A Construct Validation Study of the Extended Listening Sections of the ECPE and MELAB

Elvis Wagner
Teachers College, Columbia University

Numerous taxonomies of second language listening ability have been created and hypothesized by researchers, yet few of these taxonomies have been empirically validated. For this study, a model of second language listening ability based on Buck's (2001) default listening construct was developed and used to investigate the construct validity of the extended listening sections of the Michigan English Language Assessment Battery (MELAB) and the Examination for the Certificate of Proficiency in English (ECPE). Analyzing the data using internal consistency reliability analysis and exploratory factor analysis provided limited empirical evidence in support of the hypothesized model.

The testing of second language (L2) listening ability presents special problems for test developers. When assessing L2 listening ability, it is first necessary to define what exactly L2 listening ability entails. However, providing an adequate definition of L2 listening ability is no easy task, and is problematic for test developers. Numerous researchers (Buck, 1991, 2001; Buck & Tatsuoka, 1998; Dunkel, Henning & Chaudron, 1993; Richards, 1983; Rubin, 1994) have described the necessity of defining the concept of L2 listening ability, yet an adequate definition is still elusive. In fact, providing a global, comprehensive definition may be impossible, in part because so many different processes and variables are involved in L2 listening ability.

Perhaps because of the inherent difficulty in providing a comprehensive definition of L2 listening, a number of taxonomies of listening comprehension skills have been created by researchers to describe the listening process (Aitken, 1978; Lund, 1991; Peterson, 1991; Richards, 1983; Weir, 1993). While these taxonomies are important for L2 listening teachers, their utilization by researchers and test developers 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 et al., 1993, p. 182). In fact, Buck (2001) criticized these taxonomies with his observation that what are described as sub-skills in these models are, in fact, skills. About these "sub-skills," Buck stated that the "research seems to suggest that we are able to identify them statistically at almost any level of detail" (p. 59), while in actuality these taxonomies are essentially hypothetical in nature, since there has been little research to examine and validate them.

The lack of a global, comprehensive definition of L2 listening ability, as well as the limited usefulness of listening taxonomies, presents serious problems for L2 listening test developers in operationalizing a construct definition for their tests. Fortunately, a number of researchers have addressed this issue in order to assist test developers in creating reliable and valid assessments. Buck (2001) gave a list of recommendations to be used when creating a listening construct, which he referred to as his "default listening construct" (p. 113). This default construct included focusing on the assessment of those skills that are unique to listening (e.g., phonological modification, accent, prosodic features, non-verbal signals); testing listeners using a variety of texts on a variety of topics; using longer texts that test discourse skills, 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 gave 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). In this default listening construct, the first component (the ability to process extended samples of realistic spoken language, automatically and in real time) refers to both text and task characteristics. The second and third components (the ability to understand explicitly stated information, and the ability to understand implied information), however, refer to different aspects of a learner's listening ability. This idea of listening ability involving the ability to comprehend both explicit and implicit information is found throughout the L2 listening literature.

Similar to Buck (2001), Brindley (1998) also described 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. He stated that the testing of listening ability presents problems for the language tester, because listeners use higher and lower level processing simultaneously when interpreting a text, and thus it is difficult to attribute the listeners' responses on a test to any one skill. In addition, Brindley questioned the feasibility of testing a listener's ability to make inferences, since "interpretations may differ between individuals according to a multiplicity of cognitive and affective factors" (p. 173). Brindley went on to cite a number of researchers who advocated the use of items that test the ability of listeners to comprehend non-literal meanings, and stated that care needs to be taken when constructing inference items, because text interpretations are always subjective, and consequently the test developer needs to design items to constrain the possible responses.

Nissan, Devnicenzi, and Tang (1996) also found evidence for the implicit/explicit information distinction. They investigated factors that affected the difficulty of dialogue items in the listening comprehension section of the Test of English as a Foreign Language (TOEFL). They identified 17 different variables they hypothesized might affect the difficulty of listening items. In a subsequent statistical analysis of listening items, however, only five of the 17 originally hypothesized variables were found to have a statistically significant effect on item difficulty. These five variables were: utterance pattern, negative in stimulus, word frequency, inference, and role of speakers. They identified the "inference" variable as being related to "whether the information tested is explicitly or implicitly stated in the stimulus" (p. 8). Explicit items were ones in which the answer to an item was explicitly stated in the stimulus, or a close paraphrase of the answer was explicitly stated, while implicit items were those which it was "necessary to go beyond what is actually stated in the stimulus" (p. 8). They cited taxonomies by Richards (1983) and Rost (1990) that distinguished between the skills of comprehending explicit and comprehending implicitly stated information. Nissan et al. hypothesized that items that tested implicit information would be more difficult for test-takers than the items that tested explicit information because "a text with an implicit proposition requires more complex processing than one in which the proposition is explicit" (p. 9). The data supported the hypothesis that testing implied information would be more difficult than testing explicit information. They hypothesized that this might be a result of "some inherent difference in the item type associated with testing inferencing" (p. 28), a hypothesis in line with other taxonomies of L2 listening ability that distinguished between these two types of processing.

Hansen and Jensen (1994) also studied the role of explicitly stated and implicit information in listening tests when they examined a listening test called the T-LAP, which included the use of two authentic academic lectures: a history and a chemistry lecture. One focus of their study was how listeners of different ability levels would be able to answer global versus detail questions. The researchers hypothesized that lower ability listeners would have more problems with global questions because of the necessity of utilizing implicit information found in the lectures. Their hypothesis was confirmed when they found that the lower ability listeners did more poorly on global questions than they did on detail questions in comparison to the higher ability group. They found evidence that lower ability level listeners relied on verbatim responses in answering their questions, perhaps because they did not have the ability to process the implicit information provided in the academic lectures.

Finally, Wagner (2002) described a model of L2 listening ability similar to Buck (2001) and Brindley (1998). Wagner created a video listening test based on a two-factor model of listening ability, with one factor corresponding to "top-down processing," and the other factor corresponding to "bottom-up processing." Top-down processing was operationalized with four different types of items (listening for gist, making text-based inferences, making pragmatic inferences, and deducing vocabulary through context), while bottom-up processing was operationalized with two different types of items (identifying details and facts, and recognition of supporting ideas). The test was then administered to 75 high school ESL students. Wagner used exploratory factor analysis (EFA) to examine the results of this test, and did not find evidence to support the theoretical model. Instead, he found that the items on a video listening test loaded on two factors, one factor corresponding to the ability to listen for information explicitly stated in the text, and the second factor corresponding to the ability to listen for implicit information. Wagner concluded that a two-factor model of listening ability similar to Buck's (2001) listening construct was consistent with the data he analyzed. The interfactor correlation matrix indicated a moderate correlation of 0.515, which he concluded was not unexpected, because listening for explicitly stated information, and listening for implicit information, can be seen as two separate, but interrelated, abilities.[1]

While the idea of "the ability to listen for explicitly stated information" is self-defining, "the ability to listen for implicit information" requires a very wide-ranging (and consequently problematic) definition, and can include many different kinds of inferencing. Hildyard and Olson (1978) classified 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

[1] Another study which found evidence for this implicit/explicit information distinction was Purpura (1999), although the focus of this study was on reading ability. In studying the reading section of the First Certificate in English (FCE) Anchor Test, Purpura found that a two-factor solution for the passage comprehension section of the reading test best represented the data, with the two factors corresponding to "reading for explicit information" and "reading for inferential information."

interpretations of the speakers and the text). Buck and Tatsuoka (1998) included low-level bridging inferencing, higher-level reasoning, and using background knowledge. They also described a particular type of inferencing ability as being "text-based." Text-based inferencing mirrors the propositional and enabling inferences described by Hildyard and Olson.

