THE CONSTRUCT VALIDATION OF ELI LISTENING PLACEMENT TESTS

JEAN YOUNG CHUN
University of Hawaiʻi at Mānoa

ABSTRACT

This study presents the validation process for the listening placement tests administered by the English Language Institute (ELI) at the University of Hawaiʻi at Mānoa. The research questions are: (a) How does the ELI define the listening comprehension construct validity? (b) How well does the ELI Listening Placement Test (ELI LPT) measure the listening comprehension construct? And, (c) how differently do test-takers perform on the dictation test and the multiple-choice test according to language group? Participants in the research included international and immigrant students enrolled over three semesters in spring and fall 2010 and spring 2011. The study was conducted using a quantitative approach including test score analysis, test item analysis, and a survey, as well as a qualitative approach including curriculum analysis and interviews with administrators and instructors. The findings from the evaluation process addresses the three research questions, the ELI listening comprehension construct, positive and negative evidence concerning construct validity, and the different test performances of language groups. Some constructive suggestions based on these results are suggested for the ELI as well as follow-up research topics.

INTRODUCTION

The University of Hawaiʻi at Mānoa (UHM) has a high percentage of international and immigrant students. The English Language Institute (ELI) provides Academic English courses for these students to prepare them for their regular courses and avoid any excessive academic burden that might be caused by English language deficiencies. In order for the ELI to offer appropriate support for students, the ELI evaluates their language proficiency and places them in appropriate levels of English courses. Inaccurate evaluation of students' language abilities can cause numerous problems such as academic struggle, unnecessary expenditure for extra courses, and conflicts among stakeholders, including professors, instructors, administrators as well as

Second Language Studies, 30(1), Fall 2011, pp. 1-47.

students. Thus, it is important to ensure that the ELI Placement Test functions well as a measure of students' language proficiency in order to prevent such negative impacts. Although we should consider a number of issues such as the reliability of the test, distribution of the test scores, and test practicality when evaluating the test, the construct validity of the test is the most fundamental concern. If the test is not an appropriate measure for assessing student proficiency, the test scores will not represent students' actual levels and will not serve as a sound basis for the ELI's decisions on student placements.

Up to now, the ELI administrators have not noticed any conspicuous problems due to the placement test. However, the ELI has not conducted any test validation research for some time even though the current ELI placement tests have been used for a long period of time. The lack of evaluation on the validity of the ELI Placement Test motivates the author to carry out this study. From among the five subtests of the ELI Placement Test, the Academic Listening Test (ALT), Dictation Test (DCT), Reading Comprehension Test (RCT), Gap Filling (GF), and Writing test, this study will investigate the two listening placement tests, that is, the Academic Listening Test (ALT) and Dictation Test (DCT). Thus, the following literature review will discuss the construct of listening comprehension in academic circumstances like the ELI, how we can operationalize it, and which methods can be utilized to evaluate whether such operationalizations accurately represent students' academic listening skills.

Construct and Construct Validity

A *construct* is not easy to measure since it is related to something that occurs in the human mind. According to Ebel and Frisbie (1991, p. 108), "The term *construct* refers to a theoretical conceptualization about an aspect of human behavior that cannot be measured or observed directly." That is, the term *construct* describes concepts, such as love, motivation, attitude, and reading comprehension, which underlie certain human behaviors, yet are hard to define. Measurement involves collecting evidence that tells us something about the construct, but there is always the question of whether particular evidence is actually relevant to a given construct. According to Ebel and Frisbie (1991, p. 108), "Construct validation is a process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure." In other words, the goal of construct validation is to ascertain with solid evidence that a test score truly represents a test-taker's capability in whatever the test developer intends to measure. Based on this definition, we will discuss below the definition of listening comprehension, how to measure this construct, and how to evaluate the measurements.

Buck (2001) suggests two steps to defining a listening comprehension construct: (a) the understanding of the construct on a theoretical or conceptual level and (b) the accumulation of information of the target language use situation. That is, he argues that the theoretical understandings of the construct should be re-interpreted in light of the specific language use situation. These two steps are discussed in the following two sections.

Theoretical understandings of the listening comprehension construct. The current literature on listening comprehension has not reached consensus on a single definition of listening comprehension. According to Carroll (1971), as cited in Dunkel, Henning, and Chaudron (1993), most of the research on listening comprehension until the mid 1990s was approached from the perspectives of classroom objectives, without attention to the genuine nature of listening and its relationships with general language behavior. In addition to discussing the deficiencies of research on listening and insufficient understandings of listening comprehension, Witkin (1990) points out that another challenge to listening-related research is confusion and disagreements about the definition of *listening*. By discussing thirty-four definitions of listening comprehension extracted from the existing communication research, Glen (1989) claims that the failure to reach one universal definition of listening comprehension leads to a limit on the research into the nature of listening comprehension and listening-related teaching methods.

However, in spite of the differing points of view on the listening comprehension construct, scholars agree on some of the basic characteristics of listening when defining the construct. Listening comprehension involves processing information from auditory and visual inputs and relating it to previous schemata (Clark & Clark, 1977). Vandergrift (1999, p. 168) included a social context in listening comprehension. He described listening as:

> A complex, active process in which the listener must discriminate between sounds, understand vocabulary and grammatical structure, interpret stress and intonation, retain what was gathered in all of the above and interpret it within the immediate as well as the larger sociocultural context of the utterance.

Some researchers attempt to classify the cognitive process in a hierarchy from lower to higher order (Buck, 1991; Rost, 1990). They categorize these cognitive skills into two classes: a "lower" order of understanding, that is, the literal meanings; and a "higher" order understanding like inference and evaluation. Along with such efforts to align listening skills in a hierarchy, listening-skill-related taxonomies have been introduced and used as guidelines for listening test development (Munby, 1978; Powers, 1986; Richards, 1983; Rost, 1990).

Van Dijk and Kintsch (1983) approach the information process of listening comprehension from the perspective of listeners' strategies by introducing two main listening strategies: local and global. These are divided based on the different locations of the clues for listening comprehension. Local strategies search for the clues by connecting the facts provided by texts. That is, meaning construction is limited to the clause or sentence levels. In contrast, global strategies, also known as macro strategies, go beyond the local clues by relating the facts from texts to previous knowledge, such as synthesizing information, drawing conclusion, and making inferences.

