhat misleading, in that the sentence in the text pertaining to this item stated "Wild Bill, along with **two** other men, killed **three** men at Rock Creek Station, but the magazine accounts of the incident credit Wild Bill with killing **ten** men all by himself." The correct answer is explicitly stated, but two of the other possible choices are also explicitly stated, and in the same sentence. This was a somewhat problematic item, and it had the lowest item-total correlation (.25) of any of the retained items, indicating that it may have been a trick question for some of the test-takers. Test-takers heard the answer explicitly stated, but they also had to distinguish between two other possible answers that were explicitly stated. It seems likely that for test-takers to answer this item correctly, they would have to be able to distinguish between different pieces of information in the same utterance, representing a high processing load. Thus, it is not surprising that this item loaded on factor 2, even though it did have the answer explicitly stated in the text.

In summary, the EFA performed on these data seems to indicate that a two-factor solution is the most appropriate. For this two-factor solution, factor one seems to relate to the ability to listen for explicitly stated information, and the second factor relates to the ability to listen for implicit information.

Method Effects

I also examined the data for the possibility of any underlying test method effects. In the two-factor solution just discussed, factor 1 (listening for explicitly stated information) had five limited-production, and one multiple-choice question with loadings higher than .3 (item 15, which was a multiple-choice item, also loaded on factor 1 at .389, although it loaded more heavily on factor 2, at .513). Factor 2 had two limited-production and five multiple-choice questions. In addition, the one multiple-choice question that loaded on factor 1 loaded more weakly on this factor than the other 5 items (it loaded at .392). This grouping of items according to question type might indicate that test method effects, particularly due to question type, was the cause for the factor groupings. I therefore ran two separate EFAs, the first with the six multiple-choice items, and the second EFA with the seven limited-production items.

The results of these EFAs provided somewhat conflicting evidence about the role of method effect in the factor grouping. For the multiple-choice question types, on the basis of the eigenvalues greater than the 1.0 criterion, a two-factor solution appeared optimal, but examination of the scree plot seemed to indicate that a one-factor solution was most appropriate. For a one-factor solution, all six items loaded at .3 or above. For a two-factor solution, items 16, 15, 13, and 8 loaded on factor 1, and item 10 loaded on factor 2 (item 14 did not load above .300 on either factor). The two-factor solution would indicate that there was not a test method effect based on question type. However, this two-factor solution based on the six multiple-choice questions does not mirror exactly the two-factor solution found in the overall, 13-item

assessment, and this EFA could also be interpreted as presenting evidence for a test method effect. The only multiple-choice question that loaded on factor 1 in this EFA was item 10, which has already been discussed as being somewhat problematic. As discussed earlier, the answer to this question could be seen as being given both explicitly and implicitly in the spoken text. This may also be the reason it (was the only item that) loaded on factor 2 in this EFA.

I then performed a second EFA on the seven limited-production questions. On the basis of the eigenvalues greater than the 1.0 criterion and examination of the scree plot, a one-factor solution was extracted. In the two-factor explicit/implicit solution, five limited-production items loaded on factor one, and two limited-production items loaded on factor two, but in the EFA conducted with only limited-production items, only one factor was extracted. Again, this could be interpreted as evidence that there is a test method effect dependent on question type, or evidence of correlated measurement error, in the assessment.

DISCUSSION

This study examined the construct validity of an assessment aimed at measuring the listening ability of second language learners in an academic listening TLU domain. The assessment itself was delivered through the use of video, and was based on a theoretical model divided into six separate scales, with a two-factor model based on bottom-up and top-down aural processing. The current study aimed to determine if the items designed to measure each skill did indeed measure what it was designed for, in an attempt to determine the construct validity of the overall assessment, and to attempt to validate this model of second language listening comprehension.

