Investigating the Construct Validity of the Grammar and Vocabulary Section and the Listening Section of the ECCE: Lexico-Grammatical Ability as a Predictor of L2 Listening Ability

Yen-Fen Liao
Teachers College, Columbia University

This study investigates the construct validity of the grammar and vocabulary (GV) section and the listening section of the Examination for the Certificate of Competency in English (ECCE), developed by the English Language Institute of the University of Michigan. Very few studies in the second language (L2) listening literature have examined the role of lexico-grammatical knowledge in listening comprehension. As a result, the current study focuses attention on exploring the issue of the relationship between lexico-grammatical knowledge and L2 listening ability in the context of the ECCE. The discussion first centers on the factorial structures of the GV and listening sections of the ECCE, and then turns to the role of lexico-grammatical knowledge in the prediction of L2 listening performance on the ECCE. A series of sophisticated statistical analyses provide ample evidence in support of lexico-grammatical knowledge as a critical predictor of L2 listening ability.

Over the years, the factors that contribute to second language (L2) listening comprehension have been of particular interest to L2 researchers and practitioners (for a review, see Buck, 2001; Rubin, 1994; Thompson, 1995). Among these factors influencing the process of understanding L2 aural input, lexico-grammatical knowledge has been widely identified as playing a critical role (Mecartty, 2000; Thompson, 1995). Lexico-grammatical knowledge allows L2 learners to derive literal meaning of the message and internalize the language structure so as to facilitate comprehension (Mecartty, 2000). Therefore, L2 learners with more linguistic knowledge and more automated processes are thought to consume less working memory, leaving more room to retain information and make inferences while listening (Vandergrift, 2004). Unquestionably, how lexico-grammatical knowledge is related to L2 listening comprehension deserves more research attention.

Lexico-grammatical knowledge can be simply defined as knowledge of the words and grammatical structures (Celce-Murcia & Larsen-Freeman, 1999). In other words, lexico-grammatical knowledge refers to lexical and grammatical knowledge. Recently, the roles of lexical and grammatical knowledge in L2 listening comprehension have emerged as a research focus. A large amount of survey research (e.g., Boyle, 1984; Hasan, 2000; Higgins, 1995; Teng, 2002; Yen, 1988) or qualitative research (e.g., Goh, 2000; Sun, 2002) have been conducted to examine L2 learners' listening problems and find out which factors most exacerbated their listening difficulties. For instance, on a survey of 30 English teachers and 60 learners, Boyle (1984) reported that lexical and syntactical knowledge were considered two of the most crucial factors in listening comprehension. Similarly, Hasan (2000), working with 81 Arabic learners, investigated their perceptions of English listening comprehension problems, and more than half of the students reported that unfamiliar words and difficult grammatical structures hindered their listening comprehension. In sum, these survey or diary studies have illustrated that lexical and grammatical knowledge are two essential factors which may exert considerable influence on listening comprehension.

Research Background

The Role of Lexical and Grammatical Knowledge in L2 Listening Comprehension

The crucial role of lexical knowledge in comprehending aural messages has been commonly accepted (Thompson, 1995). Rost (1990) highlighted the significance of lexical knowledge in processing aural input from the aspect of lexical effect. That is, when hearing an ambiguous or isolated word out of context, listeners tend to identify it as a plausible word. Once a word is recognized, listeners may activate their schemata to associate the word with related concepts and in turn better comprehend the text and predict the incoming messages.

Given that vocabulary plays a critical role in auditory perception, recent attention has been directed to examine whether lexical knowledge is one of the potential factors that cause listening comprehension problems. A number of research efforts (e.g., Goh, 2000; Hasan, 2000; Kelly, 1991; Sun, 2002) have examined this issue. For example, Kelly (1991) examined both EFL teachers' and students' comprehension problems and categorized their errors into three types: perceptual, lexical, and syntactical. The results showed that lexical errors accounted for 65.5% of all the errors where comprehension impairment occurred, suggesting that lexical ignorance was the main obstacle to listening comprehension with advanced learners. More recently, Hasan (2000) examined listening problems encountered by Arabic EFL learners, and found that they had difficulty in predicting missing words, which might be attributed to their word-by-word processing approaches and limited lexical knowledge.

Some qualitative studies have been undertaken to investigate learners' listening difficulties as well. For instance, Goh (2000) analyzed the data elicited from 40 ESL university students' self-reports on their listening difficulties. She found that failure to recognize words or to parse input efficiently caused listening comprehension problems. Along the same lines, Sun (2002) looked at 40 EFL learners' listening diaries and reported that about 70% of the students were unable to automatically associate sounds with words stored in their long-term memory, contributing to their listening problems.

Taken together, these studies have found that a lack of vocabulary was one of the primary causes which exacerbated listening difficulties. Nevertheless, a complete picture of the relationship between lexical knowledge and listening performance has not been obtained because most studies have explored the role of vocabulary in listening comprehension from either teachers' or learners' perceptions of listening difficulties. As a result, no consensus on the role of lexis in L2 listening performance has been drawn.

In addition to lexical knowledge, grammatical knowledge has also been hypothesized to contribute to students' success in comprehending aural input (Thompson, 1995). Rost (1990) maintained that when a listener's attention is directed to a spoken text, grammatical information is available to constrain the syntactic structures of upcoming input. On the other hand, if listeners lack syntactic knowledge, they may have trouble segmenting streams of speech with many words linked together (Sun, 2002). This has given rise to an ongoing discussion of the grammatical characteristics of spoken input (e.g., Glisan, 1985), and of how morphological and syntactic modifications make input more comprehensible (e.g., Blau, 1990; Cervantes & Gainer, 1992; Chaudron, 1983; Chiang & Dunkel, 1992; Kelch, 1985; Long, 1985; Pica, Young, & Doughty, 1987; Teng, 2001).

However, relatively few empirical studies have investigated the extent to which grammatical knowledge actually contributes to comprehension (Mecartty, 2000), or how well grammatical knowledge serves as a predictor of L2 listening ability. No conclusive answer has been found to the question: To what extent does a listening test also test grammar or vice versa? These issues indicate that there is still a need for more extensive research.

In fact, in the listening literature very few empirical studies have specifically explored the relationship between lexico-grammatical knowledge and L2 listening ability. One of these studies was carried out by Conrad (1985). She investigated whether nonnative listeners paid more attention than natives to syntactic information as opposed to semantic information in listening. The findings showed that native listeners used primarily semantic units to process spoken input, while nonnative listeners tended to direct more attention to syntactic information.

More recently, Mecartty (2000) worked with 154 high beginners of Spanish and found that although both grammatical knowledge and lexical knowledge were significantly correlated with listening and reading comprehension, respectively, only lexical knowledge explained a significant proportion of the variance in reading and listening comprehension. Although both studies provide some insights into this issue, their preliminary results should be interpreted with caution because of some flaws in the research designs such as low reliability and validity of the measurement instruments used in both studies.

In sum, the research to date suggests that while the significance of lexico-grammatical knowledge in comprehending L2 input has received increased attention, much less effort has been expended on examining how lexico-grammatical knowledge contributes to L2 listening ability, particularly in the testing context. Clearly, more empirical research is needed to obtain a more complete picture of the role of lexico-grammatical knowledge in L2 listening test performance.

Defining the Construct of Lexico-Grammatical Knowledge

As mentioned earlier, lexico-grammatical knowledge generally refers to knowledge of words and grammatical structures (Celce-Murcia & Larsen-Freeman, 1999). Specifically, lexico-grammatical knowledge involves lexical form and meaning as well as syntactic form. Grammatical knowledge is another term often used interchangeably with lexico-grammatical knowledge. Grammatical knowledge refers to a set of informational structures related to grammatical form and meaning available for use in long-term memory (Purpura, 2004). According to Purpura’s (2004) model of grammatical knowledge, knowledge of words and structures involves two dimensions: form and meaning. In this respect, the two terms *grammatical knowledge* and *lexico-grammatical knowledge* are interchangeable. Therefore, for the current study, both lexico-grammatical knowledge and grammatical knowledge are interchangeably used.

Grammar has traditionally been considered as a syntactic system that decides how words are arranged in sentences. This view of grammar as form has been questioned by several researchers who posited that grammar involves not only formal patterns of the language in terms of morphology and syntax, but also meanings expressed through the use of forms (Bolinger, 1977) in certain language use contexts (Leech, 1983). In other words, linguistic accuracy is not the only concern to grammar. How meaning is conveyed by the use of linguistic forms and governed by pragmatic principles also needs to be considered.

Some researchers have attempted to propose preconceived models of grammatical knowledge to incorporate the different aspects of grammar. For example, in Rea Dickins' (1987, 1991, 1997) hierarchical model, grammatical competence is defined in terms of two components: knowledge of grammatical rules, and knowledge of rules in use. Knowledge of rules refers to morphosyntactic and semantic knowledge which concerns linguistic structures, word formation, and meaning conveyed by those linguistic forms. Knowledge of rules in use refers to the ability to use grammatical knowledge in an appropriate context for a communication purpose. In fact, Rea Dickins' (1987) notion of grammar is very similar to Leech's (1983) definition of language. In this regard, the distinction between language proficiency and grammatical knowledge has not been clearly identified by Rea Dickins (Chang, 2004; Purpura, 2004). The other unresolved issues involved in Rea Dickins' model is that the hierarchical relationship between knowledge of grammatical rules and of rules in use and their higher-order construct (i.e., grammatical competence) has not yet been empirically confirmed.

The other influential grammatical model is Larsen-Freeman's (1991) three-dimensional grammar framework for pedagogical purposes. In this model the complexity of grammar involves three dimensions of language: form/structure, meaning/semantics, and pragmatics. Specifically, grammatical structures "not only have a morphosyntactic form, they are also used to express meaning (semantics) in context-appropriate use (pragmatics)" (Celce-Murcia & Larsen-Freeman, 1999, p. 4). These three dimensions are interconnected and "are not hierarchically arranged as many traditional characterizations of linguistic strata depict," such as Rea Dickins' (1987) model (Larsen-Freeman, 1991, p. 280). Although some research efforts (e.g., Chang, 2004) have been made to inquire into these three dimensions, very little is known about the scope of each language dimension, the empirical distinctiveness among them, and how each dimension is interconnected with each other.

Different from Rea-Dickins' (1987) and Larsen-Freeman's (1991) models, Purpura (2004) attempted to separate grammatical knowledge from pragmatic knowledge. Grammatical knowledge, as depicted in Purpura's (2004) model, embodies two closely related but not identified dimensions: grammatical form and semantic meaning. Grammatical knowledge comprises grammatical form and semantic meaning, which are highly interrelated. Grammatical form involves grammatical structures associated with phonology, lexis, morphosyntax, cohesion, information management, and interaction on both the sentential and the suprasentential levels. Semantic meaning refers to the literal and intended meaning expressed by one or more grammatical forms. Recent research findings (e.g., Chang, 2004; Liao, 2006; Purpura, 2006; Saito, 2003) have lent empirical support to the plausibility of Purpura's (2004) model. Thus, the present study adopts Purpura's (2004) model to define the construct of lexico-grammatical knowledge as the mental representation of grammatical form and semantic meaning stored in long-term memory.

Defining the Construct of L2 Listening Ability

In the past decades, listening researchers have made a concerted effort to understand the nature of L2 listening ability from various perspectives. Influenced by a structuralist linguistics view of language ability in the 1960s, listening has traditionally been seen as a language skill, which involves recognition of the signaling linguistic elements in a communication context (Lado, 1961). Listening was perceived as an activity of receiving and recognizing information passively. In the late 1970s, the active role of a listener in communications was more precisely specified. Listening involved not only receiving input but also assigning meaning to that stimuli and developing anticipation for the coming information (Neisser, 1976; Wolvin & Coakley, 1982). In the early 1990s, L2 listening researchers tended to define listening in terms of information processing. L2 listening comprehension was viewed as an act of active information processing where listeners actively construct a mental representation of aural information (Morley, 1990; Rubin, 1990; Fischer & Farris, 1995). More recently, researchers (e.g., Buck, 2001) also defined L2 listening comprehension as an active process of receiving, attending to, and constructing meaning and distinguished listening skill from other language skills in terms of: (1) the acoustic input, (2) the real-time nature, and (3) the linguistic features of spoken texts.

In order to describe the complex listening process and define L2 listening ability in a more precise way, several researchers (e.g., Lund, 1990; Peterson, 1991; Richards, 1983) have created a number of taxonomies of listening comprehension skills. While these taxonomies may be useful for pedagogical purposes, their lack of empirical support limits their utilization by L2 listening researchers and test developers (Buck, 2001; Wagner, 2004).

The paucity of an agreed definition of the listening construct (Dunkel 1991; Joiner 1997), of limited usefulness of listening taxonomies (Wagner, 2004), and of adequate listening research (Rubin 1994) not only impedes work in L2 listening research (Mendelsohn, 1998), but also raises serious questions about operationalization of constructs for listening tests (Wagner, 2004). As a result, an increasing number of researchers have attempted to propose a theoretical model to characterize the components of L2 listening ability. For instance, Dunkel, Henning, and Chaudron (1993) proposed a model of L2 listening comprehension for assessment purposes in a social context. L2 listening ability was defined in light of the test taker's competence and test task characteristics. However, how these two components interact was not precisely specified in their model.