These two strategies are reported to require listeners to approach information in different ways according to their listening abilities (Hildyard & Olson, 1982). When transitioning from the global to the local, listeners can verify the hypotheses they have already made from facts. The other way that listening comprehension proceeds, from local strategies to global strategies, is by building up data to construct a conclusion. The former is called knowledge-based analysis, while the latter is termed data-based analysis. The use of the different strategy processes can be determined by a listener's proficiency. It is argued that more skillful listeners are more likely to follow the knowledge-based analysis, while less skillful listeners adhere to the details for general understandings. This claim is buttressed by the study done by Shohamy and Inbar (1991) showing that lower level test-takers found it more difficult to answer questions that referred to global rather than local cues.

The complexity of the nature of listening comprehension itself and the various interpretations and approaches by a myriad of researchers do not allow for one ultimate answer for developing or evaluating listening tests. Hence, decisions concerning the scope of the listening comprehension construct are left to test developers based on the purposes of the test and their own specific circumstances. Dunkel, Henning, and Chaudron (1993) mentioned the importance of delimiting the listening construct when developing a test in their tentative model for test specification and development. According to them, listening comprehension follows the prerequisite steps of orientation, attention, perception, and recognition, while it precedes the subsequent steps application, analysis, synthesis, and evaluation. They pointed out that listening comprehension is separate from prerequisite and subsequent stages. Listening comprehension is cannot occur without the former steps, and the subsequent steps play a significant role in facilitating the listeners to fully understand the contents.

Understanding target-language use situations. As mentioned above, it is necessary to apply these theoretical understandings of the listening comprehension construct to the target-language use situation (Buck, 2001). With academic listening tests, it is essential to carefully consider which listening skills and/or tasks for successful performances in academic situations are needed.

Some researchers have suggested a list of listening skills which are necessary for successful achievement in academic circumstances. For example, Powers' (1986) survey presented important listening skills needed for students' academic success in class from the perspectives of university lecturers. The findings from the survey included: identifying main ideas or topic of a lecture; the relationships among main ideas in a lecture; retaining and retrieving what they listened to by taking notes; and inferring relationships in information provided in a lecture. Richards (1983) also provided a taxonomy of micro-skills needed in academic listening, including identifying the purpose of a lecture, identifying the key lexical items on a topic, and being aware of various lecturing styles. Among various listening skills regarded as crucial by researchers, Rost (1994) emphasized the importance of inference by claiming that lecture input can be changed into memorable intake mostly based on inference.

As an important tool to assist listeners with listening comprehension in academic situations like a lecture, note taking has gained much attention by researchers. A survey done by Ferris and Tagg (1996) indicates the importance of note-taking skills in students' subject-matter courses. The survey was distributed to faculty members across majors in four tertiary education institutions in USA. The survey addressed the importance of each main task in the participants' classes. Surprisingly, note-taking skills ranked highest, above tasks such as oral presentations and small group discussion. On a four-point Likert scale [from Always (1) to Never (4)], the average of importance for the note-taking skill ranged from 1.00 to 2.02, while oral presentations and small group work ranged between 2.33 and 3.51 and between 2.13 and 3.38, respectively. However, the diversity of class sizes and course levels in that study should be considered when applying the findings from this survey in other contexts since the different environments and contexts of each program might affect the importance of various tasks in class.

As a follow-up study, Ferris (1998) later approached note-taking skills from students' perspectives by asking students in the same four tertiary institutions about the relative importance of each main task in their class. Interestingly, even though students still pointed out the importance of note-taking in the survey, the degree of importance for note-taking was lower than the instructors' responses had been. The students put formal speaking as the most important skill with the mean score of 3.35 out of 7 on a Likert scale [from 1 (most important) to 7 (least important)]. This was followed by general listening comprehension (3.41), pronunciation (3.44), and communications with peers (3.60). However, note-taking showed an average score of 3.93. The answers shown in the student survey present a different rank from those which the instructors posed for each task's importance. The rank for the relative importance of oral/aural tasks in class answered by the students in this survey paralleled the results of Kim (2006)'s study, reporting on participants who were graduate students in non-science and non-engineering fields, who chose presentation skills and general listening comprehension as the most important skills to possess, while pronunciation and note-taking were considered the least important.

Operationalizing the construct. How a listening comprehension construct, defined based on theoretical and situational understandings, can be measured by making decisions on texts and tasks (test formats) of listening tests, is a process known as *operationalization* (Buck, 2001). In this study, the construct validity of tests will be approached with a focus on format. A discussion of test formats, especially multiple-choice questions and dictation will follow.

Currently, a number of listening formats have been developed and used such as listening cloze, sentence repetition, dictation, note-taking, and interpretive formats (Brown, 2006). However, among the myriad listening tasks, it is crucial to carefully consider which task best operationalizes the listening-comprehension construct. Such decisions for an appropriate task or test format must be based on clear understanding of what each task is designed to measure.

The dictation and multiple-choice listening test formats used as the ELI LPTs (Listening Placement Test) include the following characteristics. First, since dictation is widely used as an integrative test, a number of studies on what dictation actually measures have discussed how dictation contributes to assessing the various language skills, rather than providing a narrow investigation on the listening skills themselves used in dictation.

In favor of dictation tests, Oller and Streiff (1975) claimed that dictation played an important role as an excellent language test to assess language learner's internalized grammar, which can be called an *expectancy grammar*. According to their claims, the dictation test provided comprehensive information on the learner’s language development across phonological, syntactic, semantic, and sociolinguistic knowledge. These arguments were supported by other researchers. According to Hio (1983), the errors found in dictation revealed what listeners comprehended from auditory inputs as well as their knowledge of phonology, morphology, syntax, and semantics. In addition, other studies have demonstrated high correlations of dictation scores with those of other language proficiency tests (Valette, 1967; Oller, 1971; Oller et al.,1974; Fouly & Cziko, 1985).

Unlike a number of reports that showed dictation as a great measure for an overall language skill, only a few studies have narrowed down the scope of their research into listening skills and strategies alone, examined by dictation. Moreover, the findings from these studies have presented conflicting results. According to Buck (2001), dictation can measure listening comprehension, although the listening comprehension measured by dictation remains in a lower-level of cognitive skills, with only a literal understanding. Cartledge (1968) also argued that general listening comprehension can be measured by dictation since listeners need to contextualize their aural inputs. On the other hand, the results of the studies done by Valette (1964) and Sugawara (1999) do not support Buck (2001) and Cartledge (1968)’s claims. Valette (1964) reported that learners who practiced dictation in class failed to improve their listening comprehension skills. She argued that sound discriminations trained through dictation are different from general listening comprehension. Sugawara (1999) also raised the possibility that dictation might interfere with the listening comprehension process. According to him, dictation does not promote listening comprehension because listeners might feel burdened by the excessive mental processes involved in dictation. Therefore, it remains unclear whether dictation can play the role of a legitimate listening comprehension test and which listening skills are actually involved in dictation.