With regard to research question 1, "What is the nature of second language listening ability as measured by a listening comprehension test delivered through the use of video?", the results seemed to present some evidence for a two-factor model of second language listening ability. This two-factor model is similar to the two-factor model often theorized in the literature, with one factor corresponding to the ability to perform processing in which the listener is required to comprehend explicitly stated aural information. The second factor corresponds to the ability of the listener to process implicit information in an aural text. Also, some qualitative evidence was presented validating the inclusion of non-verbal communication in L2 listening models.

With regard to research question 2, "To what extent do the scores from the assessment provide evidence for the construct validity of the theoretical model?", the scores provide limited evidence for the construct validity of the six-skill theoretical model of second language listening comprehension. In addition, the results showed that many of the items did not measure the underlying skill that they were specifically designed to measure. This may be because the items were poorly designed, or miscoded. As Buck (1991) noted, items designed to measure one skill might end up testing quite another skill. Limited-production questions are especially problematic, because test-takers can give different answers to the same questions, and thus items meant to test top-down processing sometimes required bottom-up processing, and vice versa. It should also be mentioned that these results are not entirely surprising. As Brindley (1998) and Buck (1991, 2001) have argued, because the two processes involved in listening are so

interrelated and act simultaneously, it is very difficult to differentiate between these levels of processing, or to attribute responses on a test to any one skill.

The results of the exploratory factor analysis may have provided some evidence for the validity of the two-factor higher order model that was hypothesized, although in a slightly modified form. A two-factor model of listening comprehension was theorized, with the two factors relating to items requiring bottom-up processing, and top-down processing. The two-factor solution presented by the exploratory factor analysis, however, may indicate that the two factors relate more to the idea of items in which the answers are explicitly stated in the text, versus items in which the answers are not explicitly stated in the text. Five of the six items that loaded on factor 1 did have the answers explicitly stated in the text. The only item that loaded on factor 1 that did not have the answer explicitly stated in the text was item 5, although there were some extenuating circumstances with this item. The six items that loaded on factor 2 did not have the answers explicitly stated in the text. These findings are very similar to the findings of Hansen and Jensen (1994), as well as more recently, Purpura (1999). In his analysis of the reading section of the University of Cambridge First Certificate in English Anchor Test, Purpura found a two-factor solution for the passage comprehension section, with the two factors representing "reading for explicit information" and "reading for inferential information".

With regard to research question 3, "Is there evidence of a test method effect caused by items designed with specific question types (multiple-choice versus limited-production)?", the results, though inconclusive, indicate that there might indeed be some sort of method effect related to item type. While this could be seen as a problem with this assessment instrument, it could also be indicative of the difficulty in creating valid and reliable listening assessment items. It also could be indicative of the fact that by their very nature, limited-production items may be more suitable for testing a listener's ability to comprehend inferential information, while multiple-choice type items may be better suited to assess a listener's ability to comprehend explicitly stated information. This idea necessitates further research in this area.

CONCLUSION

In this paper, the development and construct validation of a listening assessment for second language learners was explored. The development of the assessment was based on theories of second language acquisition, and research in language testing, especially pertaining to testing second language listening ability with video listening texts. While the present study did not present sufficient evidence to validate the construct definition of second language listening ability posited here, the exploratory factor analyses did present some evidence for the validation of a two-factor model of listening based on the ability to comprehend explicitly stated information, and the ability to comprehend implicit information in aural texts. In turn, this two-factor model should be examined further in an attempt to validate it. This study also indicated that a method effect based on question type may have been inherent to the assessment instrument used, and a more detailed analysis of this method effect is warranted.

Although second language listening comprehension (and its assessment) has attracted increasing attention in recent years, it remains an area with many unanswered questions. The process of second language listening needs to be more adequately described, and this will help establish a more widely-accepted and hence useful definition of L2 listening ability. While such

work is in progress, the testing field should proceed with research to inform and enable the creation of better L2 listening tests. Further work in specific areas of L2 listening such as the effect of question type, text type, and the use of video to deliver the aural texts should lead to more reliable and valid listening assessments.