Another type of inferencing ability found in the L2 literature includes 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 Tatsuoka (1998) referred 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).

Examination for the Certificate of Proficiency in English and the Michigan English Language Assessment Battery

This study examines the underlying construct of the extended listening sections of the Michigan English Language Assessment Battery (MELAB) and the Examination for the Certificate of Proficiency in English (ECPE). The MELAB and ECPE are two assessment instruments in the English Language Institute's (ELI) Testing and Certification Division at the University of Michigan. Although designed for somewhat different purposes (the MELAB is intended to measure English language proficiency for admission to North American colleges and universities, while the ECPE provides test-takers with an official certificate that indicates proficiency in English for use in the examinee's home country), the two tests are similar in that they both are aimed at advanced-level learners, and are designed to measure the following language abilities: speaking, listening, writing, reading, and lexical grammar.

The listening sections of the MELAB and the ECPE are composed of similar listening tasks. The MELAB includes tasks based on listening to questions, tasks based on listening to short statements, tasks based on listening to phrases or short questions spoken with special emphasis, and tasks based on listening to a short lecture. The ECPE includes tasks based on listening to questions, tasks based on listening to short conversational exchanges, and tasks based on listening to more extended talk on different topics.

The purpose of this study is to investigate the extended listening sections of the MELAB and ECPE. Specifically, this study investigates whether the MELAB and ECPE test learners' ability to listen for explicit stated information, and the ability to listen for implicit information, a construct that is commonly found in the L2 listening literature.

The current study addresses the following research questions:

1. To what extent do the items in the extended listening section of the MELAB perform as a homogenous group?
2. Does the MELAB assess test-takers' ability to listen for explicitly stated information, and the ability to listen for implicit information?
3. To what extent do the items in the extended listening section of the ECPE perform as a homogenous group?
4. Does the ECPE assess test-takers' ability to listen for explicitly stated information, and the ability to listen for implicit information?

Method

Participants

MELAB

The data are from the 1999 administration of the MELAB, administered at approximately 40 test centers in the United States and Canada. The participants included 823 learners of English as a second language. The participants spoke more than 60 first languages, with the largest groups being native speakers of Tagalog (33.1%), Farsi/Persian (13.6%), Chinese/Cantonese/Mandarin (9.4%), Spanish (8.3%), Korean (4.6%), Russian (2.7%), Tamil (2.6%), "Other" Asian Languages (2.4%), Arabic (2.3%), English (1.3%), Portuguese (1.2%), and Serbo-Croatian (1.1%).

Females made up approximately 58.7% of the examinees (N = 483), and males, 41.3% (N = 340). The mean age of the examinees was 28.0 years.

ECPE

The data are from the 1999 administration of the ECPE, administered at 114 test centers throughout the world. The participants included 17,099 learners of English as a foreign language. Of these, the majority (80.2%) of the test-takers were native Greek speakers (N = 13,718). 12.8% of the test-takers spoke Portuguese as their first language (N = 2,190), 5.1% spoke Spanish as their first language (N = 870), and 0.7% spoke Arabic as their first language (N = 118).

Approximately 65.5% (N = 11,212) of the test-takers were male, while 33.5% (N = 5,726) were female, and about one percent (N = 161) of the test-takers did not report their gender. The mean age of the test-takers was 21, and the median age was 19. Over all, almost 82% of the test-takers were under 25 years of age.

The MELAB Test

The MELAB was developed by the ELI at the University of Michigan, and is intended to measure English language proficiency. Test-takers are adult learners of advanced English ability who take the test in order to meet admission requirements to North American colleges and universities, or for non-native English speakers who need to demonstrate English language proficiency for employment purposes. Test-takers have approximately 150 minutes to complete the four sections of the test: written composition; listening comprehension; grammar/cloze/vocabulary/reading; and speaking (optional). The description of the test can be seen in Table 1.

Extended Listening Section of the MELAB

The listening section of the MELAB form used for this study is composed of four different sub-sections. In the first sub-section (10 items), the test-taker hears a question, and has to choose the best answer to that question. In the second sub-section (18 items), the speaker hears a statement, or a very short conversation, and the test-taker has to choose the answer which means about the same thing as the statement or conversation that was presented. The third sub-section (7 items) involves emphasis items, in which the test-taker hears a statement that is spoken in a certain way with a special emphasis. The test-taker must choose the answer that tells what the speaker would probably say next. The fourth sub-section of the listening section is composed of two parts that can be described as

Table 1. Description of the MELAB

Section	Task Type	Time (minutes)	Number of Items
Writing	200-300 word composition	30	1
Listening	MC	30	50
GCVR		75	100
Grammar	MC		(30)
Cloze	MC		(20)
Vocabulary	MC		(30)
Reading	MC		(20)
Speaking	Oral Interview	10-15	1

"Extended Listening." In the first part a short lecture of about two minutes long is given, and test-takers must then complete five multiple-choice comprehension questions on the lecture. In the second part test-takers hear a longer interview, in the form of a radio "news magazine," of about three minutes in length, and test-takers must answer ten multiple-choice comprehension questions on the text. The description of the listening section of the MELAB is given in Table 2.

Table 2. Description of the Listening Section of the MELAB

Sub-Section Text Type	Number of Items	Text Length
Question	10	One sentence question
Short Text	18	
Statement	(9)	One sentence statement
Two-person dialogue	(9)	Two-turns
Emphasis	7	Incomplete sentence
Extended Listening	15	
Lecture	(5)	Approximately 2 minutes
Radio "News magazine"	(10)	Approximately 3 minutes

In the item specifications for the extended listening section of the MELAB, the test developers describe how the items on this section of the test aim to assess comprehension of extended discourse, including the understanding of the main points, supporting details or examples, and the relationship between the different elements of the text. Questions may also require the test-taker to distinguish a point from an aside, understand the metadiscoursal elements of the text, and comprehend both macro and micro-structures in the discourse, as well as demonstrate both global and local understanding. Importantly, the test developers also state that the questions might require the comprehension of "explicit propositions as well as implicit or inferred meaning" (J. S. Johnson, personal communication, January 31, 2003).