Recently, adopting both the competence-based approach and the task-based approach for different testing purposes, Buck (2001) argued that L2 listening comprehension should be defined in terms of the interaction between competence and test task. He posited the idea of "default listening construct" (p. 113), which he referred to as the ability to (1) process extended samples of realistic spoken language, automatically and in real time, (2) understand the linguistic information that is unequivocally included in the text, and (3) make whatever inferences are unambiguously implicated by the content of the passage (p. 114). This definition suggests that listening ability is composed of the ability to comprehend explicitly and implicitly stated information. This notion of explicit/implicit distinction is, in fact, similar to Sanford and Garrod's (1981) idea of explicit/implicit focus, and to Brindley's (1998) view of identifiable listening skills, involving lower order skills to process literal meaning and higher order skills to make inferences or evaluation.

The L2 listening literature has demonstrated some empirical evidence (e.g., Hansen & Jensen, 1994; Liao, 2006; Nissan, DeVincenzi, & Tang, 1996; Wagner, 2002, 2004) in support of this general model of L2 listening ability, with the ability to comprehend explicit information and the ability to comprehend implicit information. Nevertheless, more research on the validation of this model is needed to advance our understanding of L2 listening ability. Therefore, the current study bases the construct of L2 listening ability on this model and attempts to validate this preconceived model in the context of a large-scale standardized test.

Context of the Current Study

Most high-stakes, standardized L2 tests (e.g., General English Proficiency Test; Michigan English Language Assessment Battery; Test of English for International Communication) are comprised of parts that include a listening section and a grammar and vocabulary section. However, there is surprisingly a paucity of research on the factorial structures of these sections and on the issue of the degree to which lexico-grammatical knowledge can predict L2 listening test performance.

Therefore, the current study attempts to investigate the relationship between lexico-grammatical knowledge and L2 listening ability in the context of the Examination for the Certificate of Competency in English (ECCE), developed and scored by the English Language Institute of the University Michigan. The ECCE is administered twice annually at over 120 test centers all over the world. The test aims at measuring high-intermediate level test takers' English language performance in the language modalities of listening, reading, writing, and speaking. A special emphasis is put on the ability of the test taker to communicate effectively in English (ECCE Information Bulletin, 2006).

Similar to many large-scale standardized English tests, the ECCE comprises four sections: speaking, listening, GVR (grammar, vocabulary, reading), and writing. A test taker who passes each section of the ECCE will be awarded a certificate as evidence of high-intermediate competence in English for personal, educational, or employment purposes. Due to its popularity and high-stakes characteristics, the ECCE was used in the present study, functioning as a measure of L2 learners' lexico-grammatical knowledge and L2 listening ability.

Purpose of the Study and Research Questions

The current study first explores the underlying structures of the grammar, vocabulary, and listening test sections. It then investigates to what extent lexico-grammatical knowledge predicts L2 listening ability in the context of this test.

The following research questions are addressed: (1) What is the factorial structure of the ECCE grammar and vocabulary section? (2) What is the factorial structure of the ECCE listening section? and (3) What is the relationship between lexico-grammatical knowledge and L2 listening ability? More specifically, what is the relationship between knowledge of form and meaning and L2 listening ability across different question types (implicit vs. explicit items)?

Method

Participants

The data used in the current study are from the 2003 administration of the Examination for the Certificate of Competency in English (ECCE), administered at approximately 120 authorized test centers in Asia, Europe, and Latin America. This involved 42,507 learners of English as a foreign language (EFL) with diverse native languages. Almost all the examinees for this test in 2003 were included, except for a few with some missing data, accounting for less than one tenth of one percent. The test takers included both males and females at various ages.

The ECCE

The Examination for the Certificate of Competency in English (ECCE) is a standardized high-intermediate level EFL test, developed and scored by the English Language Institute of the University Michigan (ELI-UM). It is intended to measure students' communicative use of English rather than their formalistic knowledge of English. Thus, the language tested in the ECCE is general, not academic. Nevertheless, the ECCE may be viewed as a bridge to a more academically oriented exam such as the Michigan English Language Assessment Battery (MELAB) and the Examination for the Certificate of Proficiency in English (ECPE) (ECCE Information Bulletin, 2006).

The ECCE tests all four skill areas of the language. It comprises four test sections: speaking, listening, GVR (grammar, vocabulary, reading), and writing. Test takers have approximately 155 minutes to complete the four sections. Table 1 describes the format and content of the ECCE.

Table 1. Format and Content of the ECCE

Tasks	Time (minutes)	Number of Items
Speaking	15	1 task
Listening	30	50
Part I: Short Conversation		30
Part II: Radio Interview		20
Grammar, Vocabulary, Reading	80	100
Grammar		35
Vocabulary		35
Reading		30
Writing	30	1 task

While all the sections of the ECCE warrant in-depth investigations of their underlying constructs, the current study focuses only on the grammar and vocabulary and listening sections. The Grammar/Vocabulary/Reading (GVR) section consists of three subsections with 100 multiple-choice items: grammar (35 items), vocabulary (35 items), and reading (30 items). The test time is 80 minutes. The first two item types (i.e., grammar and vocabulary), consisting of incomplete sentences followed by a choice of words or phrases to complete them, were designed to measure EFL learners' lexico-grammatical knowledge. Given that the reading comprehension items were beyond the scope of this study, only the grammar and vocabulary (GV) subsections were examined. Hence, the GV section in this paper refers to the first two subsections in the GVR section of the ECCE.

The listening section is delivered via audio recording. It comprises two test item types with 50 multiple-choice items: short conversation (30 items) and radio interview (20 items). In the first type (i.e., short conversation), each short recorded conversation is followed by a question. Answer choices are shown as pictures. In the second type (i.e., radio interview), a recorded radio interview is broken into several segments, each followed by groups of questions, with brief printed answer choices. The test time is approximately 30 minutes.

Data Collection and Scoring Procedures

The ECCE is generally administered twice annually at ELI-UM–approved test centers around the world. The listening, GVR, and writing sections are given during a single administration period. The speaking section may be given before or after the test takers sit for the other sections of the ECCE, depending on the local test center's schedule.

Since this paper examines only the GV and listening sections, the data from other test sections were not collected or analyzed for the current study. It should also be noted that the data used in this study included only the items used for pass-or-fail decisions (i.e., some items were not scored to determine examinee proficiency). Consequently, only 24 grammar items, 22 vocabulary items, and 44 listening items were analyzed in this paper.

Assuming that intermediate-level learners typically do not learn all language skills at the same pace, the ECCE is scored using aggregate scoring, which makes it possible for test takers to fail one of the test sections and still pass the exam if their scores in other sections are well above the minimum passing level. Examinees who fail two or more sections will not pass the exam (ECCE Information Bulletin, 2006).

Generally, examinees must answer about 65% of the multiple-choice GVR and listening items correctly in order to pass those sections. Since this study examined only the GV and listening sections, with a special emphasis on the relationship between lexico-grammatical knowledge and L2 listening ability, the data for the entire test were not used. Instead, this study used 65% of the listening items correct (i.e., 29 out of the 44 listening items) as the cutoff score to classify the test takers into groups of masters (pass) and nonmasters (fail) in L2 listening comprehension. For this study, all the multiple-choice items were scored dichotomously, with only one key to each question.

Coding Procedures

Before the statistical analyses were performed, the ECCE GV items and the listening items were coded independently by four trained and experienced ESL teachers to determine what they appeared to measure. The reviewers were given an item coding packet with models of lexico-grammatical knowledge and of L2 listening ability and some example GV and listening items. The author explained the models to the reviewers and discussed the example items with them to ensure that the reviewers had an understanding of the coding scheme. They were then requested to code all the GV and listening items independently.

The GV items were coded based upon Purpura's (2004) model of grammatical knowledge, as illustrated in Figure 1. In this model, language knowledge comprises grammatical knowledge and pragmatic knowledge. Grammatical knowledge embodies two closely related dimensions: grammatical form and literal meaning. Knowledge of form and meaning relates to phonology, lexis, and morphosyntax at the sentential level and to cohesion, information management and interaction at the suprasentential level.

The items were categorized first in terms of form and meaning. Then, the coders indicated what aspect of form and meaning was substantively being tapped into. According to Purpura's (2004) model of grammatical knowledge, lexical form items measure knowledge of the features of words that encode grammar rather than those that reveal meaning, such as part of speech, word formation, a co-occurrence restriction, and so on. Lexical meaning items measure knowledge of literal meaning of the word and formulaic or lexicalized expressions. Another frequently tested component is morphosyntactic form, which refers to knowledge of the morphological and syntactic forms of the language, such as word order, sentence structure, mood, voice and modality. Morphosyntactic meaning items measure knowledge of literal meanings derived from morphosyntactic forms such as aspect and time, negation and agency, voice and word order, and so on. The other two components measured in the GV tasks are cohesive form and meaning. Cohesive form items measure knowledge of phonological, lexical and morphosyntactic features of the language on the sentence and/or the discourse levels. Cohesive meaning items measure knowledge of referential meanings derived from cohesive links.

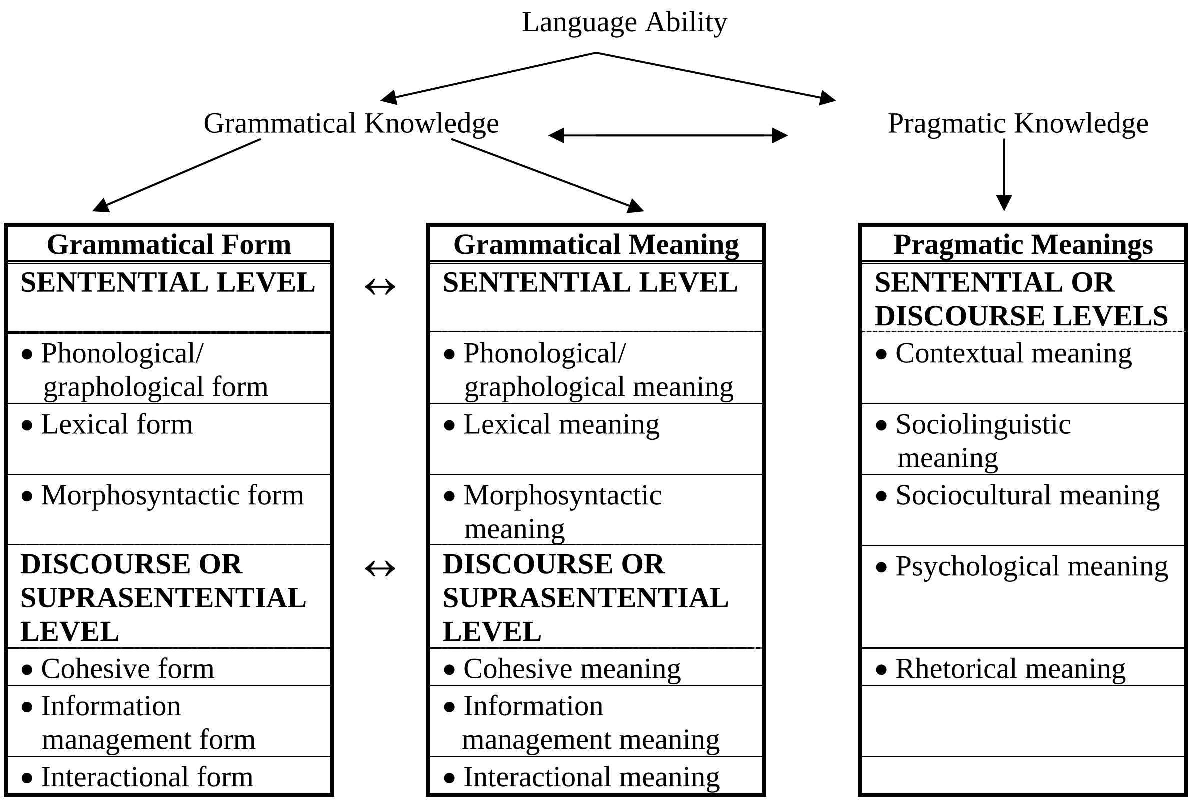


Figure 1. A Theoretical Definition of Language Ability (adapted from Purpura, 2004, p. 91)

Table 2 presents the final coding for the GV items. The Cohen’s Kappa agreement coefficient was 0.74 significant at the 0.001 level, indicating a high agreement between coders on item coding. Items G5 and G6 were coded differently by all four raters, and the author determined the final coding of these two items.

Although the GV items were categorized within the framework of form and meaning, it should be noted that the line between the components of form and meaning may be blurred and difficult to specify (Purpura, 2004). In fact, whether the GV items measure grammatical form and meaning separately or together is an empirical question to be answered as part of a test validation argument, which will be addressed in this study.

It should also be noted that most researchers tend to define knowledge of form as morphosyntactic knowledge, or simply grammatical knowledge and regard knowledge of meaning as vocabulary knowledge. In contrast, Purpura (2004) maintains that form and meaning involve more than morphosyntactic rules and lexical meanings. Although many studies may simplify lexico-grammatical knowledge to be vocabulary and morphosyntactic knowledge, the current study will operationalize the construct of lexico-grammatical knowledge through the form and meaning items, drawing on Purpura's (2004) definitions of grammatical knowledge.

Table 2. Coding of the GV Items

Components	Items		Totals
	Task 1 Grammar Task: Item G1-G24	Task 2 Vocabulary Task: Item V1-V22	
Form			24
Lexical form	G1, G 8, G10, G12, G15, G19, G22,		7
Morphosyntactic form	G2, G3, G4, G7, G9, G11, G13, G14, G16, G17, G20, G21, G23, G24		14
Cohesive form	G5, G6, G18		3
Meaning			22
Lexical meaning		V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19, V20, V21, V22	22

The listening items were coded based upon Buck's (2001) and Wagner's (2002, 2004) theoretical models of L2 listening ability, as graphically presented in Figure 2. In this model, L2 listening ability comprises two underlying traits: the ability to listen for explicitly stated information , and the ability to listen for implicitly stated information. In this coding scheme, explicit items refer to the items that have the answers explicitly stated in the spoken input, while implicit items pertain to those items that require the test taker to make an inference based on the form of the argument or on the content of the text in order to get the correct answer (Hildyard & Olson, 1978).