REFERENCES

Aitken, K. (1978). Measuring listening comprehension. *English as a second Language. TEAL Occasional Papers,* Vol. 2. Vancouver: British Colombia Association of Teachers of English as an Additional Language. ERIC Document No. ED 155 945.

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

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Baltova, I. (1994). The impact of video on comprehension skills of core French students. *Canadian Modern Language Review, 50*, 507-531.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton, NJ: Educational Testing Service.

Benson, M. (1989). The academic listening task: A case study. *TESOL Quarterly, 23*, 421-45.

Berne, J. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania, 78*, 316-329.

Blau, E. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly, 24*, 746-753.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18*, 171-191.

Brown, G. (1986). Investigating listening comprehension in context. *Applied Linguistics, 7,* 284-302.

Brown, G. (1995). Dimensions of difficulty in listening comprehension. In D. Mendelshon & J. Rubin (Eds.), *A Guide for the Teaching of Second Language Listening* (pp. 59-73). San Diego: Dominie Press.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*, 67-91.

Buck, G. (1994) The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*, 145-170.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-157.

Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment.* Tampere: University of Jyvaskyla.

Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly, 19*, 765-781.

Canale, M., Child, J., Jones, R., Liskin-Gasparro, J., & Lowe, P. (1984). The testing of reading and listening proficiency: a synthesis. *Foreign Language Annals, 17*, 389-392.

Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*, 354-374.

Chiang, C., & Dunkel, P. (1992). The effect of speech modification, prior knowledge and listening proficiency on EFL lecture learning. *TESOL Quarterly, 26*, 345-74.

Chaudron, C. (1983). Simplification of input: Topic reinstatements and their effects on L2 learners' recognition and recall. *TESOL Quarterly, 17*, 437-458.

Clark, H., & Clark, E. (1977). *Psychology and language: An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Conrad, L. (1985): Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition, 7*, 59-72.

Conrad, L. (1989). The effects of time-compressed speech on listening comprehension. *Studies in Second Language Acquisition, 11*, 1-16.

Douglas, D. (1988). Testing listening comprehension in the context of the ACTFL proficiency guidelines. *Studies in Second Language Acquisition, 10*, 345-61.

Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL Quarterly, 22*, 259-281.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*, 180-191.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension—an overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7-29). Cambridge: Cambridge University Press.

Gruba, P. (1997). The role of video media in listening assessment. *System, 25*, 335-345.

Hadley, A. (2001). *Teaching language in context*. Third edition. Boston: Heinle and Heinle.

Hale, G., & Courtney, R. (1994). The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing, 11*, 29-48.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241-268). New York: Cambridge University Press.

Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes, 1*, 91-107.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29*, 135-149.

Kellerman, S. (1992). "I see what you mean." The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics, 13*, 239-58.

Kline, R. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

Longeran, J. (1984). *Video in language teaching.* Cambridge: Cambridge University Press.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal, 75*, 196-204.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13-103). Third edition. New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 242-256.

Mplus (2001). Muthen and Muthen, Inc.

O'Malley, J., Chamot, A., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*, 418-37.

Parker, K., & Chaudron, C. (1987). The effects of linguistic simplification and elaborative modifications on L2 comprehension. *University of Hawai'i Working Papers in ESL, 6*, 107-133.

Pennycook, A. (1985). Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly, 19*, 336-43.

Peterson, P. (1991). A synthesis of methods for interactive listening. In M. Celce-Murcia, (Ed.), *Teaching English as a second or foreign language*, 2nd Edition (pp. 106-122). New York: Newbury House.

Pica, T., Young, R., & Doughty, D. (1987). The impact of interaction on comprehension. *TESOL Quarterly, 21*, 737-758.

Progosh, D. (1996). Using video for listening assessment opinions of test-takers. *TESL Canada Journal, 14*, 34-43.