The ECPE Test

The ECPE was developed by the ELI at the University of Michigan, and is intended to measure English language proficiency. Test-takers are adult learners of advanced English ability who take the test in order to obtain an official certificate that indicates proficiency in

English for use in the examinee's home country. Test-takers have approximately 155 minutes to complete the five sections of the test: speaking; written composition; listening comprehension; reading/cloze; and grammar/vocabulary/reading. The description of the test can be seen in Table 3.

Table 3. Description of the ECPE

Section	Task Type	Time (minutes)	Number of Items
Speaking	Oral Interview	10-15	1
Writing	200-300 word composition	30	1
Listening	MC	25-30	40
Reading/Cloze	MC cloze	25	40
GVR		60	100
Grammar	MC		(40)
Vocabulary	MC		(40)
Reading	MC		(20)

Extended Listening Section of the ECPE

The listening section of the ECPE is composed of three different sub-sections. In the first sub-section (17 items), the test-taker hears a question, and has to choose the best answer to that question. The second sub-section (13 items) has two different types of texts. With the first text type, the speaker hears a very short conversation of either two or three turns, and the test-taker has to choose the answer that means about the same thing as the statement or conversation that was presented. The second type of text in this section is a longer two-person conversation. For the 1999 administration, this text has seventeen turns, and five comprehension questions. The text is broken up into two parts. The first part is presented, then two comprehension questions are given. The text then continues, and three comprehension questions are given. The third sub-section of the listening section (10 items) involves longer texts. Although in these texts there is more than one speaker, it is not an interactive conversation. Rather, they are presented as a radio program, in which more than one speaker presents the information. Each of these texts has five multiple-choice comprehension items connected with it. The description of the listening section of the ECPE is given in Table 4. For the purpose of this study, the “Extended two-person dialogue” and the “Radio News magazine” text will be considered “extended listening.”

Table 4. Description of the Listening Section of the ECPE

Sub-Section Text Type	Number of Items	Text Length
Question	17	One sentence question
Conversation		
Two-person dialogue	8	Two or three turns
Extended two-person dialogue	5	Approximately 20 turns
Radio “News magazine”	10	
Text 1	(5)	Approximately 3 minutes
Text 2	(5)	Approximately 3 minutes

No item specifications were available for the ECPE, but it is assumed that since the extended listening sections of the ECPE are similar to the extended listening section of the MELAB, the item specifications described earlier for the MELAB would also apply to the ECPE.

Procedures

MELAB

In 1999 the MELAB was administered by about 100 approved MELAB examiners around the world under uniform test administration procedures. All four sections of the MELAB were administered in a single administration period. First, the test answer sheets were distributed to test-takers, and then the actual test packet. Instructions for the overall test were read aloud in English by the test administrator. Test-takers wrote their compositions on the composition test form, and then filled in their responses to the listening and GCVR items on the answer sheet. At the end of the test, all test materials were collected, the compositions were rated by University of Michigan ELI raters, and the listening and GCVR answer sheets were machine scored.

ECPE

The 1999 ECPE was administered at over 120 testing sites in over 20 countries, under uniform test administration procedures. The writing, listening, reading, and grammar/vocabulary/reading sections were administered in one sitting. The speaking section was administered at a different time. First, the test answer sheets were distributed to test-takers, and then the actual test packet. Instructions for the overall test were read aloud in English by the test administrator. Test-takers wrote their compositions on the composition form, and filled in their responses to the listening, reading/cloze, and GVR items on the answer sheet. At the end of the test, all test materials were collected, and the answer sheets were machine scored.

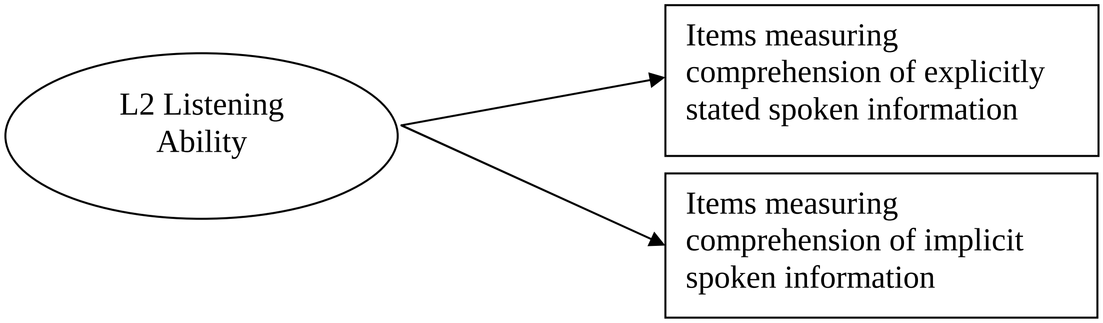
Analyses

Coding of the Items

Before performing the statistical analyses, the items in the extended listening sections of both the MELAB and ECPE were coded to determine what these items were measuring. The coding was based on a theoretical model of L2 listening ability proposed by Wagner (2002), which was in turn based on a construct definition of L2 listening ability given by Buck (2001). According to this model, listening ability can be seen as encompassing two factors: the ability to listen for explicitly stated information, and the ability to listen for implicit information. This model can be seen in Figure 1.

Using this model, the items in the extended listening section were categorized according to whether they required the ability to listen for explicitly stated information, or the ability to listen for implicit information. In this coding scheme, items that had the answer explicitly stated in the text were coded as *explicit* items, and items that required the listener to make an inference to answer correctly were coded *implicit*. This coding scheme matches the scheme found in Nissan et al. (1996). However, coding the items as *explicit* or *implicit* proved challenging, and it is necessary to go into some detail here to describe the coding process and provide examples. Unfortunately, because the test items examined for this study are still secure, the examples used here are not actual items from the MELAB or ECPE.

Figure 1. Operationalization of a Model of Second Language Listening Ability


L2 Listening Ability
Items measuring comprehension of explicitly stated spoken information
Items measuring comprehension of implicit spoken information

An item coded *explicit* is one in which the answer to the item is explicitly stated in the text. The following item is an example:

> *The striped skunk is the most common type of skunk found in North America. Skunks are omnivorous mammals, about 2 feet in length.*
> *Q: A skunk is a type of _____.*
> *a. bird*
> *b. insect*
> *c. reptile*
> *d. mammal**

The correct answer, *mammal,* is explicitly stated in the text, and thus is coded *explicit*. The next example item coded *explicit* is less obvious:

> *And humans have also benefited from the presence of skunks, because skunks eat bugs like grasshoppers and insect larvae that often eat human agricultural crops.*
> *Q: Grasshoppers and larvae are examples of _____.*
> *a. different types of skunks*
> *b. agricultural pests that skunks eat**
> *c. chemicals that cause a skunk's odor*
> *d. predators that sometimes eat skunks*

While the exact wording of the correct answer, *agricultural pests that skunks eat,* is not explicitly stated word for word in the text, the information found in the text (*skunks eat bugs like grasshoppers and insect larvae…*) is a close paraphrase or rewording of the correct answer, and thus is coded *explicit*. This idea of a "close paraphrase" will be examined more thoroughly after describing the coding of the *implicit* items.