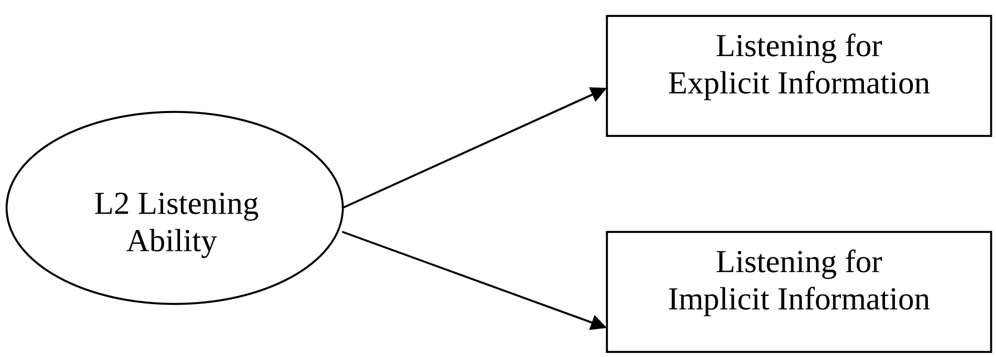


Figure 2. Operationalization of a Theoretical Model of L2 Listening Ability (adapted from Wagner, 2004)

Table 3 presents the results of the final coding. The intercoder agreement coefficient, as calculated using Cohen's Kappa, is 0.54 at the significant level of .001, indicating the coders were in moderate agreement. Items 1, 2, 4, 6, 17, 32, 33, 41, 44 were coded differently by the coders (i.e., these items were coded as explicit items by two of the coders and as implicit items by the other two). The author determined the final coding of these items.

Table 3. Coding of the Listening Items

Components	Listening items		Totals
	Task 1 Short Conversation: Item L1–L29	Task 2 Radio Interview: Item L30–L44	
Explicit	L7, L8, L9, L10, L11, L12, L13, L14, L15, L17, L18, L20, L21, L22, L23, L24, L25, L26, L27, L28, L29	L30, L31, L34, L35, L37, L38, L39, L40, L42, L43	31
Implicit	L1, L2, L3, L4, L5, L6, L16, L19	L32, L33, L36, L41, L44	13

Computer Equipment and Software

For this study a Pentium IV IBM clone was used for inputting and analyzing the data. The coding data were input and calculated using Microsoft Excel 2002. SPSS version 13.0 for Windows was employed to compute descriptive statistics, reliability analysis, and discriminant analysis. In addition, item-level exploratory factor analysis and all subsequent structural equation modeling were performed using Mplus version 3.01 for Windows (Muthén & Muthén, 2004).

Data Analysis

A series of statistical analyses were conducted to address the research questions, as illustrated in Figure 3. Descriptive statistics were first calculated to examine the central tendency, variability, and the variable distributions. Then, internal consistency reliability estimates using Cronbach's alpha were computed to obtain information on the consistency of measurement in the GV section and the listening section. After that, item statistics were computed to examine the difficulty and discriminability of individual test items.

To address the first two research questions concerning the factorial structure of the GV section and the listening section, item-level exploratory factor analyses (EFAs) were performed in an attempt to probe the underlying traits of the tests as well as the relative significance of each item as an indicator of its underlying factor. Because of the dichotomous nature of the variables, the matrix of tetrachoric correlations was first produced, and the determinant of the matrix was examined to assess the appropriateness of the data to proceed with the EFAs. Then, the EFAs were performed to extract the initial factors, using unweighted least squares analysis. The extractions were then rotated to an orthogonal solution using a varimax rotation and to an oblique solution using a promax rotation to enhance the interpretability of the factors. The best final solution achieved was based on simple structure and meaningful interpretation.

Subsequent to the EFAs, structural equation modeling (SEM) was performed to investigate the adequacy of the models that represent the hypothesized interrelationships between observed variables and latent variables and among latent variables (Purpura, 1999). SEM was conducted in five steps: model specification, identification, estimation, testing fit, and respecification (Bollen & Long, 1993). First, all the models under investigation were specified in advance at a conceptual level, based on substantive theory. These models were then examined in terms of their properties of identification, and were, in turn, converted into a number of mathematical equations, called statistical models. Following that, unknown parameters in these models were estimated. Each model was empirically assessed for the meaningfulness of model parameters and overall model-data fit, using fit indexes such as the χ^2 statistic, the Bentler Comparative Fit index (CFI), the Tucker-Lewis index (TLI), the Standardized Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA)[1] (Kline, 1998). Finally the model might be respecified, if necessary.

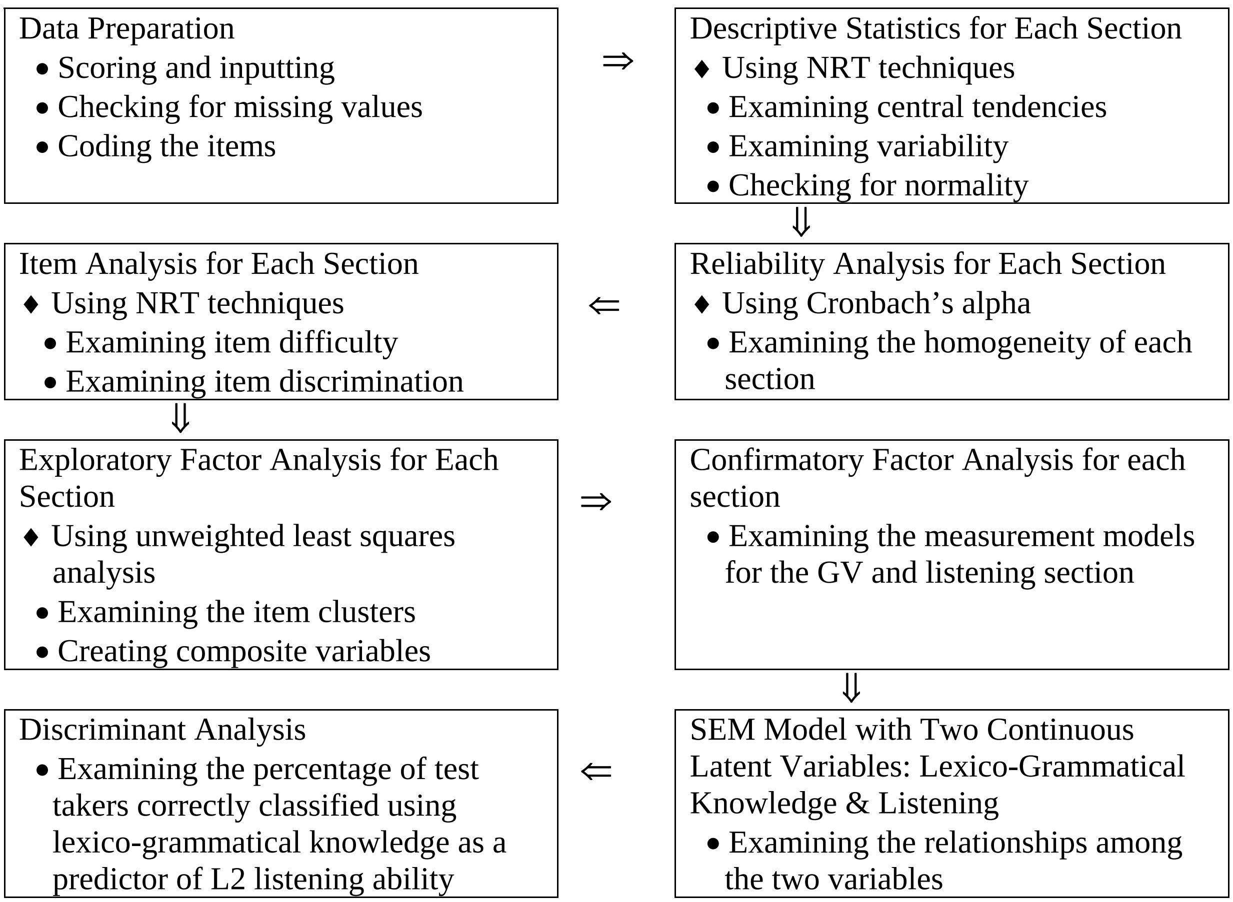


Figure 3. Flow Chart of Statistical Procedures Used in the Current Study (Adapted from Purpura, 1999)

[1] The criteria used for model-data fit are as follows: low and nonsignificant values of the χ^2 statistic; CFI greater than .95; SRMR less than .10; RMSEA less than .05 (Kline, 1998).

To address the third research question concerning the relationships between lexico-grammatical knowledge and L2 listening ability, a number of models were examined to probe the role of lexico-grammatical knowledge in listening comprehension.

Since the ECCE aimed to distinguish masters (i.e., examinees who pass the exam) from nonmasters (i.e., examinees who fail the exam), discriminant analysis using the discriminant function coefficient was performed. The purpose of discriminant analysis is to predict membership in two or more mutually exclusive groups (George & Mallery, 2003). In other words, discriminant analysis is used primarily to classify test takers into groups on the basis of a battery of measurements (Stevens, 2002). In the current study, discriminant analysis was performed to examine to what degree each predictor variable (form and meaning) correctly classified the test takers into groups of masters and nonmasters in listening.

Results

Descriptive Statistics

The descriptive statistics for the GV section and the listening section were first calculated. As shown in Table 4, the overall mean of the GV tasks was 30.92, out of 46 possible points (i.e., 67.22% correct) and the standard deviation was 7.95, indicating a wide distribution of scores. The median was 32 and the mode was 35. The fact that the most frequent score on the GV section was 76% correct suggests that this test was not very challenging. The minimum score on the GV section was 0 and the maximum score was 46, producing a wide range of 46, suggesting again a large degree of variability among the test takers' lexico-grammatical knowledge. The skewness was -0.49, indicating that there were more students who did well on the test than those who did not. The kurtosis was -0.37, indicating a flat distribution of scores which demonstrates a considerable amount of variability or heterogeneity of examinees' lexico-grammatical knowledge. Both values of skewness and kurtosis are within the acceptable limit $\pm$ 2 (Bachman, 2004), suggesting a normal distribution.

Table 4. Descriptive Statistics for the GV Section (N=42,507)

GV Tasks					
Mean	30.92	Minimum	0	Skewness	-0.49
Median	32	Maximum	46	Std. Error of Skewness	0.01
Mode	35	Range	46	Kurtosis	-0.37
Total possible	46	Std. Deviation	7.95	Std. Error of Kurtosis	0.02

With respect to the listening section, out of possible 44 points the overall mean of the listening section was 26.46 (i.e., 60.14% correct) and the standard deviation was 7.48. The median of the test was 26 and the mode was 25. The minimum score was 0 and the maximum score was 44, yielding a range of 44. Again, the wide range indicates a large degree of variability among the test takers' listening ability. The skewness was -0.01, implying that a higher frequency of test takers got higher scores. The kurtosis was -0.77, suggesting heterogeneity in terms of the test takers' listening ability. Nevertheless, both values of skewness and kurtosis were within the acceptable range, indicating that a normal distribution. These results can be seen in Table 5.

Table 5. Descriptive Statistics for the Listening Section (N=42,507)

Listening Section					
Mean	26.46	Minimum	0	Skewness	-0.01
Median	26	Maximum	44	Std. Error of Skewness	0.01
Mode	25	Range	44	Kurtosis	-0.77
Total possible	44	Std. Deviation	7.48	Std. Error of Kurtosis	0.02

Internal Consistency Reliability

An internal consistency reliability analysis was performed on the 46-item GV section of the ECCE. The reliability for the overall GV section was 0.87, suggesting a high level of internal consistency reliability (see Table 6). A number of reliability analyses were then performed on the two different GV tasks. Both the grammar task and the vocabulary task yielded moderately high reliability coefficients of 0.79 and 0.78, respectively.

Table 6. Internal Consistency Reliability Estimates for the ECCE GV Section

GV Section	Number of Items	Reliability Estimates (α)
Task 1: Grammar	24	0.79
Task 2: Vocabulary	22	0.78
Total	46	0.87

The internal consistency reliability of the entire ECCE listening section was also estimated using Cronbach’s alpha. As presented in Table 7, the reliability estimate was 0.85. The high internal consistency reliability estimate suggests that the listening items appeared to consistently measure the same construct. A series of reliability analyses were then performed on the two different listening tasks. The radio interview task produced a much lower reliability coefficient ($\alpha = 0.60$) than the short conversation task ($\alpha = 0.81$). The moderate degree of reliability estimate may be attributed to the limited number of items in the radio interview task where each item affected the reported estimate greatly.

Table 7. Internal Consistency Reliability Estimates for the ECCE Listening Section

Tasks	Number of Items	Reliability Estimates (α)
Task 1: Short Conversation	29	0.81
Task 2: Radio Interview	15	0.60
Total	44	0.85

Item Analysis

Following the reliability analysis, the item-level descriptive statistics for the 46 GV items were computed. As shown in Table 8, the means for the GV items ranged from 0.34 (Item V21) to 0.90 (Items G4 and V1), suggesting a wide range of item-difficulty levels. Six of the items (G4, G6, G7, G10, V1, V4) were easy for the examinees, all with a mean of 0.85 or higher, while three of the items (V14, V21, V22) were difficult, producing means of 0.45 or lower. In fact, the grammar items (mean = 17.10, Std. Dev. = 4.35) appeared to be easier than the vocabulary items (mean = 13.82, Std. Dev. = 4.29) for the examinees.