Purpura, J. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Richards, J. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly, 17,* 219-40.

Rost, M. (1990). *Listening in language learning.* New York: Longman.

Rubin, A. (1980). Theoretical taxonomy of the difference between oral and written language. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 411-438). Hillside, NJ: Erlbaum.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78*, 199-221.

Samuels. S. (1984). Factors influencing listening: Inside and outside the head. *Theory into Practice, 23*, 183-89.

Secules, T., Herron, C., & Tomasello, M. (1992). The effect of video context on foreign language learning. *Modern Language Journal, 76*, 480-490.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing, 14*, 185-213.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question-type. *Language Testing, 8*, 23-40.

SPSS Version 10.0.5 (1999). SPSS Inc.

Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 1-16). Norwood, NJ: Ablex.

Tyler, L., & Warren, P. (1987). Local and global structure in spoken language comprehension. *Journal of Memory and Language, 26*, 638-657.

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Von Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second language research and teaching. *Canadian Modern Language Review, 36*, 225-237.

Weir, C. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Wilkinson, R. (1984). Video-based learning activities. *TESL Canada Journal, 1*, 83-87.

Willis, J. (1983). The role of the visual element in spoken discourse: Implications for the exploitation of video in the EFL classroom. *ELT Documents, 114*, 29-43.

Wu, Y. (1998). What do tests of listening comprehension test?--A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*, 21-44.

APPENDIX A

Test Task Characteristics for Task One

I. Setting

A. Physical characteristics--classroom

B. Participants: Test-takers—High School ESL students from multi-ethnic backgrounds, but predominantly Latino, most from lower socio-economic backgrounds, with intermediate to advanced levels of proficiency in English

Test Administrators—Classroom teacher

C. Time of task--6 minutes—1 minute for question preview, 1.5 minutes for first playing of text, 3 minutes to finish answering questions

II. Input

A. Format

1. Channel--visual and aural text (video), and written instructions

2. Form--language and nonlanguage

3. Language--target language (English)

4. Length--1 minute and 30 seconds

5. Item Type--5 limited-production (short answer) questions are given to test-takers (TTs), TTs have 1 minute to review the questions, then text is shown, then TTs have 3 minutes to finish answering questions

6. Speededness--unspeeded

7. Vehicle—video tape presented on television monitor

B. Language characteristics

1. Organizational characteristics

    a. Grammatical—varied; many different grammatical forms used

    b. Textual—conversational turn-taking

2. Pragmatic characteristics

    a. Functional--ideational and manipulative (describing, persuading)

b. Sociolinguistic—Standard American English; moderately informal, natural, some vernacular and use of cultural references

3. Topical characteristics—personal and academic (one student talking to another student about a grade he had received in his class)

III. Expected response

A. Format

1. Channel--visual

2. Form--language

3. Language--target language (English)

4. Length—short, maximum of 25 words in length

5. Type—limited-production

6. Speededness--not designed to be speeded, though some students may not "finish" the task according to their own perception

B. Language characteristics

1. Organizational characteristics

a. Grammatical--variety of expected responses, ranging from low to high level general vocabulary; low to high level of proficiency in standard English morphology and syntax; graphology = handwritten

b. Textual—n.a.

2. Pragmatic characteristics

a. Functional--ideational and manipulative (describing, arguing)

b. Sociolinguistic--variety of expected responses using full range of registers (informal to formal); wide variety of degrees of naturalness; figurative language and cultural references not expected

3. Topical characteristics—related to the specific questions

IV. RELATIONSHIP BETWEEN INPUT AND RESPONSE

A. Reactivity--non-reciprocal

B. Scope of relationship-- broad and narrow scope, depending on the input item

C. Directness of relationship—direct and indirect, depending on the input item

APPENDIX B

Test Task Characteristics for Task Two

I. Setting

A. Physical characteristics--classroom

B. Participants: Test-takers—High School ESL students from multi-ethnic backgrounds, but predominantly Latino, most from lower socio-economic backgrounds, with intermediate to advanced levels of proficiency in English