Much of the research on listening for implicit information and making inferences in the L2 literature is based on the definitions of inferences given by Hildyard and Olson (1978).[2] They described three different types of inferences that listeners must make: propositional inferences, enabling inferences, and pragmatic inferences. Propositional

[2] In their study, Hildyard and Olson (1978) were describing the types of inferences that readers and listeners have to make.

inferences are “those which are the necessary implications of explicit propositions” (Hildyard & Olson, 1978, p. 92), and include transitive relations or syllogisms, and comparative terms. Examples of this type of inference include the syllogism:

A is bigger than B.
B is bigger than C.
A is bigger than C. (Propositional inference)

Another example includes the use of comparative terms:

Alex has more than Deb.
Deb has less than Alex. (Propositional inference)

Other types of propositional inferences according to Hildyard and Olson (1978) include the use of implicative verbs, and class inclusion relations. According to them, what all of these different types of propositional inferences have in common is that the inference follows from the form of the argument rather than from the content of the text.

A multiple-choice listening test item example might help to illustrate the idea of propositional inferences:

There are probably more skunks alive today than there were a thousand years ago. This is mostly because human development usually involves clearing the land of tree cover. This is fortunate for skunks, because they like to live in open areas.
Q: According to the speaker, skunks have _____.
a. been raised by humans as pets
b. evolved in the last thousand years
*c. benefited from the presence of humans**
d. almost gone extinct because of human development

This example can be seen as a type of syllogism. The listener hears, “Human development usually involves clearing the land of tree cover.” The listener also hears “Skunks like to live in open areas.” From these statements, the listener is able to make the inference “Skunks have benefited from human development,” which is a close paraphrase of the correct answer choice. It is important to note that with the propositional inference the listener does not need to utilize his or her world knowledge. Instead, the inference is a necessary implication drawn from the explicitly stated proposition. The inference is made from the form of the argument, rather than the content of that argument.

In contrast to propositional inferences, enabling inferences are those that allow the listener to link otherwise separate concepts in the text. They are made not as a result of an explicit, verbal proposition, but instead are made based on the listener’s knowledge of text structure (Hildyard & Olson, 1978), and these enabling inferences allow the listener to make sense out of the spoken utterances. Hildyard and Olson use the following example to illustrate (p. 94):

a. John threw the ball through the window.
b. Mr. Jones came running out of the house.
c. John broke the window (Enabling inference)

The listener's world knowledge (that a ball can break a window, and that the owner of the house would be angry and would try to find the person responsible), and knowledge of the structure of spoken texts (that lines *a* and *b* are somehow related, and not independent) allow the listener to make the enabling inference that John had broken the window. Another multiple-choice item is given below as an example of an enabling inference:

> *A skunk's odor works as a very good defense against predators. Skunks spray their musk at predators, causing a really bad smell.*
> *Q: Why is the skunk's smell a good defense against predators?*
> *a. The predators can smell the skunks.*
> *b. Skunks use their odor to blend in to their environment.*
> *c. Skunks do not taste very good to predators because their odor is so bad.*
> *d. The predators know that if they attack the skunks, they will end up smelling very bad.**

The listener is forced to rely on his or her world knowledge to make the enabling inference required to answer the question correctly. The listener hears:

> *a. Skunks spray their musk at predators*
> *b. This causes a really bad smell.*

From what he or she hears, the listener is able to make the following (rather extended) enabling inference to answer the question correctly:

> *c. The predators will end up smelling very bad if they attack the skunk. These predators will not want to smell bad, so they don't attack skunks. Therefore, this is a good defense mechanism for the skunk.*

Unlike propositional inferences, which are more or less independent of the content of the text, enabling inferences are made by listeners based on both the content and the structure of the spoken language (Hildyard & Olson, 1978).

The third type of inference described by Hildyard and Olson (1978) are pragmatic inferences. Like enabling inferences, they are made based on the listener's world knowledge, but unlike enabling inferences, they are not essential for interpreting the basic content of the spoken text. Instead, they elaborate on the information given in the text, and allow the listener to make more subjective interpretations, based on the listener's set of expectations about people, customs, and human behavior.

The three types of inferences described here all require the listener to utilize different knowledge sources. Hildyard and Olson (1978) summarize the three different types of inferences with the following:

> Propositional inferences are the necessarily true implications of explicit propositions. Enabling inferences are inferences which must be drawn to make discourse coherent and therefore, comprehensible. Pragmatic inferences are probably true, nonessential elaborative inferences. (p. 95)

Using the theoretical model given above, instructions about how to code items, as well as examples of *explicit* and *implicit* coded items, four doctoral students in the Applied Linguistics program at Teachers College, Columbia University, were asked to code the 30 items examined here. Of the 30 items coded, only seven of the ECPE items were coded similarly by all four of the coders, and nine of the MELAB items were coded similarly by the four coders. In addition to these 16 items that all four coders agreed on, six of the ECPE items were coded similarly by three of the four coders, and five of the MELAB items were coded similarly by three of the four coders. With these 11 items, the researcher closely examined the items in question, and in 10 of the cases it was decided to code the items as the majority had coded them. Two of the ECPE items had an even split, with two of the coders coding them *explicit*, and the other two coding them *implicit*. Similarly, one of the MELAB items had an even split of coders. The results of the coding can be seen in Table 5.

Table 5. Coding of the 30 MELAB and ECPE items

	Initial Coding	Final		Initial Coding	Final
Item	(4 raters)	Coding	Item	(4 raters)	Coding
MELAB 36	E	E	ECPE 26	I/E	I
MELAB 37	I	I	ECPE 27	I/E	E
MELAB 38	I/E	I/E	ECPE 28	I/E	E
MELAB 39	E	E	ECPE 29	I/E	E
MELAB 40	I/E	I	ECPE 30	I	I
MELAB 41	I/E	I	ECPE 31	E	E
MELAB 42	E	E	ECPE 32	E	E
MELAB 43	E	E	ECPE 33	I/E	I/E
MELAB 44	I	I	ECPE 34	I/E	E
MELAB 45	I/E	I/E	ECPE 35	I/E	I
MELAB 46	I	I	ECPE 36	E	E
MELAB 47	I/E	E	ECPE 37	I	I
MELAB 48	I/E	E	ECPE 38	I/E	E
MELAB 49	E	E	ECPE 39	I/E	I/E
MELAB 50	E	E	ECPE 40	I/E	I

Totals	Explicit	Implicit	Implicit/Explicit
MELAB	8	5	2
ECPE	8	5	2

The coding of the items was more difficult than anticipated, as is evidenced by the large number of disagreements between the coders. Part of this was due to the vagueness of the definitions in the literature. Deciding when an answer was implicit in the text, or if “a close paraphrase” (Nissan et al., 1996) was provided, proved difficult. Because of this, it was decided to code items 38 and 45 of the MELAB, and items 33 and 39 of the ECPE as double-coded, and to keep this double-coding in mind when examining how these items loaded on the particular factors.