Most values for skewness and kurtosis were within the acceptable limit ± 2, indicating univariate normality. Some items (G4, G6, G9, G10, V1, V4) yielded large skewness and kurtosis values beyond the level of ± 2. Since all these items, except for Item G9, had a mean of 0.89 or higher, as noted earlier, the values for kurtosis and skewness of these items were expected to be high. It should be noted that although Item G9 (which was measuring knowledge of complex sentences) produced a moderately high mean of 0.70, the kurtosis (-2.54) of this item was large. These values can be seen in Table 8.

The item discrimination (ID) was then calculated, using corrected item-total correlation (point biserial correlation). The ID values for the GV items ranged from 0.04 (Item G16) to 0.46 (Item G14), indicating some items worked better than the others in discriminating high-ability examinees from low-ability examinees (see Table 8). Item G16 (which was measuring the infinitive) had a relatively low ID of 0.04, indicating that it did not discriminate well between lower- and higher-ability examinees, and also contributed to the lower reliability for the overall GV section. If this item were deleted, the Cronbach's Alpha would have increased from 0.873 to 0.876. Given that the increase of reliability coefficient was small, Item G16 was kept for the subsequent analyses.

Table 8. Descriptive Statistics for the 46 GV Items

Item	Original Coding	Mean	Std. Dev.	Skewness	Kurtosis	Item Discrimination (Corrected Item-Total Correlation)
G1	Form	0.67	0.47	-0.72	-1.49	0.40
G2	Form	0.84	0.37	-1.84	1.39	0.21
G3	Form	0.65	0.48	-0.65	-1.58	0.25
G4	Form	0.90	0.30	-2.62	4.85	0.40
G5	Form	0.83	0.38	-1.75	1.07	0.44
G6	Form	0.89	0.32	-2.43	3.89	0.34
G7	Form	0.85	0.36	-1.91	1.67	0.36
G8	Form	0.63	0.48	-0.55	-1.70	0.42
G9	Form	0.70	0.46	-0.87	-2.54	0.22
G10	Form	0.89	0.31	-1.25	4.47	0.30
G11	Form	0.79	0.41	-1.44	0.08	0.37
G12	Form	0.55	0.50	-0.19	-1.96	0.26
G13	Form	0.56	0.50	-0.26	-1.93	0.38
G14	Form	0.72	0.45	-1.00	-1.00	0.46
G15	Form	0.80	0.40	-1.52	0.30	0.40
G16	Form	0.64	0.48	-0.60	-1.64	0.04

G17	Form	0.50	0.50	0.02	-2.00	0.37
G18	Form	0.76	0.43	-1.20	-0.56	0.28
G19	Form	0.71	0.45	-0.94	-1.12	0.45
G20	Form	0.76	0.43	-1.24	-0.45	0.37
G21	Form	0.59	0.49	-0.36	-1.87	0.30
G22	Form	0.50	0.50	-0.01	-2.00	0.35
G23	Form	0.58	0.49	-0.33	-1.89	0.38
G24	Form	0.78	0.41	-1.38	-0.10	0.41
V1	Meaning	0.90	0.30	-2.72	5.41	0.21
V2	Meaning	0.71	0.46	-0.90	-1.19	0.43
V3	Meaning	0.48	0.50	0.07	-2.00	0.43
V4	Meaning	0.89	0.32	-2.43	3.90	0.30
V5	Meaning	0.79	0.41	-1.40	-0.03	0.37
V6	Meaning	0.73	0.45	-1.03	-0.95	0.33
V7	Meaning	0.80	0.40	-1.51	0.28	0.40
V8	Meaning	0.64	0.48	-0.59	-1.65	0.45
V9	Meaning	0.76	0.43	-1.22	-0.51	0.36
V10	Meaning	0.60	0.49	-0.41	-1.83	0.40
V11	Meaning	0.61	0.49	-0.46	-1.79	0.31
V12	Meaning	0.58	0.49	-0.34	-1.88	0.39
V13	Meaning	0.68	0.47	-0.77	-1.40	0.36
V14	Meaning	0.45	0.50	0.19	-1.96	0.20
V15	Meaning	0.57	0.50	-0.27	-1.93	0.27
V16	Meaning	0.48	0.50	0.09	-1.99	0.38
V17	Meaning	0.57	0.50	-0.27	-1.93	0.40
V18	Meaning	0.75	0.43	-1.16	-0.66	0.35
V19	Meaning	0.55	0.50	-0.21	-1.96	0.34
V20	Meaning	0.49	0.50	0.05	-2.00	0.34
V21	Meaning	0.34	0.47	0.67	-1.55	0.19
V22	Meaning	0.45	0.50	0.21	-1.96	0.25

Table 9. Descriptive Statistics for the 44 Listening Items

Item	Original Coding	Mean	Std. Dev.	Skewness	Kurtosis	Item Discrimination (Corrected Item-Total Correlation)
L1	Implicit	0.74	0.44	-1.08	-0.83	0.26
L2	Implicit	0.78	0.41	-1.36	-0.16	0.31
L3	Implicit	0.53	0.50	-0.11	-1.99	0.37
L4	Implicit	0.68	0.47	-0.77	-1.40	0.37
L5	Implicit	0.73	0.45	-1.01	-0.98	0.23
L6	Implicit	0.79	0.41	-1.42	0.02	0.41
L7	Explicit	0.65	0.48	-0.63	-1.60	0.42
L8	Explicit	0.77	0.42	-1.30	-0.31	0.35
L9	Explicit	0.41	0.49	0.38	-1.85	0.35
L10	Explicit	0.46	0.50	0.16	-1.97	0.27

L11	Explicit	0.66	0.47	-0.68	-1.54	0.44
L12	Explicit	0.40	0.49	0.42	-1.83	0.29
L13	Explicit	0.74	0.44	-1.09	-0.82	0.33
L14	Explicit	0.72	0.45	-0.99	-1.03	0.40
L15	Explicit	0.58	0.49	-0.32	-1.90	0.32
L16	Implicit	0.66	0.47	-0.68	-1.53	0.27
L17	Explicit	0.28	0.45	0.96	-1.07	0.20
L18	Explicit	0.43	0.50	0.30	-1.91	0.33
L19	Implicit	0.61	0.49	-0.45	-1.80	0.34
L20	Explicit	0.69	0.46	-0.83	-1.32	0.45
L21	Explicit	0.54	0.50	-0.18	-1.97	0.40
L22	Explicit	0.69	0.46	-0.80	-1.36	0.42
L23	Explicit	0.49	0.50	0.04	-1.94	0.24
L24	Explicit	0.62	0.49	-0.47	-1.78	0.33
L25	Explicit	0.39	0.49	0.45	-1.79	0.22
L26	Explicit	0.38	0.48	0.51	-1.74	0.24
L27	Explicit	0.42	0.49	0.33	-1.89	0.36
L28	Explicit	0.46	0.50	0.15	-1.98	0.34
L29	Explicit	0.70	0.46	-0.85	-1.21	0.30
L30	Explicit	0.45	0.50	0.20	-1.96	0.28
L31	Explicit	0.86	0.34	-2.11	2.45	0.25
L32	Implicit	0.67	0.47	-0.69	-1.09	0.36
L33	Implicit	0.78	0.42	-1.33	-0.23	0.18
L34	Explicit	0.59	0.49	-0.37	-1.86	0.08
L35	Explicit	0.64	0.48	-0.60	-1.64	0.35
L36	Implicit	0.60	0.49	-0.40	-1.84	0.18
L37	Explicit	0.65	0.48	-0.65	-1.58	0.16
L38	Explicit	0.70	0.46	-0.85	-1.28	0.28
L39	Explicit	0.42	0.49	0.35	-1.88	0.11
L40	Explicit	0.49	0.59	0.05	-2.00	0.27
L41	Implicit	0.74	0.44	-1.12	-0.75	0.37
L42	Explicit	0.54	0.50	-0.17	-1.97	0.30
L43	Explicit	0.72	0.45	-0.99	-1.02	0.41
L44	Implicit	0.63	0.48	-0.56	-1.69	0.37

Note.L1-L29 = short conversation, L30-L44 = radio interview.

The item-level descriptive statistics for the 44 listening items were also calculated. As presented in Table 9, the means for the listening items ranged from 0.28 (Item L17) to 0.86 (Item L31), suggesting a wide range of item difficulty levels. However, compared to the GV items, the listening section contained fewer items with high means above 0.85. Only Item L31 produced a mean above 0.85, at 0.86. Nine of the items (L9, L12, L17, L18, L25, L26, L27, L30, and L39) were difficult, producing means of 0.45 or lower.

All the items had values for skewness and kurtosis were within the accepted range except for Item L31 (see Table 9). Since Item L31 had a high mean of 0.86, the values for skewness (-2.11) and kurtosis (2.45) were expected to be high.

Corrected item-total correlations were computed to obtain the ID values for each listening item. As shown in Table 9, the ID values for the listening items ranged from 0.08 (Item L34) to 0.45 (Item L20). Among these 44 items, five items (L33, L34, L36, L37, and L39), all in the radio interview section, had ID values lower than 0.2. Item L34 had a very low ID value of 0.08, indicating that it did not function well in terms of discriminating examinees, and contributed to the lower reliability estimate for the overall listening section. The Cronbach's Alpha would have increased from 0.846 to 0.848 if Item L34 were deleted, and because the increase was small, Item L34 was kept for the subsequent analyses.

Exploratory Factor Analysis

A series of EFAs was conducted to investigate the trait structures of the GV section and the listening section. The EFAs were performed to determine the number of latent variables needed to explain the correlations among the items (Muthén & Muthén, 2004). A summary of the findings of the GV section is presented first, followed by a discussion on the EFA results of the listening section.

The first step was to analyze the 46 GV items together to examine how lexico-grammatical knowledge was measured. A two-factor promax rotation was produced to maximize parsimony and interpretability, suggesting the GV section measures two underlying traits. Based on the Purpura's (2004) model of grammatical knowledge, the two underlying factors can be hypothesized as form and meaning.

As shown in Table 10, most grammar items, coded as form items, loaded on the first factor (loadings larger than 0.30), which was hypothesized as the form factor. These items appeared to measure knowledge of grammatical form. This was expected because these items seemed to be designed to measure grammatical form, particularly lexical form, morphosyntactic form, and cohesive form. However, it should be noted that Item G22, coded as a form item, produced a significant loading (0.427) on the meaning factor, indicating that this item might be measuring semantic meaning rather than grammatical form. This result was actually not surprising because this item was measuring part of speech (e.g., "respect" vs. "respectful")[2], which involves literal meaning to some degree.

On the other hand, most vocabulary items, coded as meaning items, loaded on the second factor, which was hypothesized as the meaning factor. These items appeared to measure knowledge of semantic meaning. Again, this was expected because these items seemed to be intended to measure lexical meaning. Nevertheless, Item V4 was coded as a meaning item but loaded on the form factor, suggesting that this item was probably measuring grammatical form, not lexical meaning. A plausible explanation would seem to be that since the four answer choices of Item V4 convey similar literal meaning, examinees need to use their knowledge of lexical form to answer this item correctly.

Six items (G12, G16, G18, V14, V18, and V21) produced extremely low factor loadings (lower than 0.30). This problem may be attributed to the content of these items. Among these items, Items G12 and G18 were measuring prepositions, Item G16 was measuring the infinitive, and the others were measuring lexical meaning. All these items produced acceptable but large skewness or kurtosis values (beyond ± 1). There was a moderately high correlation (r = 0.714) between the two underlying factors (Table 11). As Purpura (2004) states, "the boundaries between the components of form and meaning may be difficult to specify" (p. 90). It was, therefore, not surprising to have items with double-

---

[2] The example used here is not from the actual item on the ECCE.

loadings (e.g., Item V18) or with higher loadings on an unanticipated factor (Items G22 and V4).

Another EFA was performed in order to examine the underlying traits of the listening section. The EFA produced a two-factor solution that appeared to maximize parsimony and interpretability, which provided evidence for the hypothesis that the ECCE listening section measures two factors, corresponding to the ability to listen for implicit information and the ability to listen for explicit information.

Although some did not, most items coded as implicit loaded on the first factor, which was hypothesized as the implicit factor (see Table 12). These items appeared to measure the ability to listen for implicit information. Also, most items coded as explicit produced large loadings on the second factor, which was hypothesized as the explicit factor.

However, it should be noted that a few items did not follow the same pattern. For example, Items L3 and L16 coded as implicit items loaded on the explicit factor. The same problem occurred to Items L8, L14, L30, L31, and L43, which were all coded as explicit items, but produced higher loadings on the implicit factor. The contradictions between the coding and the EFA results might be attributed to the coding problem (the moderate intercoder agreement coefficient, as noted earlier), the content of these items, or the correlation between the two traits. In fact, an inspection of the correlation matrix indicated that there was a moderate correlation ($r = 0.620$) between implicit factor and explicit factor (see Table 13).

Table 12 also shows some items (L4, L7, L11, and L20) produced double loadings, suggesting that these items might measure both implicit and explicit skills. This was not surprising because these items, all in the short conversation task, involve not only the ability to comprehend explicitly stated information but also the ability to understand implied information. Furthermore, it is worth noting that nine items (L1, L24, L29, L34, L36, L37, L38, L39, and L40) produced extremely low factor loadings (lower than 0.30). Six of these items (L34, L36, L37, L38, L39, and L40) were in the radio interview section. The low loadings might be due to the moderate reliability estimates for the radio interview section ($\alpha = 0.60$) which indicates that these items did not perform homogeneously.