Test Administrators—Classroom teacher

C. Time of task--11 minutes—1 minute for question preview, 3 minutes for first playing of text, 1 minute to answer questions, 3 minutes for second playing of text, 2 minutes to finish answering questions

II. Input

A. Format

1. Channel--visual and aural

2. Form--language and nonlanguage

3. Language--target language (English)

4. Length--3 minutes

5. Item Type--11 selected response (multiple-choice) questions are given to test-takers (TTs), TTs have 1 minute to review the questions, then text is shown, TTs have 1 minute to answer questions, text is shown again, then TTs have 2 minutes to finish answering questions

6. Speededness—unspeeded

7. Vehicle—video tape presented on television monitor

B. Language characteristics

1. Organizational characteristics

    a. Grammatical—many different grammatical forms used

    b. Textual--cohesive sentences, academic lecture rhetorical organization

2. Pragmatic characteristics

    a. Functional--ideational and manipulative (describing, instructing)

b. Sociolinguistic--Standard American English; moderately informal; natural; some figurative language and cultural references

3. Topical characteristics—academic and informational ("Wild Bill Hickock")

III. Expected response

A. Format

1. Channel--visual (written test items)

2. Form—non-language (circle the correct response)

3. Language—n.a.

4. Length—n.a.

5. Type—selected response

6. Speededness--not designed to be speeded, though some students may not "finish" the task according to their own perception

B. Language characteristics

1. Organizational characteristics

a. Grammatical—n.a.

b. Textual—n.a.

2. Pragmatic characteristics

a. Functional—n.a.

b. Sociolinguistic—n.a.

3. Topical characteristics—n.a.

IV. RELATIONSHIP BETWEEN INPUT AND RESPONSE

A. Reactivity--non-reciprocal

B. Scope of relationship--broad and narrow scope, depending on the input item

C. Directness of relationship—direct and indirect, depending on the input item

APPENDIX C
Test Task Characteristics for Task Three

I. Setting

A. Physical characteristics--classroom

B. Participants: Test-takers—High School ESL students from multi-ethnic backgrounds, but predominantly Latino, most from lower socio-economic backgrounds, with intermediate to advanced levels of proficiency in English

Test Administrators—Classroom teacher

C. Time of task--6 minutes—1 minute for question preview, 2 minutes for first playing of text, 3 minutes to finish answering questions

II. Input

A. Format

1. Channel--visual and aural text (video), and written instructions

2. Form--language and nonlanguage

3. Language--target language (English)

4. Length--1 minute and 50 seconds

5. Item Type--4 limited-production (short answer) questions are given to test-takers (TTs), TTs have 1 minute to review the questions, then text is shown, then TTs have 3 minutes to finish answering questions

6. Speededness--unspeeded

7. Vehicle—video tape presented on television monitor

B. Language characteristics

1. Organizational characteristics

    a. Grammatical—varied; many different grammatical forms used

    b. Textual—conversational turn-taking

2. Pragmatic characteristics

    a. Functional--ideational and manipulative (describing, persuading)

    b. Sociolinguistic—Standard American English; moderately informal, natural, some vernacular and use of cultural references

3. Topical characteristics—personal and academic (one student describing to the other student what happened in biology class)

III. Expected response—same as for Task One

IV. RELATIONSHIP BETWEEN INPUT AND RESPONSE—same as for Task One

APPENDIX D

The Assessment Instrument

LISTENING TEST (three sections)

Part 1 BOB AND JULIE
Watch the video, and then answer the following questions as completely as you can in 25 words or less. You will see the video one time.