Descriptive Statistics

Descriptive statistics for each of the test items in the listening sections of the MELAB and the ECPE were calculated and assumptions of normality were analyzed using SPSS version 10.0.5 for the PC. First, the mean, median, and standard deviation for each of the 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 assessments 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. To check if the items were normally distributed, the skewness and kurtosis for each of the items were analyzed. Because the statistical analyses employed in this study assumed a normal distribution, if an item had extreme skewness or kurtosis, it was considered for deletion from further analyses.

Reliability Analysis

A series of internal consistency reliability estimates for the overall listening sections of the MELAB (50 items) and the ECPE (40 items) using Cronbach's alpha were computed. All of these calculations were computed using SPSS version 10.0.5 for the PC. The internal consistency reliability estimates were also calculated for each of the sub-sections (including the extended listening sub-section) of the listening sections of the MELAB and the ECPE, again using Cronbach's alpha.

For the extended listening sections (composed of 15 items for each of the tests), item-total correlations were also computed. This was done in order to investigate how each individual item performed in relation to the other items in the extended listening sections. Those items that performed poorly were examined in order to determine if they should be considered for deletion from further analysis.

Exploratory Factor Analysis

After the items were coded, a number of exploratory factor analyses (EFAs) were performed in an attempt to examine the patterns of correlations among the items in order to explore the basic underlying factors of the extended listening sections of the tests. For this study, it was theorized that the factor analysis for each of the tests would result in a two-factor solution: one factor would correspond to the ability to listen for explicitly stated information, and the second factor would correspond to the ability to listen for implicit information. The items that were coded *explicit* should have loaded on a factor corresponding to the ability to listen for explicitly stated information, and the items that were coded *implicit* should have loaded on a factor corresponding to the ability to listen for implicit information.

The data from the extended listening sections of both of the tests were based on answers scored dichotomously, and thus the variables were treated as categorical. As a result, tetrachoric correlations were required in performing the EFAs. These EFAs were performed using Mplus for Windows, version 2.02, which computes tetrachoric correlations. First, the correlation matrix was prepared, and the determinant of the matrix was examined to determine the appropriateness of the data for factor analysis.

The EFAs were then performed, using unweighted least squares analysis (which is an appropriate analysis to use with categorical data) to extract the initial factors. The eigenvalues and the scree plots for the two sections were examined as indicators of the number of factors represented by the data, and then this information was used in an attempt to determine if the underlying factors represented by the data were consistent with the theoretical construct described earlier. 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, the interfactor correlation matrices were examined, and meaningful interpretations were made as the final criteria for deciding the best number of factors to extract.

Results

MELAB

Descriptive Statistics

The descriptive statistics for the item-level data from the extended listening sections of the MELAB were computed first, in order to investigate the dispersion of the scores, using SPSS version 10.0.5 for the PC. The overall mean for the extended listening section of the test was 9.64 out of a possible 15 (which equates to an overall mean of 64%), and a standard deviation of 3.25. The item facilities (IF) for the 15 items ranged from 0.33 to 0.93, which suggests a rather wide range of difficulty level for the various items. Three of the items (items 36, 38, and 46) were very difficult, with IFs of less than 0.40. Since there were only three options for the multiple-choice items, this indicates that the test-takers scored little better than chance on these three items. In addition, one of the items (number 50) had a very low difficulty level, with an IF of 0.93. Of these four items, only item 50 had extreme skewness (-3.48) and kurtosis values (10.10). Subsequent analyses were undertaken with the awareness that this item might be problematic because of the threat to a normal distribution that it posed. These values can be seen in Table 6.

Table 6. Descriptive Statistics for the Extended Listening Section of the MELAB

Item	Mean	Std Dev	Skewness	Kurtosis
36	0.367	0.482	0.553	-1.698
37	0.729	0.445	-1.033	-0.936
38	0.334	0.472	0.705	-1.507
39	0.520	0.500	-0.080	-1.998
40	0.634	0.482	-0.559	-1.692
41	0.672	0.470	-0.734	-1.465
42	0.831	0.375	-1.771	1.138
43	0.849	0.358	-1.957	1.833
44	0.734	0.442	-1.061	-0.877
45	0.660	0.474	-0.676	-1.547
46	0.393	0.489	0.441	-1.810
47	0.657	0.475	-0.664	-1.562
48	0.611	0.488	-0.457	-1.796
49	0.717	0.451	-0.965	-1.072
50	0.933	0.250	-3.476	10.104

Internal Consistency Reliability

An internal consistency reliability analysis was performed on the 50-item listening section of the MELAB, in order to examine how each listening item correlated with the other listening items. The internal consistency reliability for the overall listening section was high (α = 0.901).

A series of reliability analyses was then performed with the four different sub-sections that made up the listening section of the exam, in order to examine how the items in each sub-section correlated with the other items in that sub-section. The "Question" sub-section, which had the most items (18), had the highest internal consistency reliability α. The "Emphasis" sub-section, which had the fewest number of items (7), had the lowest α. This is not surprising, because the number of items is a factor in determining internal consistency reliability. While the reliability for the 15-item extended listening section (α = 0.766) was lower than the reliability for the entire listening section, it is still moderately high, especially considering that it is composed of only 15 items. These values can be seen in Table 7.

Table 7. Internal Consistency Reliability Estimates for the Listening Section of the MELAB

Sub-Section Text Type	Number of Items	Reliability α =
Question	10	0.757
Short Text	18	0.856
Emphasis	7	0.549
Extended Listening (Lecture and Interview)	15	0.766

Overall 50-item Listening Section Reliability α = 0.901.

Item-total correlations, which are an indication of the item's ability to discriminate between higher and lower scorers, were also computed for each of the items in the extended listening section. For these 15 items, only one item (number 38) had a very low item-total correlation (0.004). This is an item that was noted in the previous section because it had a very low IF (0.334). The low IF indicates that the item was very difficult for test-takers, and the low item-total correlation indicates that the item did not discriminate well between lower- and higher-ability test-takers. Because of this, item 38 was considered for deletion before further analysis.

Exploratory Factor Analysis

In order to investigate the trait structure of the extended listening section of the MELAB, a series of EFAs was conducted. First, a matrix of tetrachoric correlations was generated using the 15 items found in this section. The first EFA was then performed using all 15 items from the extended listening section of the MELAB. Another EFA was performed with 14 items, excluding item 50, which (as noted earlier) had extreme skewness (-3.476) and kurtosis (10.104) values. Kline (1998) recommends excluding items that have an absolute skewness value of more than 4, and an absolute kurtosis value of more than 8. Another EFA was conducted excluding items 50 and 38, because item 38 had a very low discrimination index (0.004), indicating that this item was not discriminating well between high and low scorers on the test.