In sum, it was hypothesized that the GV section of the ECCE measures two factors (form and meaning) and that the listening section of the ECCE measures two traits (explicit and implicit). The EFAs performed provides some evidence for these two hypotheses. Nevertheless, the EFA results should be interpreted with caution as all these items in EFA were allowed to correlate with each factor (Kline, 1998). The results, therefore, were just exploratory.

Table 10. Factor Loadings for the GV Section: Promax Rotation

Item	Original Coding	EFA Results	Factor 1 (Form)	Factor 2 (Meaning)
G4	F (MF)	F	<u>0.668</u>	0.074
G19	F (LF)	F	<u>0.618</u>	0.053
G15	F (LF)	F	<u>0.600</u>	0.051
G24	F (MF)	F	<u>0.585</u>	0.067
G8	F (LF)	F	<u>0.574</u>	0.041
G5	F (CF)	F	<u>0.534</u>	0.180
G6	F (CF)	F	<u>0.513</u>	0.099

G10	F (LF)	F	0.498	0.065
G14	F (MF)	F	0.494	0.190
G7	F (MF)	F	0.484	0.118
G11	F (MF)	F	0.484	0.106
V4	**M (LM)**	**F**	0.430	0.112
G3	F (MF)	F	0.423	-0.047
G2	F (MF)	F	0.392	-0.028
G1	F (LF)	F	0.376	0.208
G21	F (MF)	F	0.374	0.065
G13	F (MF)	F	0.359	0.196
G20	F (MF)	F	0.341	0.232
G23	F (MF)	F	0.336	0.215
G9	F (MF)	F	0.327	0.003
G17	F (MF)	F	0.322	0.222
V16	M (LM)	M	-0.109	0.662
V12	M (LM)	M	-0.078	0.646
V3	M (LM)	M	0.034	0.594
V17	M (LM)	M	-0.018	0.593
V7	M (LM)	M	0.110	0.527
V2	M (LM)	M	0.114	0.516
V10	M (LM)	M	0.125	0.463
V8	M (LM)	M	0.201	0.457
V9	M (LM)	M	0.114	0.450
V11	M (LM)	M	0.011	0.450
V13	M (LM)	M	0.107	0.435
G22	**F (LF)**	**M**	0.083	0.427
V14	M (LM)	M	-0.009	0.398
V20	M (LM)	M	0.108	0.375
V22	M (LM)	M	-0.004	0.371
V19	M (LM)	M	0.141	0.349
V5	M (LM)	M	0.249	0.342
V1	M (LM)	M	0.075	0.335
V6	M (LM)	M	0.184	0.319
V18	M (LM)		0.278	0.255
G18	F (CF)		0.273	0.164
G12	F (LF)		0.166	0.211
G16	F (MF)		0.069	-0.008
V21	M (LM)		0.038	0.235
V14	M (LM)		0.011	0.272

Note. LF = Lexical Form, MF = Morphosyntactic Form, CF = Cohesive Form, LM = Lexical Meaning

Table 11. GV Section Factor Correlation Matrix

	Form	Meaning
Form	1.000	
Meaning	0.714	1.000

Table 12. Factor Loadings for the Listening Test: Promax Rotation

Item	Original Coding	EFA Results	Factor 1 (Implicit)	Factor 2 (Explicit)
L6	Implicit	Implicit	0.518	0.178
L8	**Explicit**	**Implicit**	0.487	0.105
L31	**Explicit**	**Implicit**	0.474	-0.013
L14	**Explicit**	**Implicit**	0.460	0.180
L43	**Explicit**	**Implicit**	0.443	0.204
L41	Implicit	Implicit	0.424	0.173
L32	Implicit	Implicit	0.413	0.152
L19	Implicit	Implicit	0.408	0.114
L5	Implicit	Implicit	0.394	-0.014
L30	**Explicit**	**Implicit**	0.380	0.058
L44	Implicit	Implicit	0.355	0.214
L33	Implicit	Implicit	0.340	-0.049
L2	Implicit	Implicit	0.311	0.206
L9	Explicit	Explicit	-0.004	0.541
L15	Explicit	Explicit	0.035	0.453
L18	Explicit	Explicit	0.088	0.424
L12	Explicit	Explicit	0.041	0.396
L26	Explicit	Explicit	-0.029	0.391
L21	Explicit	Explicit	0.220	0.390
L22	Explicit	Explicit	0.272	0.381
L10	Explicit	Explicit	0.044	0.367
L28	Explicit	Explicit	0.156	0.365
L23	Explicit	Explicit	-0.006	0.364
L17	Explicit	Explicit	-0.045	0.356
L35	Explicit	Explicit	0.167	0.356
L27	Explicit	Explicit	0.223	0.337
L3	**Implicit**	**Explicit**	0.243	0.322
L25	Explicit	Explicit	0.019	0.312
L16	**Implicit**	**Explicit**	0.108	0.303
L42	Explicit	Explicit	0.161	0.291
L13	Explicit	Explicit	0.246	0.290
L20	**Explicit**	**Implicit & Explicit**	0.381	0.318
L11	**Explicit**	**Implicit & Explicit**	0.299	0.378
L4	**Implicit**	**Implicit & Explicit**	0.295	0.290
L7	**Explicit**	**Implicit & Explicit**	0.295	0.346
L38	Explicit		0.270	0.162
L24	Explicit		0.256	0.243
L29	Explicit		0.234	0.237
L1	Implicit		0.207	0.213
L36	Implicit		0.173	0.099
L37	Explicit		0.155	0.087
L40	Explicit		0.147	0.256
L34	Explicit		0.026	0.091
L39	Explicit		-0.030	0.183

Table 13. Listening Section Factor Correlation Matrix

	Implicit	Explicit
Implicit	1.000	
Explicit	0.620	1.000

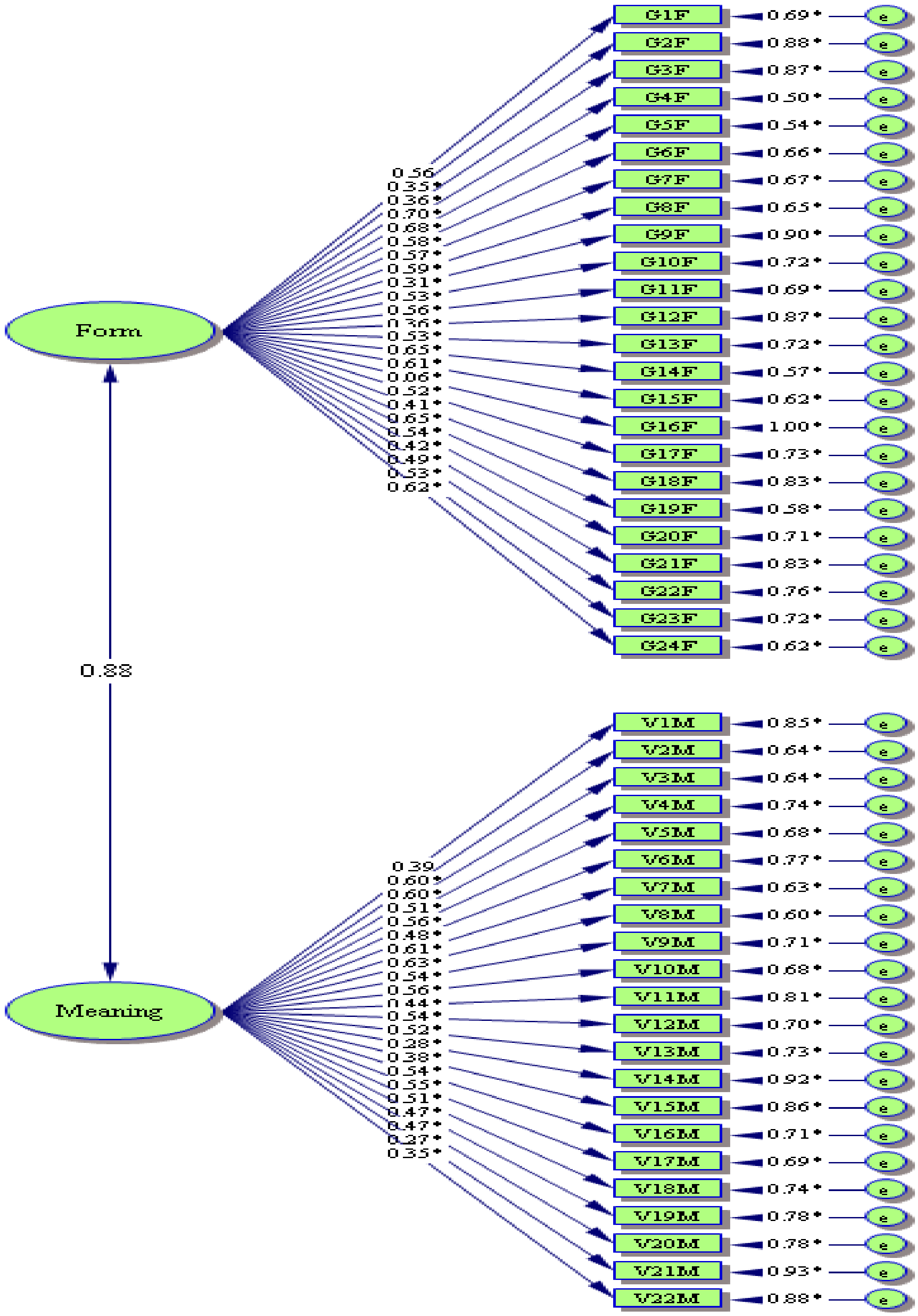
Structural Equation Modeling

Based on the results of the EFAs, the GV items appeared to measure two factors: grammatical form and semantic meaning. The listening section also appeared to have two underlying traits: listening for implicit information and listening for explicit information. However, there was not a very good correspondence between the original coding and the EFA results and the EFA only provided exploratory results (as noted earlier). Given that the item coding was grounded in substantive theories of language ability, item-level SEM, using the original item coding, was performed to obtain more information about the factorial structures of the GV and listening sections.

Since the model of lexico-grammatical knowledge using form and meaning as observed variables is underidentified (Kline, 1998), it is appropriate to perform item-level SEM using form and meaning as underlying factors while the items are treated as observed variables (as illustrated in Figure 4). Model 1 was evaluated to examine to what extent the model fit the sample data for the item level analysis. As can seen in Table 14, the model data fit statistics for Model 1 produced a Chi-square of 15276.484 with 909 degrees of freedom, significant at the .001 level, indicating a misfit of this model, which is not surprising because of the sensitivity of the χ^2 statistic. This model resulted in a CFI of 0.957, a SRMR of 0.029, and a RMSEA of 0.019. The values of CFI, SRMR and RMSEA are all favorable, indicating that this model represents the data well.

The feasibility of the individual parameter estimates was then evaluated. The results show that all the parameter estimates appeared to be substantively reasonable and statistically significant at the 0.05 level, indicating that the two underlying traits (form and meaning) were reasonably measured by the observed variables. As graphically represented in Figure 4, except for Item G16, almost all the factor loadings for Model 1 were moderate, ranging from a moderate of 0.31 to a high of 0.70. Item G16 produced an extremely low loading (0.06) on the form factor. This is actually expected because this item yielded a very low ID value (0.04). It should also be noted that the form factor was highly correlated with the meaning factor (r = 0.88), which is not surprising as both the form and meaning items were intended to measure the same construct (i.e., lexico-grammatical knowledge), and the boundary between form and meaning is sometimes blurred (Purpura, 2004).

In brief, Model 1 provided evidence in support of the two-factor solution of the GV section as a reasonable explanation of the correlations among the observed variables. In other words, the GV section measures two underlying factors: grammatical form and literal meaning.