1. Describe Bob's mood. (GIST)
2. Do you think Bob had a good reason for missing class? Why or why not? (TEXT INF)
3. How did Bob do on his final exam? (TEXT INF)
4. What advice does Julie give Bob? (SUPPORT)
5. Will Bob do what Julie advises? Why or why not? (PRAG INF)

Part 2 WILD BILL HICKOK
Watch the video, and then answer the following questions. Circle the best answer. You will see the video two times.

1. The word "dandy" describes a person who _____. (VOCAB)
   a. moves from town to town
   b. kills people without remorse
   c. is very concerned with the way they look
   d. uses the "underhand" or "twist" draw in a gunfight

2. Wild Bill's "twist" draw was an unusual draw, because most gunfighters_____. (SUPPORT)
   a. only used one gun
   b. drew their guns without looking
   c. had their gun handles facing backward
   d. thought that the "twist" draw was quicker

3. From the video, we might conclude that Wild Bill moved around so often because he _____. (TEXT INF)
   a. made a lot of enemies
   b. wanted to see the world
   c. owed people a lot of money
   d. got tired of living in the same town

4. In the video, the word "exploits" means _____.(VOCAB)
   a. credits
   b. traditions
   c. adventures
   d. exaggerations

5. How many people were killed at Rock Creek Station? (DETAIL)
   a. 2
   b. 3
   c. 5
   d. 10

6. We might conclude that magazines wrote exaggerated stories about Wild Bill because _____. (PRAG INF)
   a. it helped sell magazines
   b. it was hard to tell fact from fiction
   c. he was famous because he was the sheriff
   d. he fought with General Custer in the Civil War

7. The word "notorious" means _____. (VOCAB)
   a. good at helping people
   b. tired of moving around
   c. well-known for something bad
   d. able to draw and shoot very quickly

8. Wild Bill was killed in _____. (DETAIL)
   a. Abilene
   b. Deadwood
   c. Kansas City
   d. Rock Creek Station

9. Jack McCall was _____. (DETAIL)
   a. a poker player
   b. the sheriff of Abilene
   c. the killer of Wild Bill
   d. a U.S. Cavalry General

10. "Dead Man's Hand" is the name of the _____. (DETAIL)
    a. gun Wild Bill used
    b. horse Wild Bill rode
    c. draw Wild Bill used in gunfighting
    d. cards Wild Bill was holding when he was shot

11. What's the best title for this passage? (GIST)
    a. "The Legend of Wild Bill"
    b. "Wild Bill: The Dandy Sheriff"
    c. "The Rock Creek Station Killer"
    d. "Wild Bill: The Best Poker Player in the West"

Part 3 DAVID AND AMY
Watch the video, and then answer the following questions as completely as you can in 25 words or less. You will see the video one time.

1. What happened to Tina in class, and why? (GIST)
2. How did the professor respond to the incident? What did he say and do?(SUPPORT)
3. What is Amy's attitude towards Tina after she hears the story? (TEXT INF)
4. What is Amy's attitude toward David when he tells the story? (PRAG INF)

APPENDIX E

Transcripts of the Aural Texts

Text 1 – Bob and Julie

J: Hey Bob.
B: Hey Julie, what's up?
J: Not much. How are you doing?
B: Oh, I'm OK.
J: You don't look OK. What's wrong.
B: Oh, I'm a bit upset because I got a "C" in that class I was taking.
J: A "C"? Why did you get a "C"?
B: I don't know. I thought I was doing really well, but the teacher gave me a "C".
J: Well, how did you do on the final?
B: I don't know for sure, because I haven't gotten it back yet. But I thought I did pretty well. I honestly don't know how I got a "C".
J: Did you miss any classes?
B: I only missed one. And that was because I went my sister's wedding in California. I mean, I had to go. It was my sister's wedding.
J: Well, maybe you should talk to your teacher?
B: Oh, you think so.
J: Yeah.

B: Uh, I don't know. I don't think she likes me.
J: Well, regardless of whether she likes you, you should go talk to her to find out why you got a "C".
B: Yeah, maybe you're right.
J: I think it would be a good idea.