Turning to multiple-choice (MC) tests, Yi’an (1998) claims that what an MC test measures depends on the questions and answer options used. That is, the kinds of questions we ask test-takers will determine which listening strategies they will use in a test. In his study, by using introspection methods, Yi’an points out the role of questions and options in MC. He argues that questions and options provide the purpose of listening for listeners and influence their listening processes. Another noticeable feature of MC is that it involves various test-taking strategies, especially guessing. Cheng (2004) found a high rate of testing strategies used in MC as indicated by the results of the posttest survey in his study. According to his survey, 97% of the test-takers responded that they prefer multiple-choice (MC) items to open-ended (OE) items because their test-taking strategies helped them to get better scores. Their preferences towards MC were also revealed in the results of the two tests, MC and OE. The mean score for MC items, 33.84, was significantly higher than that for OE items, 25.23. Thus, he concluded that correct answers did not represent true listening comprehension, since 97 %, again, answered that guessing worked as a very important strategy in a test.

In addition to making the right decisions on test methods for operationalizing the listening construct, a test developer should consider the intervening variables involved with test performances when interpreting the test scores (Buck, 2001). The intervening variables refer to construct-irrelevant factors which might cause variability between test-takers. The variances caused by the factors which are not directly related to the test construct might mislead the inference of the test scores and threaten the construct validity of the test. Thus, a number of researchers have paid attention to which factors affect test performances and discussed how to avoid them in a test (Briere, 1968, 1973; Briere & Brown, 1971; Farhady, 1979; Chen & Henning, 1985; Hansen-Strain, 1987; Zeidner, 1986, 1987; Kunnan, 1990; Rubin, 1994; Buck, 2001). A variety of construct-irrelevant variables have been found including test method, text, and test-takers' characteristics. Among these factors, interesting findings on different characteristics of test-takers such as native language, ethnicity, and educational background have been reported and raise some concerns on the issue of test bias. For example, Kunnan's (1990) study showed how items function differently according to test-takers' native languages and their educational backgrounds. In his study, Chinese and Japanese groups showed strong preferences towards grammar multiple-choice questions, while Spanish groups preferred vocabulary to the other four test items of grammar, listening comprehension, reading comprehension, and writing error detection. Based on these results, he suggested that the different characteristics of ethnic groups, like instructional backgrounds and native language, might provide some advantages or disadvantages for examinees. Another study conducted by Hansen-Strain (1987) strengthened these arguments of test-bias by comparing the cognitive style and the cloze-test scores between two groups, Asian students and students from South Pacific Islands. According to the results of a field dependence/independence measure, and the cloze-test, Asian students turned out to be more field independent than Island students, and outperformed Island students in the cloze test. Based on these findings, she claimed that Asian students might have a great advantage taking the cloze-test due to their cognitive style of field independence, as compared to Island Students.

Evaluating Construct Validity

A number of approaches to construct validation have been used. They include differential-group studies, intervention studies, structural equation modeling and statistical methods using internal correlations, MTMM (Multi-Trait Multi-Method) and factor analysis (Bachman & Palmer, 1979, 1982; Alderson, Clapham, & Wall, 1995; Kunnan 1995; Sasaki 1996). In addition, other empirical studies have suggested different ways of evaluating construct validity by analyzing reliability and item difficulties and by correlating a test with another official test focusing on criterion-related validity (Fouly & Cziko, 1985; Chapell, Jamieson, & Hegelheimer, 2003). Among the approaches, internal correlations, MTMM, and analysis of reliability and item difficulties used in this study will be discussed in detail.

Alderson et al. (1995) introduced ways of using internal correlations to evaluate construct validity. One of them is to check the relationship between subtests and the whole test. They mentioned Classic Test Theory in their book, claiming that the correlations between subtests and the whole test should be more than + 0.70 in order to show good construct validity. They noted that the correlation between the subtest and the overall score (minus the score of test in question) should be used in order to prevent inflation of the correlation. However, since the internal correlation merely provides us with a general picture of how subtests are related to each other, we need to use more refined methods in order to determine what really leads to such correlations.

In this case, more advanced statistical analysis such as MTMM can be used. MTMM is based on the theory that tests measuring similar traits will demonstrate “higher intercorrelations (convergent validity) than those measuring different traits (divergent validity)” (Alderson et al., 1995, p. 186). In one example case using MTMM, Bachman and Palmer (1979) used MTMM to discover that scores from an interview were less influenced by test methods than scores from other tasks (translation and self-rating), and the interview satisfied both convergent validity and discriminant validity. However, Brown (2001) has pointed out that MTMM is rarely used as a construct validity measure for language tests since the mutual relatedness of language skills is in conflict with the precondition that the traits measured should be different from each other. Similarly, although some empirical research has attempted to use MTMM in language-related studies, the result had not been very successful. For example, Buck (1992) failed to prove with MTMM that listening and reading skills are distinct traits.

Chapelle, Jamieson, and Hegelheimer (2003) studied the construct validity of a web-based ESL test showed how reliability and item difficulties were used as part of the validation process. The web-based test was designed to place students into a proper ability level after taking the 10-15 minute Interest and Ability Finder. This test was the initial check for test-takers' interest and language proficiency. With the design of a computer adaptive test, the test-takers were supposed to spend most of their test-taking time at their level. Thus, the researchers needed to discuss whether test-takers can be distinguished reliably and items demonstrate proper level of difficulties. By examining reliability and item difficulties at each level, they found that test scores of advanced level examinees showed unsatisfactory levels of reliability like .70 for the reading test, .69 for the listening test, and .63 for the writing test. Thus, in order to work on reliability, they suggested improving their items and carrying out further qualitative research on test authenticity and strategies of test-takers.

Statement of Purpose

This paper will evaluate the construct validity of two subtests on the ELI Listening Placement Tests (LPTs): the multiple-choice ALT (Academic Listening Test) and the Dictation Test (DCT). This study is motivated by a variety of issues. As mentioned in the introduction, placement tests have important consequences. Moreover, there is a lack of construct validity evaluation for the ELI LPTs, and ELI administrators have noticed apparent differences in test performance among ethnic groups [Harsch, personal communication, October, 5, 2010]. Thus, this study will first examine what listening comprehension means for the ELI and how the dictation and multiple-choice tests function as measures of this ELI listening comprehension construct. Finally, this study will investigate how three different language groups, classified and based on their official languages, performed in the ALT and DCT. To those ends, the following research questions were posed:

1. How does the ELI define the construct validity of listening comprehension?
2. How well does the ELI LPT measure what the ELI defines as academic listening?
3. How differently do test-takers perform in the dictation test and the multiple-choice test according to language group?