* = Freely estimated
Figure 4. Model 1: Two-Factor Model of the GV Section

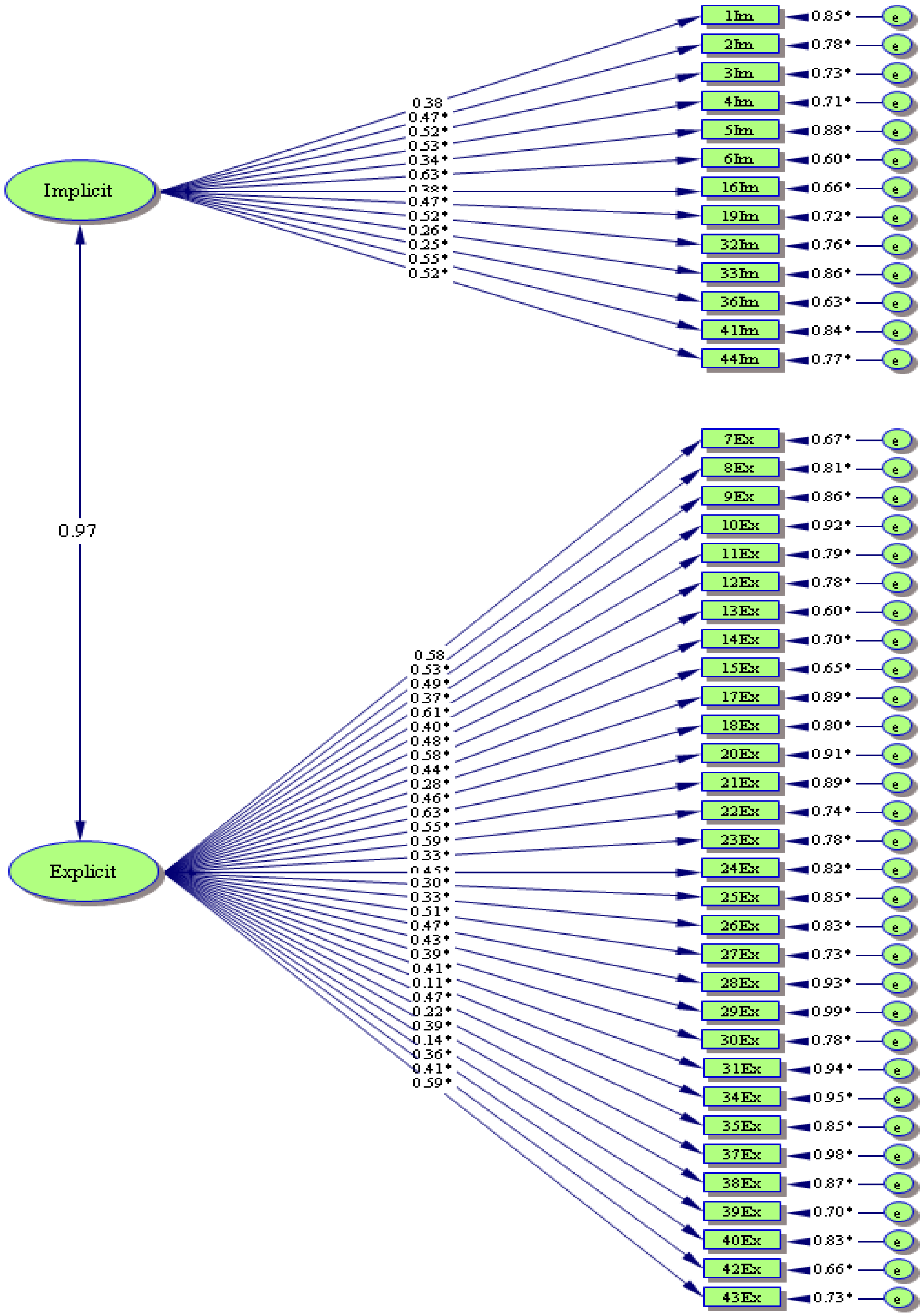
Table 14. Tests of Model Fit for the GV Section Two-Factor Model: Model 1

Goodness of fit summary:	
Comparative fit index (CFI)	0.957
The Tucker-Lewis index (TLI)	0.983
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.029
Root Mean Square Error of Approximation (RMSEA)	0.019
Chi-square test of model fit:	
Value	15276.484
Degrees of Freedom	909
P-Value	0.0000

Similar to the GV section, the EFA results also indicated that the listening section appeared to measure two factors: listening for implicit information and listening for explicit information. At first, the first-order confirmatory factor analysis was considered, using implicit and explicit as observed variables. However, there are only two observed variables in the one-factor model, so the model is underidentified which will produce unreliable parameter estimates. In order to solve the problem of underidentification, item-level SEM was implemented to investigate the factorial structure of the listening section. *Implicit* and *explicit* were viewed as underlying factors, while the listening items were treated as observed variables in this model (see Figure 5). Model 2 was evaluated to examine to what extent the model fit the sample data for the item level analysis. As presented in Table 15, the goodness of fit index for Model 2 yielded a Chi-square of 12682.544 with 850 degrees of freedom, significant at the 0.001 level, suggesting a misfit of this model. In fact, a significant Chi-square value seems not surprising because the χ^2 statistic is very sensitive to large sample size (N = 42,507 in this case). This model produced a CFI of 0.960, a SRMR of 0.026, and a RMSEA of 0.018, indicating that this model is a good representation of the sample data.

The feasibility of the individual parameter estimates was then evaluated. All the parameter estimates appeared to be substantively reasonable and statistically significant at the 0.05 level. This indicates that the two underlying traits (implicit and explicit) were reasonably measured by the observed variables. As illustrated in Figure 5, most factor loadings for Model 2 were somewhat moderate, higher than 0.3. Among the listening items, Items L34 and L39 produced extremely low loadings on the explicit factor (0.11 and 0.14, respectively). Again, this was not surprising because these two items yielded very low ID values (0.08 and 0.11, respectively). Besides, it is worth noting that the correlation coefficient between the implicit factor and explicit factor was relatively high (0.97), indicating these two factors are closely interrelated, but still not identical.

In sum, Model 2 provided evidence for the two-factor solution of the listening section as a reasonable explanation of the correlations among the observed variables. In other words, the listening section was measuring two underlying factors: listening for implicit information and listening for explicit information.



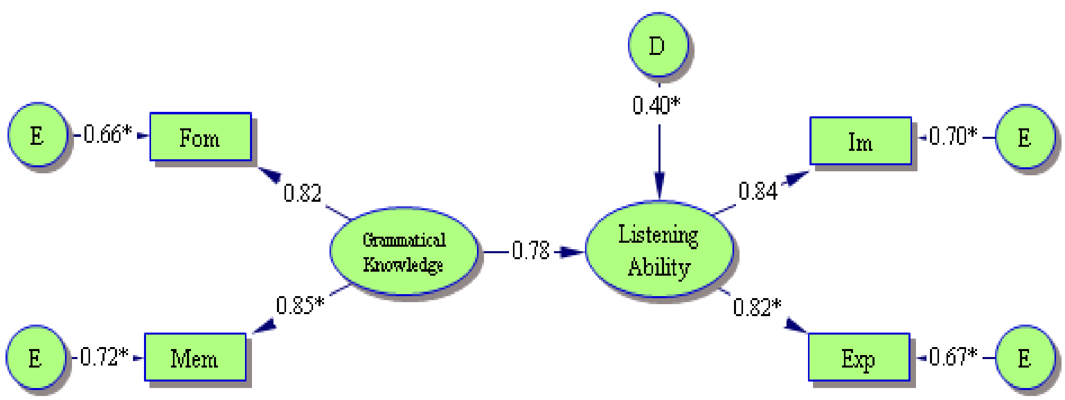
* = Freely estimated

Figure 5. Model 2: Two-Factor Model of the Listening Section

Table 15. Tests of Model Fit for the Listening Section Two-Factor Model: Model 2

Goodness of fit summary:	
Comparative fit index (CFI)	0.960
The Tucker-Lewis index (TLI)	0.979
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.026
Root Mean Square Error of Approximation (RMSEA)	0.018
Chi-square test of model fit:	
Value	12682.544
Degrees of Freedom	850
P-Value	0.0000

Since the second purpose of this study is to investigate the relationship between lexico-grammatical knowledge and L2 listening ability in the context of the ECCE, the SEM was implemented again to inquire into this issue. The SEM was first conducted using form and meaning as the factor indicators to examine how L2 listening ability regressed on the predictors. Although this model seemed plausible, the Mplus program was not able to produce the parameter estimates efficiently due to the relatively large sample size (N=42,507) and the large number of test items (k = 90). In order to solve this problem, composite scores of the 24 form items and of the 22 meaning items were computed separately, serving as the two indicators of the latent variable (lexico-grammatical knowledge). Composite scores of the 13 implicit items and of the 31 explicit items were also calculated separately, which were treated as the two indicators of the other latent variable (L2 listening ability). Following that, a regression model was examined using the SEM approach (see Figure 6).



* = Freely estimated
Figure 6. Model 3: Lexico-Grammatical Knowledge as Predictors

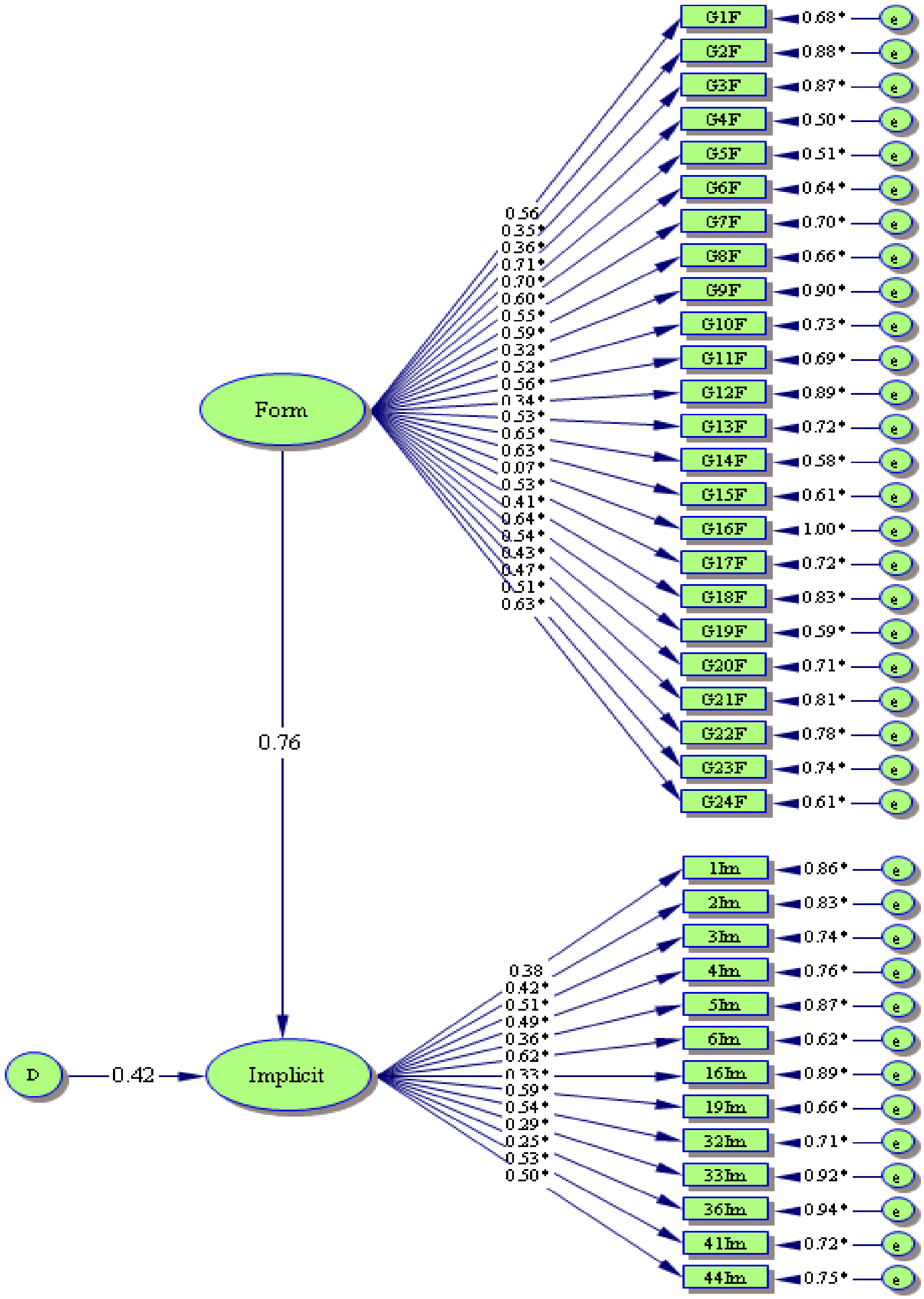
Table 16. Tests of Model Fit for the SEM Model: Model 3

Goodness of fit summary:	
Comparative fit index (CFI)	1.000
The Tucker-Lewis index (TLI)	0.998
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.002
Root Mean Square Error of Approximation (RMSEA)	0.025
Chi-square test of model fit:	
Value	26.629
Degrees of Freedom	1
P-Value	0.0000

Table 16 shows the model-data fit statistics. Model 3 yielded a significant value of the χ^2 index (26.629, df = 1, $p < .001$), indicating a model-data misfit. However, a perfect value of CFI (1.000), and relatively low values of SRMR (0.002) and RMSEA (0.025) suggest that Model 3 is a good representation of the sample data and provides evidence for lexico-grammatical knowledge as a good predictors of L2 listening ability. As depicted in Figure 6, L2 listening ability regressed positively on grammatical knowledge, with a standardized coefficient of 0.78, indicating that the test takers scores on the listening items would increase if they answered the GV items correctly.

Given that the GV items appear to measure the form factor and the meaning factor, and the listening items measure the implicitly stated information and the explicitly stated information, many different SEM models were further examined to investigate the relationship between knowledge of form and meaning and L2 listening ability across various question types (implicit vs. explicit items). However, most were underidentified or misfitting. In fact, a plausible but complicated SEM regression model on the item level was preferred, using form and meaning as the predictor variables, and implicit and explicit as the dependent variables. Unfortunately, this model was not examinable, perhaps in part because of the large sample size (42,507) and the large number of test items on each latent variable (90). As a result, the author performed separate SEM regression models on each latent variable, and these models seemed to work well.

First, the SEM regression model using form as the predictor variable and implicit as the dependent variable was examined (see Figure 7). Table 17 shows that the goodness of fit index for Model 4 produced a Chi-square value of 11209.670 with 579 degrees of freedom, significant at the 0.001 level. Although Model 4 yielded a significant χ^2 index, which indicates a misfit of this model, a high value of CFI (0.954), and low values of SRMR (0.031) and RMSEA (0.021) suggest that Model 4 represents the data well. As illustrated in Figure 7, the implicit factor regressed positively on the form variable, with a standardized coefficient of 0.76. The results suggest that the examinees' scores on the implicit items would increase for every correct response on the form items.