Text 2 – Wild Bill Hickok

James Butler Hickok—probably not too many people recognize this name. But this is the real name of one of the most famous gunfighter of the American West, "Wild Bill" Hickok.

(Picture of Wild Bill shown for 2 seconds)

Wild Bill was really something to see. He was very tall and thin, with long blond hair and a big drooping mustache. And he was something of a dandy. He always wore finely tailored suits and frock coats, He liked to make a big entrance, and for people to know who he was. And he was easy to recognize, especially because of his guns. He always wore a holster with two Colt pistols, with their ivory handles turned forward. The handles were turned forward because Wild Bill used the underhand, or "twist" draw when he drew on other gunmen. This was a very unconventional way to draw. Almost all other gunmen had their guns with the handles facing backwards, because they believed that this way of drawing was quicker. And in gunfighting, quickness is everything.

(Picture of Wild Bill with guns drawn shown for 2 seconds)

But this "underhand" or "twist" draw worked very well for Wild Bill. He was in a lot of gunfights, and he never lost. He always won. He killed a lot of men. He fought in the Civil War, and afterwards he worked as a Scout for General Custer and the US Cavalry. He worked as a sheriff in Abilene, Kansas, probably the most dangerous town in the West. As sheriff, he dispensed "frontier justice", which usually meant justice delivered with a gun. He was in gunfights in places all over the west, including Missouri, Nebraska, Kansas, and Wyoming. Wild Bill moved around a lot. Gunfighters can't stay too long in one place. Killing people often wears out a welcome.

No one really knows for sure how many men he killed, because it's very hard to distinguish fact from legend with Wild Bill. Newspapers and magazines wrote about him, and often greatly exaggerated his exploits. An example of this is the story of the gunfight at Rock Creek Station. Wild Bill, along with 2 other men, killed three men in a gunfight at Rock Creek Station in Nebraska. But a magazine story about the gunfight credited Wild Bill with killing 10 men all by himself. People liked to read stories about gunfighters and killings. The more killings the better.

Wild Bill was one of the most notorious and feared gunfighters in the West for almost ten years. But it all ended in 1874 when Wild Bill was shot and killed in a casino in Deadwood South Dakota. Wild Bill was playing poker with three other cowboys, when a man named Jack McCall snuck up behind him and shot him in the head. Wild Bill was holding cards with a pair of aces, and a pair of eights, and this hand has come to be known as "Dead Man's Hand". Another Wild Bill legend to add to the list.

Text 3 - David and Amy

A: Hi David, what's up?
D: Hey Amy. How's it going?
A: Oh, pretty good, except I have so much studying to do for mid-terms.
D: Tell me about it. This Biology class I'm taking is killing me.
A: Yeah? My friend Tina is in that class too. Do you know her? She's always complaining about how much work it is. I'm glad I'm not taking it.
D: You're friends with Tina?
A: Yeah, why?
D: Have you talked to her since class on Monday?
A: No, why? What happened?
D: Oh, it was crazy. I don't know if I should be telling you, but I guess everybody knows about it anyway. You know that class is right after lunch, after everyone has just eaten, and sometimes you can get pretty sleepy in there. Anyway, the professor was lecturing, and he was going on and on and on, when all of the sudden there was this big crash in the back of the room, and everybody turned around to see what happened. And Tina was lying on the floor. Apparently she'd fallen asleep, and fell out of her desk.
A: Oh no.. What did the professor do?
D: It WAS bad. Tina kind of got up, and said she was sorry, and then a bunch of people started laughing. The professor was kind of mad, and said it was nothing to laugh about, and then he just sort of started lecturing again. I don't think she'll be falling asleep in class again anytime soon.
A: Oh, that's horrible, the poor thing. She must have felt awful, and you're laughing at her.
D: I don't mean to be mean. It's just that it was pretty funny. You had to be there.
A: Well I don't think it's so funny. I hope she's all right.