METHOD

Participants

The participants in this study were international students and immigrant students who required further assistance to develop their English academic skills. They participated in ELI placement tests during one of the three semesters between spring 2010 and spring 2011. At UHM, international and immigrant students typically submit an official language proficiency test score like TOEFL (*Test of English as a Foreign Language*) for admission. If they achieve a score of 600 or higher, they are exempted from ELI courses. If they score below 600, or if they do not have any official language proficiency test score, they are required to take ELI placement tests. The scores of a majority of the students usually range between 500 and 600 based on the paper-based TOEFL. With the results of the ELI Placement Test, the test-takers can be divided into three levels: Exempted, 80 level, and 70 level. Those who receive a *T*-score of 60 or above on either the ALT or the DCT are exempted from ELI listening and speaking courses as are students with TOEFL scores of 600 or higher. Next, those who receive *T*-scores between 50 and 60 on either the ALT or DCT are placed into the ELI 80 speaking and listening course 80, while the rest of students who score below 50 are assigned to the ELI 70 speaking and listening course.

Although the total number of the test-takers during the three semesters was 374, the scores of 70 ELI test-takers were deleted due to missing or incomplete test scores. A total of 304 test-takers had complete scores for the four subtests of the ELI placement tests: Academic Listening Test, Dictation Test, Gap-fillings, and Reading Comprehension. The decisions made for these test-takers included 111 students who were exempted from the ELI, 94 students in ELI 80 courses, and 99 students in ELI 70 courses.

According to their official language information, this population can be classified into three broad categories as well: first, those from countries that use English as an Official Language (EOL); second, those from countries whose official languages belong to the Indo-European language family (IE); and lastly, those with non-Indo-European language (NIE) backgrounds. This grouping is based on how close their official languages were to English which is a part of the Indo-European language family. Out of 304 test-takers, only 277 students' backgrounds were accessible due to missing or incomplete demographic data. The first group, EOL, includes the following countries: Zimbabwe, Botswana, Hong Kong, Philippine, India, Palau, American

Samoa, Micronesia, Malaysia and Papua New Guinea. The second group consists of Switzerland, Norway, Germany, Poland, Denmark, Sweden, Russia, Angola, Macau, Bangladesh, Nepal, Chile, Brazil, Latvia, Iran, and Timor-Leste. The third group includes Burma, Morocco, China, Taiwan, Korea, Thailand, Japan, Mongolia, Vietnam, Cambodia, and Indonesia. Table 1 presents percentages of each language background that placed in each level of the speaking and listening courses. EOL has 34 students who were exempted (69%), 12 in ELI 80 (24%), and four in ELI 70 (8%) out of a total of 50. Next, in the IE group, 23 were exempted (53.49%), 13 placed in ELI 80 (30.28%), and seven in ELI 70 (16.28%) for a total of 43. Finally, in the NIE group, 32 were exempted (17.39%) with 66 (35.87%) placed in ELI 80 and 86 (46.74%) in ELI 70 for a total of 184.

Table 1

Percentages of Each Language Group in 70, 80, and Exempted Levels

	EOL	IE	NIE
Exempted	34(69%)	23(53.49%)	32(17.39%)
ELI 80	12(24%)	13(30.23%)	66(35.87%)
ELI 70	4(8%)	7(16.28%)	86(46.74%)
Total	50	43	184

Materials

Several sets of materials were used for this study: a survey questionnaire (see Appendix 1); the ELI placement test forms; listening test audio CDs; the students' demographic information such as language backgrounds and TOEFL scores; a survey of Teachers' Confidence Level (2009); and test scores for the four placement subtests: (a) Dictation Test (DCT), (b) Academic Listening Test (ALT), (c) Gap fillings (GF), and (d) Reading Comprehension Test (RCT). The author also examined the ELI speaking and listening course curriculum. More details regarding the survey questionnaire, Teachers' Confidence Level, and ELI listening tests are provided below.

The survey questionnaire consists of two parts. Part 1 collects demographic information, and Part 2 includes 14 items addressing the factors affecting test performances. The purpose of the survey is to analyze which factors influenced the students' performances on the listening tests and to investigate whether these factors are related to the listening construct or not. The factors were classified based on the two categories, construct-relevant variables and construct-irrelevant variables. As discussed in the literature review above, test-takers' performances are influenced by various factors including construct-relevant and construct-irrelevant variables. The construct-relevant factors will be the examinee's listening skills in the case of listening tests. In contrast, the construct-irrelevant variables include test method effects, test environment, different characteristics of test-takers, etc. The survey items includes items covering text familiarity (lecture listening), reading effect, test-strategy, writing effect, topic effect, understandings of the contents, word recognitions, test formats, memory effect, vocabulary, anxiety, speech delivery rate, and test instructions. These factors were classified based on the listening comprehension construct as it is defined by the ELI curriculum. The ELI curriculum puts an emphasis on note-taking skills and vocabulary for comprehensive understanding of academic sources such as lectures; thus, the construct-relevant factors include recognizing words, understanding the content, lecture listening in class, note-taking, and new vocabulary. Construct-irrelevant factors are reading effect, test strategies, anxiety, lecture listening in a test, memory effects, speech delivery rate, writing effect, topic, and test instruction.

Teacher Confidence Level is a survey for ELI instructors asking how confidently they feel that each student in their ELI 70 or 80 courses was placed in the right class. This survey was conducted by the ELI at the end of the Fall semester in 2009. Teachers' judgments on student placements were quantified as percentages.

The ELI listening placement tests includes dictation and multiple-choice sections. For the dictation (DCT), students listen to a 50-word speech at normal speed at first and listen to the same text divided into seven chunks, with pauses and at a slow speaking rate, and finally, they listen to the text at normal speed. The multiple-choice test (ALT) includes five listening testlets (three short lectures and two long lectures) and 35 items. Test takers listen only once and are allowed to take notes while they listen to the lectures. The lectures cover various academic topics and include features of a natural lecture in class like false starts, pauses, and fillers.