* = Freely estimated

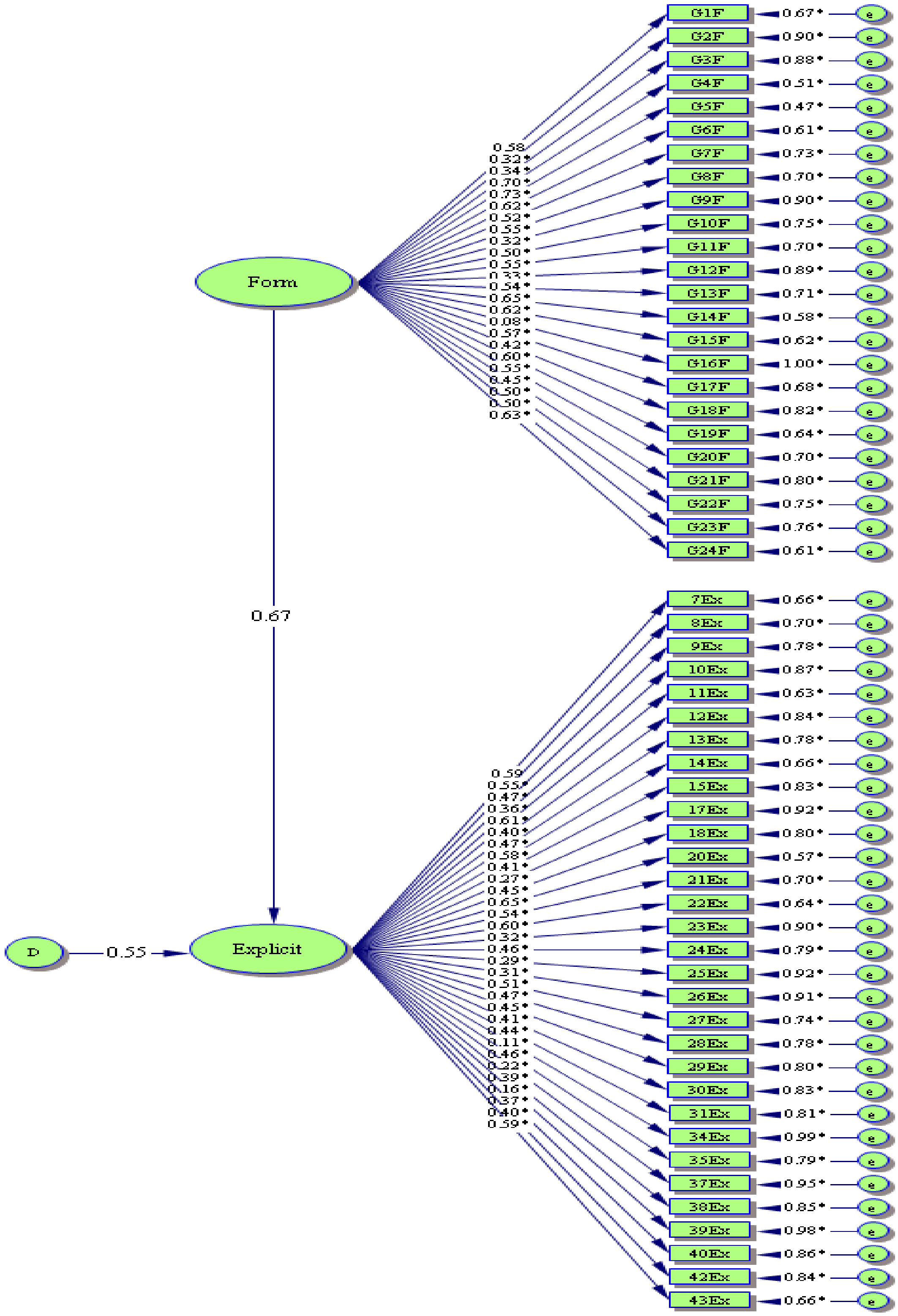
Figure 7. Model 4: Form as the Predictor and Implicit as the Dependent Variable

Table 17. Tests of Model Fit for the SEM Model: Model 4

Goodness of fit summary:	
Comparative fit index (CFI)	0.954
The Tucker-Lewis index (TLI)	0.976
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.031
Root Mean Square Error of Approximation (RMSEA)	0.021
Chi-square test of model fit:	
Value	11209.670
Degrees of Freedom	579
P-Value	0.0000

Next, the regression model using form as the predictor variable and explicit as the dependent variable was evaluated (see Figure 8). As shown in Table 18, the model-data fit statistics for Model 5 produced a Chi-square of 23895.311 with 1213 degrees of freedom, significant at the 0.001 level, indicating a misfit of this model, which is not surprising due to the sensitivity of the χ^2 statistic, as noted earlier. This model resulted in a CFI of 0.937, a SRMR of 0.032, and a RMSEA of 0.021. Although a higher value of CFI is preferred, the values of SRMR and RMSEA are favorable, indicating that this model represents the data well. As graphically represented in Figure 7, the explicit factor also regressed positively on the form variable, with a standardized regression coefficient of 0.67, suggesting that the test takers' scores on explicit items would increase in relation to their correct answers on the form items.

To summarize, the SEM results provided evidence in support of knowledge of form as a good predictor of L2 listening ability. Specifically, there was a positive close relationship between knowledge of form and the ability to listen for implicit and explicit information. Examinees who performed well on the form items tended to score higher on the listening items as well, but knowledge of form seemed to contribute more to the implicit items than to the explicit ones.



* = Freely estimated

Figure 8. Model 5: Form as the Predictor and Explicit as the Dependent Variable

Table 18. Tests of Model Fit for the SEM Model: Model 5

Goodness of fit summary:	
Comparative fit index (CFI)	0.937
The Tucker-Lewis index (TLI)	0.968
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.032
Root Mean Square Error of Approximation (RMSEA)	0.021
Chi-square test of model fit:	
Value	23895.311
Degrees of Freedom	1213
P-Value	0.0000

For the next analyses, the SEM regression model using meaning as the predictor variable and implicit as the dependent variable was examined (see Figure 9). As shown in Table 19, Model 4 produced a Chi-square value of 7870.094 with 526 degrees of freedom, significant at the 0.001 level. Again, it may be in part because the χ^2 index is very sensitive to large sample sizes. In fact, Model 6 yielded a high value of CFI (0.967), and low values of SRMR (0.026) and RMSEA (0.018), suggesting that this model fits the data well. As demonstrated in Figure 9, the implicit factor regressed positively on the meaning variable, with a standardized coefficient of 0.79, indicating that the test takers' scores on the implicit items would increase for each correct response on the meaning items.

* = Freely estimated

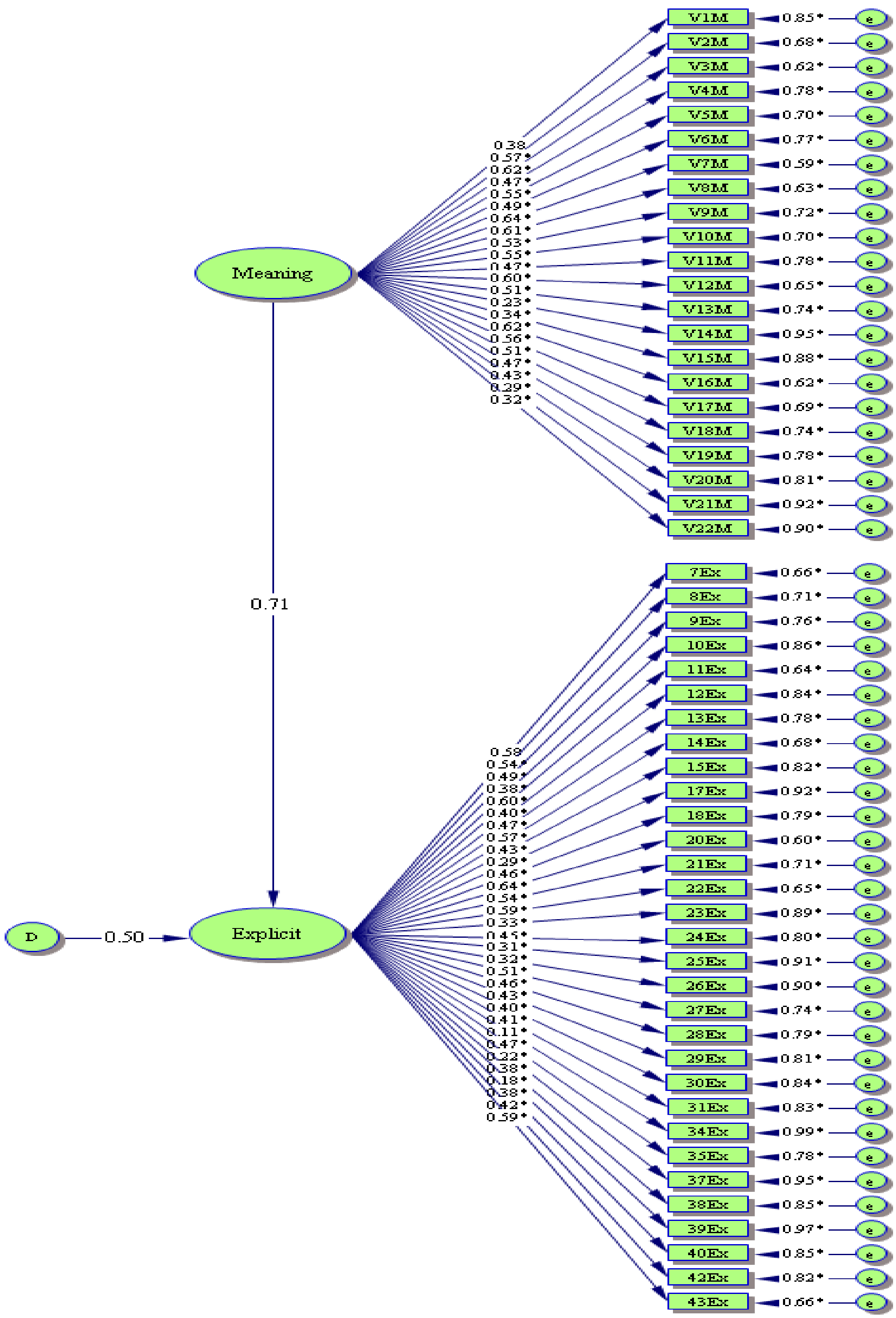
Figure 9. Model 6: Meaning as the Predictor and Implicit as the Dependent Variable

Table 19. Tests of Model Fit for the SEM Model: Model 6

Goodness of fit summary:	
Comparative fit index (CFI)	0.967
The Tucker-Lewis index (TLI)	0.981
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.026
Root Mean Square Error of Approximation (RMSEA)	0.018
Chi-square test of model fit:	
Value	7870.094
Degrees of Freedom	526
P-Value	0.0000

Finally, the regression model using meaning as the predictor variable and explicit as the dependent variable was evaluated (see Figure 10). Table 20 shows the model-data fit statistics for Model 7. This model produced a Chi-square of 17573.148 with 1169 degrees of freedom, significant at the 0.001 level, indicating a model-data misfit. Nevertheless, this model resulted in a CFI of 0.953, a SRMR of 0.027, and a RMSEA of 0.018. These values are favorable, indicating that this model represents the data well. As graphically represented in Figure 10, the explicit factor regressed positively on the meaning variable, with a standardized regression coefficient of 0.71, suggesting that the examinees' scores on explicit items would increase for their correct answers on the meaning items.

In sum, the SEM results provided some evidence for the crucial role of knowledge of meaning as a predictor of L2 listening ability. Specifically, there was a positive close relationship between knowledge of meaning and the ability to understand implicitly and explicitly stated information. Examinees who performed well on the meaning items tended to score higher on the listening items as well. Similar to knowledge of form, knowledge of meaning also seemed to contribute more to the implicit items than to the explicit ones.



* = Freely estimated

Figure 10. Model 7: SEM with *Meaning* as the Predictor and *Explicit* as the Dependent Variable

Table 20. Tests of Model Fit for the SEM Model: Model 7

Goodness of fit summary:	
Comparative fit index (CFI)	0.953
The Tucker-Lewis index (TLI)	0.976
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.027
Root Mean Square Error of Approximation (RMSEA)	0.018
Chi-square test of model fit:	
Value	17573.148
Degrees of Freedom	1169
P-Value	0.0000

Discriminant Analysis

Given that the ECCE aims to differentiate masters (i.e., test takers who pass the exam) from nonmasters (i.e., test takers who fail the exam), discriminant analysis was performed to investigate to what extent each predictor variable (form and meaning) correctly classify the examinees into groups of masters and nonmasters in listening comprehension. First, the groups of masters and nonmasters were identified using the cut score on the listening section (29 listening items correct[3]). Next, the composite scores on the GV items were computed, serving as the independent variables. Discriminant analysis was then conducted with a group variable (masters, nonmasters) as a dependent variable and two predictors (form and meaning).

Before implementing discriminant analysis, the F values of the two predictor variables were examined. Both form and meaning produced the F values at the level of 0.001 (form: $F = 10298.379$ (1, 42505, $p < 0.001$); meaning: $F = 12843.259$ (1, 42505, $p < 0.001$), indicating that the masters and the nonmasters differed significantly for each variable.

Next, the assumption of multivariate normality was measured, using the Box's test of equality of covariance matrices. The result showed that the value of Box's M was 2351.541 with an approximate F value of 783.805 at the 0.001 significant level, which violates the assumption of the equal covariance matrices for the two groups when performing discriminant analysis. This violation was not surprising due to the relatively large sample size. Since discriminant analysis is a very robust statistical technique, it was decided to continue the analysis (Sharma, 1996).

Table 21 displays the discriminant analysis results. The large discriminant function coefficient ($\lambda = 0.34$), analogous to the eta value in the analysis of variance, suggests that form and meaning function well in discriminating the masters from the nonmasters. Also, the value of Wilks' Lambda (0.75) and the high Chi-square value (12497.85, $p < 0.001$) indicates that the two variables discriminate between the test takers well. Based on the standardized canonical discriminant function coefficients, the equation for the discriminant function is as follows: D = 0.43 (Form) + 0.68 (Meaning), suggesting that meaning contributes to the discriminant function more than does form. Also, meaning yields a larger structure function coefficient (0.94) than form (0.84), which indicates that meaning has a stronger impact on the discriminant function than form.