Although it was hypothesized that the extended listening section of the MELAB would measure two factors, corresponding to the ability to listen for explicitly stated information, and the ability to listen for implicit information, the EFAs performed did not provide evidence for this hypothesis. Instead, a one-factor solution seemed to maximize parsimony and interpretability. All 13 items (items 38 and 50 had been deleted as noted earlier) loaded at 0.333 or higher on the single factor, as can be seen in Table 8. This would seem to indicate that the items on the extended listening section of the MELAB were measuring a single construct.

Table 8. Factor Loadings for the MELAB

Item	Loading
36	**0.333**
37	**0.576**
39	**0.437**
40	**0.555**
41	**0.687**
42	**0.965**
43	**0.886**
44	**0.760**
45	**0.730**
46	**0.447**
47	**0.662**
48	**0.535**
49	**0.353**

It was theorized that the items would be measuring two separate but related constructs, the ability to listen for explicitly stated information, and the ability to listen for implicit information. Therefore, even though a one-factor solution appeared to fit the data well, another EFA with two factors was performed. An examination of this two-factor solution indicated that this solution was not as parsimonious or interpretable as a one-factor solution. Because the two factors were moderately correlated (0.667), it was determined that using a Promax rotation to obtain an oblique solution was appropriate for these data. As can be seen in Table 9, ten of the 13 items loaded on factor 1 at the 0.3 level or higher, and six of the 13 items loaded on factor 2 at the 0.3 level or higher (items 40, 41, and 45 loaded on both factors).

After coding the items, it was hypothesized that items 37, 40, 41, 44 and 46 would load on one factor corresponding to the ability to listen for implicit information, and that items 36, 39, 42, 43, 47, 48, and 49 would load on another factor corresponding to the ability to listen for explicitly stated information (item 45 was double-coded). Examining the two-factor solution shown in Table 9, there is limited evidence supporting the hypothesized two-factor model described earlier. If factor one in Table 9 is interpreted as corresponding to the ability to understand explicitly stated information, it can be seen that the items 36, 39, 42, 43, and 47, which were all coded *explicit* items, loaded at the 0.3 level or higher on this factor. If factor two is interpreted as corresponding to the ability to understand implicit information,

Table 9. Factor Loadings 2-Factor Oblique Solution for the MELAB

		Loadings	
Item	Coding	F1	F2
36	E	**0.351**	-0.006
37	I	0.129	**0.544**
39	E	**0.439**	0.016
40	I	**0.309**	**0.303**
41	I	**0.371**	**0.391**
42	E	**0.927**	0.093
43	E	**0.889**	0.033
44	I	**0.531**	0.291
45	I/E	**0.492**	**0.301**
46	I	**0.430**	0.039
47	E	**0.584**	0.116
48	E	0.157	**0.458**
49	E	-0.261	**0.734**

items 37, 40, and 41, which were coded as *implicit* items, loaded at the 0.3 level or higher on this factor. In addition, item 45, which was double-coded because two of the coders thought it was *implicit*, and two of the coders considered it *explicit*, actually cross-loaded on both factors. This would seem to present evidence in support of the hypothesized model.

However, items 44 and 46, which were coded as *implicit* items loaded on factor 1, and items 48 and 49, which were coded as *explicit*, loaded on factor 2, which is the opposite of what was theorized. Because only nine of the 13 items loaded on the factors that corresponded to the hypothesized model, the data from the extended listening section of the MELAB provides only limited evidence in support of this model.

ECPE

Descriptive Statistics

The descriptive statistics for the item-level data from the extended listening sections of the ECPE were computed first, in order to investigate the dispersion of the scores. The overall mean for the extended listening section of the ECPE was 12.160 out of a possible 15 (which equates to an overall mean percent correct of 0.811), and the standard deviation was 2.240. The IFs for the 15 items ranged from 0.650 to 0.940, which suggested a narrower range of difficulty level for the various items than found in the MELAB. Four of the items (items 26, 28, 29, and 33) were very easy for the test-takers, all with an IF of more than 0.896. Of these four items, only item 33 had extreme skewness (-3.708), although all four items had inflated kurtosis values, as can be seen in Table 10. Because extreme skewness and kurtosis values indicate a threat to the normality of the sample, item 33 was considered for deletion before further analysis, when statistical techniques are used that require a normal distribution.

Internal Consistency Reliability

An internal consistency reliability analysis was performed on the 40-item listening section of the ECPE, in order to examine how each listening item correlated with the other

Table 10. Descriptive Statistics for the Extended Listening Section of the ECPE

Item	Mean	Std Dev	Skewness	Kurtosis
26	0.901	0.298	-2.691	5.241
27	0.871	0.335	-2.219	2.925
28	0.896	0.306	-2.587	4.692
29	0.911	0.285	-2.879	6.291
30	0.720	0.449	-0.981	-1.039
31	0.847	0.360	-1.931	1.727
32	0.830	0.376	-1.752	1.070
33	0.940	0.237	-3.708	11.750
34	0.825	0.380	-1.709	0.922
35	0.861	0.346	-2.083	2.338
36	0.697	0.460	-0.858	-1.264
37	0.723	0.447	-0.998	-1.004
38	0.650	0.477	-0.628	-1.606
39	0.777	0.416	-1.331	-0.228
40	0.712	0.453	-0.934	-1.264

listening items. The internal consistency reliability for the overall listening section was moderate ($\alpha = 0.768$).

A series of reliability analyses was then performed with the three different sub-sections that made up the listening section of the exam, in order to examine how the items in each sub-section correlated with the other items in that sub-section. The "Extended Listening" sub-section had the highest internal consistency reliability α, even though it had two fewer items than the "Questions" sub-section. The "Conversation" sub-section, which had the fewest number of items (8), had the lowest α. Again, this is not surprising, because the number of items is a factor in determining internal consistency reliability. The reliability for the 15-item extended listening section ($\alpha = 0.605$) was lower than the reliability for the entire listening section. These values can be seen in Table 11.

Table 11. Internal Consistency Reliability Estimates for the Listening Section of the ECPE

Sub-section Text Type	Number of Items	Reliability α =
Question	17	0.560
Conversation	8	0.539
Extended Listening[3]	15	0.605

Overall 40-item Listening Section Reliability $\alpha = 0.768$.

Item-total correlations were also computed for each of the items in the extended listening section. For these 15 items, three items had an item-total correlation lower than 0.200. Item 33 had an item-total correlation of 0.169, item 37 had an item-total correlation of 0.168, and

[3] For this analysis, the "Extended Listening" includes the "Extended two-person dialogue" sub-section, which on the test was part of the "Conversation" sub-section.

item 40 had an item-total correlation of 0.188. In addition, five items (28, 31, 32, 34, and 39) had item-total correlations ranging from 0.207 to 0.216. The relatively low item-total correlation values indicate that the items do not discriminate well between lower- and higher-ability test-takers, and also are responsible for the relatively low internal consistency reliability for the “Extended Listening” sub-section, and also contribute to the lower reliability for the overall listening section. Of these seven items with low item-total correlations, only items 28 and 33 were also flagged earlier for having inflated skewness and/or kurtosis values. These items had high IFs, (0.896 and 0.940, respectively), indicating that these items were relatively easy for the test-takers, which can probably account for the low item-total correlations. However, the other five items with low item-total correlations (31, 32, 37, 39, and 40) had lower IFs, ranging from 0.712 to 0.847.