Procedures

This study has five main concerns:

1. The academic listening comprehension construct of interest to the ELI was examined by conducting document analysis. After that, two ELI listening and speaking course

instructors and the author analyzed ALT test items and compared them with ELI listening strategies and skills described in the ELI curriculum in order to see how well test items reflected what students learn in an ELI course.

2. The reliability of ALT and DCT and item discriminations of ALT were examined.
3. The internal consistencies of the ALT and DCT were investigated and the test scores of each of three groups (EOL, IE, NIE) were compared.
4. The Teacher Confidence Level survey and interviews with administrators and listening and speaking course instructors were utilized for evaluating the external validity of the ALT and DCT.
5. A survey was distributed to test-takers who took ELI placement tests in spring 2011 right after the ELI placement test was completed on each test day. The author visited the test room each time (The placement test was administered on three different days), distributed the paper survey, and collected it. It took about ten minutes to finish the survey. Moreover, I observed the whole test procedure from the beginning to the end.

Analysis

First, in terms of ELI curriculum analysis, I examined the ELI speaking and listening course curriculum including course goals, expected student outcomes, objectives, as well as listening strategies and tasks. After analyzing the ELI curriculum, the results were compared with the ALT test item analysis. The test item analysis was carried out based on the two categories of global and local questions. An ELI instructor and the author categorized each of the 35 items on the ALT into global and local questions, as is suggested by a study of Hansen and Jensen (1994), and the items that the two evaluators disagreed on were evaluated by another ELI instructor. The number of test items in each category was counted and converted into percentages.

The reliability of the ALT was examined with Cronbach Alpha, while Kuder-Richardson formula 21 (KR-21) was used for DCT reliability. Regarding the DCT test, each student's total score was the only accessible data unlike the ALT. Cronbach Alpha reliability was not an option for calculating the DCT reliability (Cronbach Alpha needs to split items into two sets, odd-numbered items and even-numbered items). However, the DCT did have information on test-takers' total scores and so KR-21 was used for obtaining DCT reliability.

Item discriminations for the 35 ALT test items were analyzed by comparing the performances of a high group and a low group. The grouping for item discriminations was based on the test-takers ALT total scores. The ALT total score refers to the number of right answers to each item. After calculating the total scores of 315 test-takers, these scores were aligned from the highest to the lowest and classified into three groups, high level (105), intermediate (105), and low (105). Then, item facilities of three levels on each item were calculated. To obtain item discriminations for each item, the item facility of the low group was subtracted from that of the high group.

The internal correlations among all four ELIPT subtests, listening multiple-choice (ALT), dictation (DCT), reading multiple-choice (RCT), and gap fillings (GF), were investigated. Correlation coefficients between each subtest and the whole test minus the subtest in question were calculated. After that, the DCT and ALT were correlated with each other, as well as the different subtests for reading and writing. In particular, if the correlation coefficients for the dictation and the listening multiple-choice tests, respectively, and the total score minus the subtest in question, show more than +.70, it indicates that DCT and ALT may be interpreted as demonstrating good construct validity according to classical theory (Alderson et al., 1995).

In order to compare how each language group, EOL, IE, and NIE, performs on the ALT and DCT, a two-way ANOVA (with the two variables being language group and language proficiency) was conducted. For this analysis, the language proficiency classification of test-takers did not follow the ELI student placements into Exempt, ELI 80, and ELI 70 since the author found that a certain portion of students were reassigned from their original placement based on their TOEFL scores. As for the NIE group, 19 students out of 184, or 10.33%, were reassigned into a new level, while 2% of the EOL group, and 2.33% of the IE group were moved up or exempted from the ELI. Thus, in this study, the grouping of the test-takers into language proficiency levels was based on the sum of their ALT and DCT scores. The test-takers were aligned from the highest to the lowest and were divided into three groups of high level (*n* = 93), intermediate (*n* = 92) and low (*n* = 92).

The survey of ELI teachers' confidence levels was used as one of the methods to judge external validity. All the percentages of confidence levels provided by ELI teachers towards each student's placement into ELI 70 and 80 listening and speaking courses were added up and divided by the number of students to calculate the average of teacher confidence levels. In addition to the average of the confidence levels, the comments of administrators and instructors in interviews regarding ELI LPTs and students placements were considered.

Finally, the responses to the survey of factors affecting test performances were analyzed based on the two classifications of construct-relevant and construct-irrelevant variables. The mean scores and standard deviations were compared for the three language groups, EOL, IE and NIE. The 59 respondents out of 68 examinees who took the ELI listening placement tests in spring 2011 returned the survey sheets right after they finished their placement test. Moreover, the scores for items 4, 6, 12, 13, and 14 that imply negative meanings were reversed-scored in order to compare the averages of each group's answers. For example, a score of 1 was converted into 5, while a score 4 was changed into 2, etc.

RESULTS

ELI Listening and Speaking Course Curriculum Analysis

The analysis of goals, student outcomes, and teaching philosophy described in the ELI speaking and listening course curriculum and interviews with administrators and instructors helped define ELI listening comprehension. The ELI listening comprehension construct can be defined as *the ability to use proper listening strategies for facilitating their understandings and appropriately responding to their listening as an autonomous listener in academic situation* (ELI Curriculum Philosophy & Mission Statement).

In order to define the listening comprehension construct, one of the approaches introduced by Buck (2001) was applied. He discussed three different approaches to defining the construct: competence-based listening construct, task-based listening construct, and a construct based on interaction between competence and task. Among them, the third approach was adopted to define ELI listening comprehension. This approach emphasizes the interaction between task and competence used in a target situation. The interaction refers to how competence is involved in task completion. For example, in an academic situation like college, major tasks will include lecture listening, discussion, and presentation. According to the third approach, the listening strategies or skills that are needed to complete these main tasks should be analyzed. Thus, this approach investigates both tasks and competencies that are used in a target situation, and does not merely focus on either competence or task.

Based on the third approach, the listening parts of the ELI listening and speaking course curriculum were analyzed to determine the major necessary listening tasks, skills, and strategies.

The findings from this analysis demonstrate that ELI curriculum includes three major tasks, *lecture, presentation,* and *discussion,* and three main listening skills, *listening comprehension, critical listening,* and *interactive listening*. The selection of these strategies and tasks is based on the findings of Ferris' studies (1996, 1998) studies on oral/aural communication skills in academic situations.

The ELI curriculum also describes in detail which listening skills and strategies are needed to complete each main task. How the listening tasks and listening skills are involved with each other will be illustrated below.