[3] Editor's note: This is an approximation of the cut score, supplied to the author by ELI-UM. Actual pass decisions are based on ability levels estimated using item response theory methods.

Table 21. Discriminant Analysis Using Form and Meaning as Predictor Variables

	Standardized Canonical Discriminant Function Coefficient	Structure Function	Canonical Discriminant Function Coefficient
Form	0.43	0.84	0.11
Meaning	0.68	0.94	0.18
(Constant)			-4.32
Discriminant function coefficient: λ = 0.34 Wilks' Λ = 0.75 Chi-square = 12497.85*			

Note. * $p < 0.001$

Table 22 summarizes the number and percentage of the test takers classified correctly and incorrectly. Using form and meaning as predictor variables resulted in 71.5% of original group cases correctly classified. More specifically, there were 8138 (32.4%) nonmasters who were incorrectly predicted as masters, and 3973 (22.9%) masters who were incorrectly predicted as nonmasters. Even though the percentage of correct classification (nonmasters: 67.6%; masters: 77.1%) was not relatively high, form and meaning appeared to be critical variables in predicting the test taker's membership in the groups of masters and nonmasters in listening comprehension.

Table 22. Classification Results Using Form and Meaning as Predictor Variables

Group	Original Group Membership	Predicted Group Membership	
		Nonmasters	Masters
Nonmasters	25,131	16,993 (67.6%)	8,138 (32.4%)
Masters	17,376	3,973 (22.9%)	13,403 (77.1%)
71.5% of original grouped cases correctly classified.			

Discussion

This study investigated three research questions pertaining to the factorial structures of the GV section and the listening section of the ECCE and the relationship between lexico-grammatical knowledge and L2 listening ability. Before performing the EFA and the SEM, the test reliability was examined to obtain information on the extent to which the items performed as a homogenous groups. The GV section yielded a high internal consistency reliability coefficient, indicating that the GV items worked consistently in measuring lexico-grammatical knowledge. Similarly, the listening section also produced a high alpha, suggesting the listening items consistently measured the same construct.

The EFA was first conducted to probe the underlying traits of the GV items and the listening items. Consistent with Saito's (2003) findings, the EFAs on the GV section produced a two-factor promax solution that seemed to maximize parsimony and interpretability, indicating that the GV section was measuring two traits which were hypothesized as grammatical form and semantic meaning. With respect to the listening section, the analysis also produced a two-factor solution, suggesting that the listening items were measuring two underlying factors: listening for implicit information and listening for explicit information. The results were in accordance with Wagner's (2004) findings on the extended listening sections of the MELAB.

Following the EFA, item level SEM based on the original item coding was separately performed on the GV items and the listening items, in order to obtain more confirmatory information about the underlying traits of these item sets. The SEM results of the GV items provided substantial evidence for the two-factor solution of the GV section. More specifically, the GV items were measuring two traits: grammatical form and literal meaning, providing empirical evidence for Purpura's (2004) theoretical model of grammatical knowledge. Another item level SEM was then implemented on the listening items. The results also provided sufficient evidence in support of the two-factor solution of the listening section. That is, the listening items were measuring two underlying factors: listening for implicit information and listening for explicit information. Similar to Nissan, DeVincenzi, and Tang's (1996) findings, this study also found evidence for the distinction between the ability to comprehend explicit and implicit information and affirmed Buck's (2001) and Wagner's (2002, 2004) models of L2 listening ability.

Next, a series of SEMs was performed to investigate the relationship between lexico-grammatical knowledge and L2 listening ability. First, SEM was conducted, using composite scores on the form items and the meaning items serving as the factor indicators, in order to examine how L2 listening ability regressed on the predictors. The results showed that lexico-grammatical knowledge was a moderately strong predictor of L2 listening ability. In other words, the test taker's predicted listening scores would increase for every correct response on the GV items. This finding basically confirmed Rost's (1990) and Vandergrift's (2004) arguments for the significance of linguistic knowledge in L2 listening comprehension.

The SEM regression models at the item level provide more detailed information about to what degree knowledge of form and meaning contribute to L2 listening ability across different question types (implicit vs. explicit). The results show that both form and meaning serve as strong positive predictors of L2 listening ability, the ability to listen for implicit information, in particular. In accordance with Mecartty's (2000) study, this study also found evidence to support a positive close relationship between lexico-grammatical knowledge and L2 listening test performance. Also, the SEM results show that the meaning predictor seems to produce larger loadings on the implicit and explicit factors than does the form predictor. This seems to confirm VanPatten's (1996) input processing theory, where L2 learners tend to process aural input for meaning before they process it for form. However, no decisive conclusion can be drawn from this study, as the SEM models on these two predictors were not performed simultaneously in the current study. Future research is needed to investigate the relative contribution of grammatical form and lexical meaning to listening comprehension.

Finally, a discriminant analysis was performed to examine to what extent each predictor variable (form and meaning) correctly classify the test takers into groups of masters and nonmasters in L2 listening comprehension. The results show that using form and meaning as predictors produces a moderately high percentage of the test takers classified correctly. Both form and meaning function well in discriminating the masters from the nonmasters, but meaning contributes more to the discriminant function than form. The discriminant analysis proves valuable in providing information about the relative significance of each predictor.

Conclusion

The current study investigated the factorial structures of the ECCE GV section and the listening section and looked into the relationship between lexico-grammatical knowledge and L2 listening ability. The two underlying traits of the GV section and of the listening section were determined and lexico-grammatical knowledge was identified as an essential predictor of L2 listening ability. This study not only provides empirical evidence in support of Purpura's (2004) framework of grammatical knowledge and Buck's (2001) and Wagner's (2004) models of L2 listening ability, but also provides insights for how lexico-grammatical knowledge is related to L2 listening test performance.

Nevertheless, a few caveats for this study should be noted and may suggest some directions for future research. First, the results may not generalize to low- or high-ability L2 learners in that the ECCE used in this study targeted high-intermediate EFL learners. L2 listeners' proficiency level has been recognized as a potential variable by a number of researchers (e.g., Chiang & Dunkel, 1992; Conrad, 1985; Glisan, 1985; Rubin, 1994; Thompson, 1995). The relationship between lexico-grammatical knowledge and L2 listening ability may vary across various proficiency levels. More research working with different ability level learners will shed some light on this issue. In addition, this study did not consider the effects of task types on the role of lexico-grammatical knowledge in listening comprehension. In other words, the relationship between lexico-grammatical knowledge and L2 listening ability may vary across different task types (e.g., short conversation, radio interview). Very few empirical studies, unfortunately, have been undertaken to tap into this question.

These issues warrant further research in an attempt to allow test developers to design more reliable and valid assessments and to advance our understanding of the relationship between lexico-grammatical knowledge and L2 listening ability. Unquestionably, the continued focus upon research on L2 grammar and listening assessments is a challenging and pressing endeavor that will continue to exert considerable influence on second language testing and pedagogy.

Acknowledgements

I would like to express appreciation to the Spaan Fellowship Committee for accepting my research proposal. I am also very grateful to anonymous reviewers at ELI-UM for their valuable suggestions and insightful comments on this study. Special thanks also goes to Professor James E. Purpura and my colleagues at Teachers College, for their full support and good cheer.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Blau, E. (1990). The effect of syntax, speed and pauses on listening comprehension. *TESOL Quarterly*, *24*(4), 746–753.

Bolinger, D. (1977). *Meaning and form*. New York: Longman.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
Boyle, J. P. (1984). Factors affecting listening comprehension. *ELT Journal*, *38*(1), 34–38.
Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171–191.
Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.
Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book*. Boston: Heinle & Heinle.
Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*(4), 767–770.
Chang, J. (2004). *Examining models of second language knowledge with specific reference to relative clauses: A model-comparison approach*. Unpublished doctoral dissertation, Teachers College, Columbia University.
Chaudron, C. (1983). Simplification of input: Topic reinstatements and their effects on L2 learners' recognition and recall. *TESOL Quarterly*, *17*(3), 437–458.
Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, *26*(2), 345–374.
Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition*, *7*(1), 59–72.
Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly*, *25*(3), 431–457.
Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: Tentative model for test specification and development. *Modern Language Journal*, *77*(2), 180–191.
ECCE information bulletin. (2006). Retrieved October 3, 2006, from University of Michigan, English Language Institute Web site, http://141.211.177.75/UofM/Content/eli/document/ECCE_IB1.pdf
Fischer, R., & Farris, M. (1995). Instructional basis of Libra. *IALL Journal of Language Learning Technology*, *28*(1), 29–90.
George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference 11.0 update* (4th ed.). Boston: Allyn and Bacon.
Glisan, E. W. (1985). The effect of word order on listening comprehension and pattern retention: An experiment in Spanish as a foreign language. *Language Learning*, *35*(3), 443–472.
Goh, C. (2000). A cognitive perspective on language learner's listening comprehension problems. *System*, *28*(1), 55–75.
Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241–268). Cambridge, UK: Cambridge University Press.
Hasan, A. S. (2000). Learners' perceptions of listening comprehension problems. *Language, Culture, and Curriculum*, *13*(2), 137–153.
Higgins, J.M.D. (1995). *Facilitating listening in second language classrooms through the manipulation of temporal variables*. Unpublished doctoral dissertation, University of Kent at Canterbury.
Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes*, *1*(1), 91–107.

Joiner, E. (1997). Teaching listening: How technology can help. In M. D. Bush & R. M. Terry (Eds.), *Teaching enhanced language learning* (pp. 77–120). Lincolnwood, IL: National Textbook Company.
Kelch, K. (1985). Modified input as an aid to comprehension. *Studies in Second Language Acquisition*, *7*(1), 81–89.
Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics in Language Teaching, 29*(2), 135–149.
Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
Lado, R. (1961). *Language testing*. New York: McGraw-Hill.
Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murica (Ed.), *Teaching English as a second language or foreign language* (pp. 279–296). Boston: Heinle and Heinle.
Leech, G. (1983). *Principles of pragmatics*. London: Longman.
Liao, Y. (October, 2006). *Investigating the construct validity of the elementary-level GEPT grammar-vocabulary section*. Paper presented at the East Coast Organization of Language Testers Conference, Washington, DC.
Long, M. H. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377–393). Rowley, MA: Newbury House.
Lund, R. (1990). A taxonomy for teaching second language listening. *Foreign language annals, 23*, 105–115.
Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning, 11*(2), 323–348.
Mendelsohn, D. J. (1998). Teaching listening. *Annual Review of Applied Linguistics, 18*, 81–101.
Morley, J. (1990). Trends and developments in listening comprehension: Theory and practice. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics* (pp. 317–337). Washington, D. C.: Georgetown University Press.
Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
Neisser, U. (1976). *Cognition and reality. Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.
Nissan, S., DeVincenzi, F., & K. L. Tang. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report 51.]
Peterson, P. W. (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). Boston, MA: Heinle & Heinle.
Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly, 21*(4), 737–758.
Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge, UK: Cambridge University Press.
Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (October, 2006). *Reconsidering the notion of grammatical and pragmatic knowledge*. Paper presented at the East Coast Organization of Language Testers Conference, Washington, DC.

Rea-Dickins, P. (1987). *The relationship between grammatical abilities and aspects of communicative competence: With specific reference to the testing of grammar*. Unpublished doctoral dissertation: University of Lancaster.

Rea-Dickins, P. (1991). What makes a grammar test communicative? In J. Alderson, & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 112–131). London: MacMillan.

Rea-Dickins, P. (1997). The testing of grammar in a foreign language. In C. Clapham, & D. Corson (Eds.), *Language testing and assessment* (pp. 87–97). Victoria, Netherlands: Kluwer Academic Publishers.

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

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

Rubin, J. (1990). Improving foreign language listening comprehension. In J. E. Alatis (Ed.), *Georgetown University round table on languages and linguistics 1990* (pp. 309–316). Washington, DC: Georgetown University Press.

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: English Language Institute, University of Michigan.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations in comprehension beyond the sentence*. Chichester, UK: Wiley.

Sharma, S. (1996). *Applied multivariate techniques*. New York: John Wiley.

Stevens, J. P. (2002). *Applied multivariate statistics for the social scienc*e (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, K. (2002). Investigation of English listening difficulties of Taiwan students. *Selected papers from the eleventh international symposium on English teaching/fourth Pan-Asian conference* (pp. 518–526). Taipei: Crane Publishing Co.

Teng, H. (2001). The influence of syntactic modification on EFL listening comprehension. *Proceedings of the eighteenth conference of English teaching and learning in the R. O. C.* (288–305). Taipei: Crane Publishing Co.

Teng, H. (2002). An investigation of EFL listening difficulties for Taiwanese college students. *Selected papers from the eleventh international symposium on English teaching/fourth Pan-Asian conference* (pp. 526–533). Taipei: Crane Publishing Co.

Thompson, I. (1995). Testing listening comprehension. *AATSEEL Newsletter*, *37*, 24–31.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University, 2/1*. Retrieved October 2, 2005, from http://www.tc.edu/tesolalwebjournal/wagner.pdf.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1–25. Ann Arbor, MI: English Language Institute, University of Michigan.

Wolvin, A. & Coakley, C. (1982). *Listening*. Dubuque, IA: Brown.

Yen, A. (1988). An investigation of listening difficulties by Taiwanese college students. *English Teaching & Learning*, *12*(4), 62–72.