Exploratory Factor Analysis

In order to investigate the trait structure of the extended listening section of the ECPE, a series of EFAs was conducted. First, a matrix of tetrachoric correlations was generated using the 15 items found in this section. The first EFA was then performed using all 15 items from the extended listening section of the ECPE. Another EFA was performed with 14 items, excluding item 33, which (as noted earlier) had extreme skewness (-3.708) and kurtosis (11.750) values.

Although it was hypothesized that the extended listening section of the ECPE would measure two factors, corresponding to the ability to listen for explicitly stated information and the ability to listen for implicit information, the EFAs performed did not provide evidence for this hypothesis. Instead, two possible solutions seemed to maximize parsimony and interpretability. Similar to the MELAB findings, a one-factor solution could be interpreted from the loadings on the EFA. Twelve of the 14 items loaded on factor 1 at a level of 0.3 or higher, and are shown in bold in Table 12. Only two of the items (37 and 40) loaded lower than 0.3 on the one-factor solution, although both of these items loaded on the factor at 0.233 or above.

Table 12. Factor Loadings 1-Factor Solution for the ECPE

Item	Loading
26	**0.755**
27	**0.686**
28	**0.474**
29	**0.668**
30	**0.518**
31	**0.393**
32	**0.348**
34	**0.372**
35	**0.420**
36	**0.436**
37	0.233
38	**0.355**
39	**0.324**
40	0.282

Again, as with the MELAB data, it was hypothesized that the items would be measuring two separate but related constructs, the ability to listen for explicitly stated information, and the ability to listen for implicit information. Therefore, another EFA was conducted, this time with a two-factor solution. Because the two factors were moderately correlated (0.583), it was determined that using a Promax rotation to obtain an oblique solution was appropriate for these data. An examination of the EFA with two factors indicated that this solution was not the most appropriate solution for these data. As can be seen in Table 13, five of the 13 items loaded on factor 1 at the 0.3 level or higher, and eight of the 13 items loaded on factor 2 at the 0.3 level or higher (items 31 did not load at the 0.3 level on either factor).

After coding the items, it was hypothesized that items 26, 30, 35, 37, and 40 would load on one factor corresponding to the ability to listen for implicit information, and that items 27, 28, 29, 31, 32, 34, 36, and 38 would load on another factor corresponding to the ability to listen for explicitly stated information (item 39 was double-coded). The two-factor solution shown in Table 13 cannot be interpreted as providing evidence supporting the theorized two-factor model described earlier. The items that were coded *implicit* did not load on a particular factor, but instead seem to have loaded more or less equally on both factors. Similarly, the items coded *explicit* also seem not to have loaded on one particular factor.

Table 13. Factor Loadings 2-Factor Solution for the ECPE

		Loadings	
Item	Coding	F1	F2
26	I	**0.930**	-0.069
27	E	**0.746**	0.013
28	E	**0.415**	0.100
29	E	**0.616**	0.112
30	I	**0.417**	0.149
31	E	0.181	0.259
32	E	0.049	**0.355**
34	E	-0.009	**0.449**
35	I	-0.025	**0.528**
36	E	-0.011	**0.532**
37	I	-0.139	**0.428**
38	E	-0.095	**0.529**
39	I/E	0.013	**0.367**
40	I	0.012	**0.317**

The pattern of loadings in Table 13, however, leads to another possible interpretation of the data, one that involves a 3-factor solution.

Another EFA was performed, this time with a 3-factor solution. According to the interfactor correlation matrix, all three of the correlations were 0.498 or higher, indicating that the use of a Promax rotation to obtain an oblique solution was the most appropriate for these data. As shown in Table 14, for the three-factor solution, five items (26, 27, 28, 29, and 30) load on factor 1 at a level of 0.3 or higher. Two items load on factor 2, and four items load on factor 3. Only 3 of the items (31, 32, and 37) do not load at the 0.3 level or higher on any of the factors. However, items 31 and 32 load at 0.228 and 0.279, respectively, on factor 2. These loadings are higher than those on factors 1 or 3.

Table 14. Factor Loadings 3-Factor Solution for the ECPE

		Loadings		
Item	Coding	F1	F2	F3
26	I	**0.955**	-0.129	0.016
27	E	**0.757**	-0.006	-0.002
28	E	**0.474**	0.101	-0.002
29	E	**0.628**	0.083	0.016
30	I	**0.427**	0.089	0.057
31	E	0.180	0.228	0.054
32	E	0.053	0.279	0.102
34	E	-0.049	**0.554**	-0.012
35	I	-0.105	**0.730**	-0.040
36	E	0.021	0.102	**0.463**
37	I	-0.124	0.231	0.227
38	E	-0.066	-0.019	**0.602**
39	I/E	0.054	-0.124	**0.509**
40	I	0.046	-0.022	**0.354**

A possible interpretation of this three-factor solution is found in the way the items cluster on the three factors. The extended listening section of the ECPE included three listening texts. The first text had items 26-30 accompanying it, the second text included items 31-35, and the third text included items 36-40. The three factors can be interpreted as representing each of the listening texts. As can be seen in Table 14, items 26-30 loaded at the level of 0.427 or higher on factor 1. Items 34 and 35 loaded strongly on factor 2, and items 31 and 32 loaded more strongly on factor 2 than on either of the other factors. Items 36, 38, 39, and 40 loaded strongly on factor 3. This solution would indicate that the extended listening section of the ECPE was measuring three different things, corresponding to the three different listening texts used in the test, and that the items were text dependent. This is very similar to the findings of Saito (2003) in her investigation of the reading section of the ECPE.

Discussion

The current study investigated four research questions pertaining to the underlying trait structure of L2 listening ability, as measured by the extended listening sections of the MELAB and ECPE. The first research question investigated the extent to which the items in the extended listening section of the MELAB performed as a homogenous group. The reliability analysis that was performed indicated that the 50 items in the overall listening section of the MELAB performed as a homogenous group. The overall internal consistency reliability α was 0.901, indicating that the items in the listening section are reliable. The internal consistency for each of the sub-sections was somewhat lower: 0.757 for the

"Question" sub-section; 0.856 for the "Short Text" sub-section; 0.549 for the "Emphasis" sub-section; and 0.766 for the "Extended Listening" sub-section. This would seem to indicate that these items are measuring reliably the different components of the construct.