First, regarding the *lecture-listening* task, listening comprehension and critical listening skills are considered important for understanding academic listening materials successfully. In detail, the more specific listening strategies for completion of this task were provided with the three steps of pre-, during- and post-listening. Each step includes activations of background information, note-taking skills, and reviewing the notes, respectively. In addition, the ELI curriculum emphasizes the importance of critical listening as well, by isolating the critical listening skill from the general listening comprehension and putting it in a separate category. Critical listening is defined as "evaluating the contents that they comprehended, using what they just heard in order to construct their own opinions, incorporating their opinions from various resources, and responding to the listening materials in a critical manner" (Goals and Objectives in ELI speaking and listening curriculum, p. 5). Thus, it can be inferred that the critical listening skill exceeds the level of general listening comprehension.

Following lecture listening, the second task, *presentation*, appears to involve the three listening skills, listening comprehension, critical listening, and interactive listening. Listeners are required to develop critical listening skills as an audience, particularly when participating in presentations. This means that they need to listen critically to the presentations, not just simply comprehending them, to respond effectively to presenters by asking questions, and, eventually, evaluate the presentations. Although it holds true that these skills, actively responding to others and asking questions, belong to speaking skills, they are still associated with listening activities as well. In addition, they are distinguished from the previous two listening skills, listening comprehension and critical listening skill, in that the other two skills process aural inputs information in one-way from speakers to listeners. However, when it comes to interactive listening skills such as communication in real life, the listener is expected to immediately respond to the speaker, which means that listening is a two-way interaction, not a one-way transaction. Thus, the author classified responding to listening materials, which can be termed *Interactive Listening* (Field, 2008), as a third category of the ELI listening comprehension construct.

Similarly, the task of *discussion* turned out to also involve three listening skills: listening comprehension, critical listening, and interactive listening. The ELI indicates that students will comprehend and critically respond to other participants' opinions in discussion by effectively asking questions in small groups as well as in whole-class discussions. Based on the analysis of these listening tasks and listening skills defined in the ELI curriculum, it appears that the ELI listening comprehension construct goes beyond general listening comprehension to include subsequent stages such as application, analysis, synthesis, and evaluation, as described by Dunkel, Henning, and Chaudron (1993).

ELI LPT Test Item Analysis

As for the ALT test item analysis, each question was classified into one of two categories: global and local questions. Global questions include synthesizing information and drawing conclusions, while Local questions refer to locating details or understanding individual words. One of the ELI listening and speaking course instructors and the author classified each of the 35 ALT items as either a global or local question. The item classification process was conducted separately by each rater so they would not influence each other's judgments. After each rater finished categorizing the items, the results of the classification were compared. Out of 35 items, it was found that the raters' judgments on six items did not match. Consequently, a third ELI listening and speaking course instructor participated in item classification and her judgments on the six items were used to resolve the differences. According to a test item analysis based on these categorizations, the ALT test items turned out to include seven global questions (20%) and 28 local questions (80%) out of 35.

Descriptive Statistics, Reliability and Item Discriminations

According to Table 2, the scores on the ALT and DCT were near-normal in distribution, which means that examinees are well-placed along the continua. First, in terms of central tendency, the mean, median and mode of ALT were reported to be very similar to one another with scores of 20.70, 21, and 20, respectively. This indicates that ALT presented almost perfect normal distribution. Likewise, the mean and median for DCT were also almost exactly the same at 30.14 and 30, respectively. However, the mode for DCT was 24, which was lower than the other two estimates of central tendency. This may mean that the DCT scores were somewhat positively skewed. Next, the dispersion of ALT and DCT scores was examined in terms of their standard deviations and ranges. As shown in Table 2, the DCT standard deviation (11.30) and range (45) were much larger than the ALT standard deviation (5.28) and range (27), which means that the DCT scores were more widely spread out than the ALT scores.

Table 2

Descriptive Statistics of ALT and DCT

	ALT	DCT
No. of items	35	50
M	20.7 (59.14%)	30.14 (60.28)
Median	21 (60%)	30 (60%)
Mode	20 (57.14%)	24 (48%)
SD	5.28	11.3
Range	27	45
Skewness	.03	.07

With regard to the reliability coefficients for the ALT and DCT, the two tests showed very different results. For the ALT, the Cronbach alpha reliability coefficient was .74, while the K-R 21 was much higher at .92 for the DCT.[1] In addition, the item discrimination (ID) indices on the ALT were interesting: 10 items out of 35 (28.57%) turned out to have low item discrimination between high- and low-scoring examinee groups. These ten items' IDs ranged between .10 and .30, and among them, four items were extremely low at below .20. According to Table 3, the questions with IDs ranging from .20 to .29 were: 10 (.26), 11 (.29), 12 (.24), 13 (.23), 31 (.25), and 34 (.29), while the questions with item discriminations below .20 were: 6 (.11), 14 (.15), 33 (.17), and 35 (.16).

[1] Editor's note: Application of K-R21 or any other internal consistency estimate to dictation scores is controversial given the lack of independence between items.

Table 3

ALT Item Facilities and Item Discriminations

Item	IF (Total)	IF (High)	IF (Intermediate)	IF (Low)	ID
1	.42	.60	.35	.30	.30
2	.75	.90	.77	.59	.30
3	.41	.65	.34	.23	.42
4	.35	.65	.31	.10	.55
5	.54	.70	.52	.40	.30
6	.73	.76	.78	.65	.11
7	.73	.94	.76	.49	.46
8	.46	.72	.48	.19	.53
9	.77	.92	.76	.62	.30
10	.71	.82	.74	.56	.26
11	.81	.95	.80	.67	.29
12	.43	.55	.42	.31	.24
13	.81	.91	.82	.69	.23
14	.60	.72	.51	.57	.15
15	.37	.58	.31	.22	.36
16	.65	.83	.67	.47	.36
17	.60	.75	.64	.41	.34
18	.80	.95	.80	.66	.30
19	.34	.59	.30	.12	.47
20	.57	.71	.57	.41	.30
21	.79	.94	.90	.53	.41
22	.41	.60	.40	.22	.38
23	.34	.53	.32	.15	.38
24	.59	.74	.64	.39	.35
25	.63	.76	.66	.47	.30
26	.62	.81	.64	.41	.40
27	.37	.62	.31	.19	.43
28	.58	.86	.59	.30	.56
29	.77	.93	.85	.54	.39
30	.65	.81	.68	.47	.34
31	.54	.67	.53	.42	.25
32	.52	.72	.48	.36	.36
33	.50	.61	.47	.44	.17
34	.81	.95	.82	.67	.29
35	.75	.85	.70	.69	.16

Internal Consistency

The ELI ALT and DCT listening placement tests had moderately good correlations with the whole test score minus each test, respectively. According to Table 4, the correlation coefficient between the ALT and the whole test minus the ALT is .74, while the correlation coefficient between the DCT and the whole test minus the DCT is .60. In addition to correlations between each listening subtest and the whole test, Table 5 indicates that the correlation coefficient between the ALT and DCT at .68 surpass that of the ALT with the RCT at .58. This may indicate that the test-takers' performances on the ALT and DCT were less influenced by testing methods than their listening skills.

Table 4

Correlations Between the ALT and Whole Test Minus the ALT

	Whole test minus ALT	Whole test minus DCT
ALT	.744*	
DCT		.601*

**p*<.01

Table 5

Correlations for the ELI PT Subtests

	ALT	DCT	RCT	GF
ALT	1.00	.681*	.584*	.514*
DCT		1.00	.508*	.345*
RCT			1.00	.600*
GF				1.00

**p*<.01

External Validity

The investigations of external validity for the ALT and DCT rely on the Teacher Confidence Level (2009) ratings and the comments of administrators and listening and speaking course instructors in interviews. The average Teacher Confidence Level was high at 93.47%, which indicates that teachers in listening and speaking courses feel highly confident about the placement of students into their class. Because this study used the 2009 data for the teacher confidence level, which does not include the particular group of participants in this study, additional interviews with one of the ELI administrators and the three current ELI listening and speaking course instructors were conducted. The administrators and instructors were interviewed individually and notes on their answers were taken. The questions asked in interviews included whether they noticed any problems of their students' placements in class. According to the administrator, "I am pretty confident about students' placements. Only a few students were actually moved up after having an interview with me. Even though we officially conducted a Teacher Confidence Level only once in 2009, I guess that the teacher confidence level still would be above 90 percent." None of the instructors mentioned any noticeable problems with student placements in their classes, either. One of them pointed out gaps in students' speaking skills, but not in their listening skills, saying that students seemed to be placed evenly based on their listening skills while their speaking skills vary. The findings of the average teacher confidence level and the comments of the administrator and the instructors can lead to the following conclusions: (a) the ELI placement tests have satisfied their primary role of assigning the test-takers to the correct course levels and (b) the placement tests reflect what the instructors have taught in their classes. If the students were placed by the test with a different construct for listening comprehension from what the ELI teaches and intends to measure, the instructors might have noticed gaps between students' readiness for their class and their instruction based on the ELI curriculum. These findings can be used as evidence to support the construct validity of the tests.

Different Performances of Three Language Groups on the ALT and DCT

Descriptive statistics and a two-way ANOVA were used to investigate how the different language groups performed on the ALT and DCT tests. First, in terms of the ALT, descriptive statistics show apparent differences in test performances among the three groups, EOL, IE, and NIE. According to Table 6, the NIE group in each level from high to low outperformed the other language groups, with the highest mean scores of 60.28 in the high level, 50.31 in the intermediate level, and 39.57 in low level. Compared to the NIE group, however, the EOL showed the lowest mean scores in high and low levels with 56.92 and 38.37, respectively. In addition, the standard deviations of the NIE in each level were narrower than the other language groups. In contrast, the EOL group shows the widest standard deviations among the three language groups. This means that the examinees in the NIE group show more similarity in their test performances on the ALT in each level when compared to the other groups, while the test scores of the EOL group were relatively spread out. Despite the noticeable contrast of test performances between the EOL and NIE groups, the mean scores and standard deviations of IE group did not show any particular similar or dissimilar patterns with the other groups. In terms of their high and low levels, the mean scores of IE group range between those of the NIE and EOL groups, and the mean score in intermediate level is the lowest, at 46.22, among the three groups. The standard deviations of the IE group in each level, 5.99, 5.23, and 5.28, respectively, fall between those of NIE and EOL as well.

Table 6

Descriptive Statistics for Each Level (Language Proficiency) and Language Group (ALT)

Level	Group	*N*	*M*	*SD*
1	1	35	56.92	6.29
	2	26	58.21	5.99
	3	32	60.28	4.93
	Total	93	58.44	5.89
2	1	11	49.50	5.63
	2	10	46.22	5.23
	3	71	50.31	5.14
	Total	92	49.77	5.30
3	1	4	38.37	7.78
	2	7	38.58	5.28
	3	81	39.57	4.76
	Total	92	39.44	4.89
Total	1	50	53.80	8.25
	2	43	52.23	9.65
	3	184	47.32	9.14
	Total	277	49.25	9.45

Note. Level 1= High, Level 2=Intermediate, Level 3= Low
Group 1=EOL, Group 2=IE, Group 3=NIE

In order to ensure that these differences between language groups were significant, a two-way ANOVA was conducted. The alpha level was adjusted to .025 from .05 since two ANOVAs were used, one for each type of ELI listening placement test (i.e., the ALT and DCT). Based on this α-value of .025, Table 7 shows that only proficiency level had significant differences while language group and the interaction of level and group did not turn out to be significant. With regard to this lack of significance, these two factors, language group and the interaction between level and language group, show relatively low effect sizes compared with proficiency level. The partial eta-squared values in Table 7 represent the effect size of the three factors: proficiency level, language group, and the interaction between level and group. Their values are .514, .025, and .013, respectively. These figures mean that the significance of differences between levels show the medium effect size as over 50% while other factors, language group and interaction of level and language group had very small effect sizes (2.5% and 1.3 %, respectively). Thus, it can be inferred that language groups and the interaction between language group and level did not significantly affect the test performances of the examinees in ALT.

Table 7

Two-way ANOVA of Language Proficiency Level and Language Group (ALT)

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Level	8012.700	2	4006.350	141.867	.000*	.514
Group	191.917	2	95.958	3.398	.035	.025
Level x Group	100.692	4	25.173	.891	.470	.013
Error	7568.374	268	28.240			

*p < 0.025

However, despite the lack of overall significance differences in test performances between language groups and their small effect size, some of the different test performances between EOL/IE groups and NIE group were found to be significant using post hoc tests and descriptive statistics (see Tables 6 & 9). Tukey's HSD test was used for the post hoc tests. According to these results, there were significant differences between EOL and NIE (Table 9). In addition, as mentioned before, the descriptive statistics demonstrated contrasts in test performances between EOL/IE groups and NIE group (Table 6).