The second research question examined the underlying trait structure of the extended listening section of the MELAB. A model of L2 listening ability was hypothesized, and then EFA was used to investigate if the MELAB assessed test-takers' ability to listen for explicitly stated information, and the ability to listen for implicit information. The two-factor oblique solution for the EFA that was performed seemed to present evidence in support of the hypothesized model. There seemed to be a meaningful pattern of variable loadings on the two factors: three of the five items coded *implicit* loaded on one factor; and five of the seven items coded *explicit* loaded on the other factor. In addition, the one item that was double-coded loaded on both factors. Still, four of the 13 items loaded on the opposite factor of what had been predicted.

The third research question investigated the extent to which the items in the extended listening section of the ECPE performed as a homogenous group. The reliability analysis that was performed indicated that the 40 items in the overall listening section of the ECPE performed as a fairly homogenous group. The overall internal consistency reliability α was 0.768, indicating that the items in the listening section are moderately reliable. The internal consistency measures for each of the sub-sections were somewhat lower: 0.560 for the "Question" sub-section; 0.539 for the "Conversation" sub-section; and 0.605 for the "Extended Listening" sub-section. This would seem to indicate that these items are measuring fairly reliably the different components of the construct. However, the internal consistency reliability α was markedly lower for the listening section of the ECPE than for the MELAB. Again, this may in part be due to the fact that the ECPE listening section (40 items) had ten fewer items than the MELAB listening section (50 items). It may also have to do with the fact that the sub-sections of the ECPE were somewhat different from the sub-sections of the MELAB. While both tests had a "Question" and "Extended Listening" sub-section, the ECPE also had a "Conversation" sub-section, while the MELAB had a "Short Text" and an "Emphasis" sub-section. It should also be noted that the "Question" sub-section of the ECPE had a much lower reliability α (0.560) for the ECPE compared with a 0.757 reliability α for this sub-section of the MELAB. Similarly, the α (0.605) for the "Extended Listening" section of the ECPE is lower than the reliability α (0.766) for the comparable section of the MELAB. Assuming that the item specifications for both of these tests were similar, investigating why the reliability of the corresponding sections of the two tests varied so markedly could be an informative area for further research.

The fourth research question examined the underlying trait structure of the extended listening section of the ECPE. A model of L2 listening ability was hypothesized, and then EFA was used to investigate if the ECPE assessed test-takers' ability to listen for explicitly stated information, and the ability to listen for implicit information. The EFA that was performed, however, presented no evidence in support of the hypothesized two-factor model. Instead, a one-factor or a three-factor solution seemed more interpretable for the data. The one-factor solution would seem to indicate that the extended listening section of the ECPE was measuring a single construct. The three-factor solution seemed to indicate that the items were text dependent. For the most part, the items clustered together with the rest of the items that were measuring the same text, and each factor could be interpreted as corresponding to a particular text.

Although the analyses of the EFAs conducted on the data from the MELAB provided only limited evidence in support of the hypothesized model, and the data from the ECPE provided little evidence, there are many factors that might explain at least some of the lack of supporting evidence. Firstly, the measures of reliability for the extended listening sections of the test were relatively low, at 0.766 for the MELAB, and 0.605 for the ECPE. Again, this is at least in part due to the small number of items (15) in these sections. Regardless of the cause, this amount of random measurement error makes it more difficult when using factor analysis to find evidence for the hypothesized model. Also, the fact that the items in the ECPE appeared to be text dependent is another factor that probably had an effect on the results of the EFA, and the investigation of this apparent text dependency could be an area in which further research might be fruitful.

Secondly, the coding of the items was a problematic process, and even though efforts were made to systematize the coding system, the process still included a degree of personal subjectivity. Although this is probably an inevitable part of any coding process, part of this inherent subjectivity could be due to the fine gradations of definitions of what making an inference actually entails. The definitions for the coding were based largely on Hildyard and Olson's (1978) work, in which the definitions of the three different types of inferences are thoroughly delineated, and numerous examples were given, and yet implementing these definitions in the coding process was still challenging, especially when trying to differentiate between an inference and a close paraphrase, as Nissan et al. (1996) described it. Also, the items from these tests were not operationalized on this model of listening ability. Instead, this model had been imposed on the items, and thus contributes to the difficulty of the coding process.

Thirdly, the distinction between the ability to listen for explicitly stated information and implicit information is at least to some extent an artificial one. Listeners still must be able to understand the explicitly stated words or phrases that will allow them to make the inference necessary to understand the implied information. At least on some levels, the ability to listen for implicit information cannot be separated from the ability to listen for explicitly stated information. The relatively high correlation (0.667) between the two factors in the two-factor solution of the MELAB may be an indication of how the two abilities might not be able to be truly separated. It may be that it is necessary for the listener to be listening for both explicitly stated and implicit information simultaneously, and trying to separate the two is an artificial division. Even with the difficulties mentioned above, there was some evidence supporting the hypothesized model found in the data from the EFA performed with the data from the extended listening section of the MELAB.

It must also be kept in mind that this study looked specifically at the extended listening sections of the tests. However, the listening section of the MELAB also had three other sub-sections ("Questions", "Short Text" and "Emphasis"), and the ECPE listening section had two other sub-sections ("Questions" and "Conversation") that were intended to test different aspects of the L2 listening process. Shin (2003) investigated the construct validity of the "Emphasis" sub-section of the MELAB, and further studies of this nature investigating the various sub-sections of the test would be useful and informative.

Conclusion

Buck (2001) urged that the taxonomies found in the L2 listening literature be treated with caution, because they were based on theory only, and had not been empirically validated. He also stated that there was an unspoken assumption that the sub-skills given in the taxonomies were in fact skills, and that the "research seems to suggest that we are able to identify them statistically at almost any level of detail" (p. 59). This study was an attempt to empirically validate a more general model of L2 listening ability, one that includes the ability to listen for explicitly stated information, and the ability to listen for implicit information. The results of the study mirror Buck's warning, that identifying statistically the different abilities involved in L2 listening is not an easy task. While only limited evidence was found here in support of the hypothesized model, the proposed model warrants further research in an attempt to allow L2 listening test developers to make more reliable and valid assessments.

References

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

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

Buck, G. (1991) The test of listening comprehension: An introspective study. *Language Testing, 8*(1), 67-91.

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*(2), 119-157.

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*(2), 180-91.

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*(1), 91-107.

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

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal, 75*(2), 196-204.

Nissan, S., Devnicenzi, F., & Tang, K. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report No. 51). Princeton, NJ: Educational Testing Service.

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

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*(2), 219-40.

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

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal*, *78*(2), 199-221.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1*, 39-82. Ann Arbor, MI: University of Michigan English Language Institute.

Shin, S. (2003). A construct validation study of emphasis type questions in the Michigan English Language Assessment Battery. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1*, 25-37. Ann Arbor, MI: University of Michigan English Language Institute.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University, 2/1,* Retrieved March 31, 2003, from http://www.tc.edu/tesolalwebjournal/wagner.pdf.

Weir, C. (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice-Hall.