Table 8

Tukey HSD Comparison for Level (ALT)

					97.5% Confidence Interval	
(I) Level	(J) Level	Mean Diff (I-J)	Std. Error	*p*	Lower Bound	Upper Bound
1	2	8.67*	.781	.00	6.62	10.72
	3	18.99*	.781	.00	16.94	21.04
2	1	-8.67*	.784	.00	-10.72	-6.62
	3	10.33*	.781	.00	8.27	12.38
3	1	-18.99*	.781	.00	-21.04	-16.94
	2	-10.33*	.784	.00	-12.38	-8.27

Note. Level 1= High, Level 2=Intermediate, Level 3= Low
*p < 0.025

Table 9

Tukey HSD Comparison for Group (ALT)

					97.5% Confidence Interval	
(I) Group	(J) Group	Mean Diff (I-J)	Std. Error	*p*	Lower Bound	Upper Bound
1	2	1.57	1.105	.330	-1.32	4.47
	3	6.49*	.848	.000	4.26	8.71
2	1	-1.57	1.105	.330	-4.47	1.32
	3	4.91*	.900	.000	2.55	7.27
3	1	-6.49*	.848	.000	-8.71	-4.26
	2	-4.91*	.900	.000	-7.27	-2.55

Note. Group 1=EOL, Group 2=IE, Group 3=NIE
* $p < 0.025$

Interestingly, the results of the DCT test score analysis provided somewhat different results from the findings of the ALT analysis. According to the descriptive statistics (Table 10), the NIE group performed the worst among the three groups in high level and intermediate level with the lowest mean scores of 55.90 and 45.45 respectively, although the mean score (33.65) in the low level falls between the EOL (31.94) and IE (38.51). These results contrast with those of the ALT, showing the highest mean scores in each level on the ALT. In addition, when comparing the mean scores of the NIE group between the ALT and DCT, the figures have declined in the three levels, high, intermediate, and low by 4.38, 4.86, and 5.79, respectively.

In contrast, the EOL group in the high level and the IE group in high and intermediate levels, the mean scores between the ALT and DCT have increased by 3.68, 4.70 and 3.56, which means that high-level examinees of EOL and IE, and intermediate-level examinees of IE group perform better at DCT than ALT. Notably, in terms of low-level test-takers in EOL group, their mean scores between ALT and DCT have decreased by 6.43, which follows the pattern of the NIE group. It can be assumed that the similarity between the different language groups, EOL and NIE, might be caused by rough grouping of examinees based on their official languages. It is highly possible that there exist a myriad of different characteristics within language groupd classified by the single factor.

Table 10

Characteristics of Each Level (Language proficiency) & Language Group (DCT)

Level	Group	*N*	*M*	*SD*
1	1	35	64.60	6.03
	2	26	62.91	4.51
	3	32	55.90	5.70
	Total	93	61.14	6.70
2	1	11	49.09	6.99
	2	10	49.78	8.36
	3	71	45.45	5.27
	Total	92	46.35	6.04
3	1	4	31.94	6.50
	2	7	38.57	7.66
	3	81	33.65	5.31
	Total	92	33.95	5.65
Total	1	50	58.58	11.91
	2	43	55.89	11.17
	3	184	42.07	9.88
	Total	277	47.20	12.72

Note. Level 1= High, Level 2=Intermediate, Level 3= Low
Group 1=EOL, Group 2=IE, Group 3=NIE

Another intriguing feature of DCT test performances is the significance of differences within the two factors, language proficiency level and language group, as well as the interaction of group and level. Contrary to the results for the ALT analysis, the two-way ANOVA results for the DCT test in Table 11 show that the *p*-values for language group (.00) and the interaction of group and level (.014) are below .025, which indicates that language group and the interaction between level and group also account for differences in test performances on the DCT, with considerable variability among test-takers. In particular, the interactions between language group and language proficiency level can be identified in Figure 1 as well. The lines for the EOL group and NIE group cross at a point between intermediate level and low level, slightly towards the lower level. On the other hand, the EOL group line is crossed at the point between high and intermediate levels, somewhat closer to the intermediate level by the IE group line. In addition to the crossing of the lines at the two points, the changes in the gap between the EOL line and NIE line, and between the EOL line and IE line at different levels are noticeable in illustrating the interaction of language group and language proficiency level.

Table 11

Two-way ANOVA of Language Proficiency Level and Language Group (DCT)

Source	Sum of Squares	*df*	Mean Square	*F*	*P*	Partial Eta Squared
Level	14438.403	2	7219.202	226.548	.000*	.628
Group	928.258	2	464.129	14.565	.000*	.098
Level x Group	403.881	4	100.970	3.169	.014*	.045
Error	8540.134	268	31.866			

$^*p < 0.025$

The difference in the mean scores of the NIE group and EOL group is the largest at the high level, and the differences get narrower as the proficiency levels drop into intermediate and low levels. At the low level, the NIE group performs slightly better than the EOL group. On the other hand, the mean score difference between the EOL and IE groups is much bigger in low level than in the other levels. The gaps between the two groups decline until the two lines reach a crossing point as language proficiency levels have increased from low to high. This presents the reverse result from the pattern between the EOL and NIE groups. Thus, the significant differences of the three factors identified in Table 11 and the different patterns of lines according to language group and levels shown in Figure 1 support the interpretation that the three factors, language group, language proficiency level, and an interaction between them are associated with the examinees' test performances in DCT.

However, despite such significant differences within level, group, and their interaction, it should be noted that the effect sizes of the three factors are different. That is, the effect sizes that explain how much each factor contributed to examinees' test performances in the DCT need to be compared. For effect size comparison, the partial eta-squared values for level (.628), group (.098), and the interaction of level and group (.045) were utilized. These figures indicate that proficiency level has a medium effect size of 62.8 %. In contrast, the other factors' effect sizes are only 9.8% and 4.5%. Compared to the ALT, it holds true that the effect sizes of language group and an interaction of language proficiency and language group, to some extent, are larger. In terms of the gaps in effect sizes for the three factors between ALT and DCT, language proficiency showed the largest difference (11%) followed by language group (7.3%) and the interaction of the two factors, level and group (3.2%). It can be assumed that the larger interaction between group and level indicate greater variability among test-takers in language group and language proficiency level. These larger effect sizes indicate that the mean differences for the DCT test performances are larger within language groups and language proficiency levels, respectively. However the influences of the two factors still remain small compared to language proficiency levels. Hence, the interpretations of these results are that the test performances on the DCT test were mostly affected by language proficiency level, not language group and an interaction of level and language group as shown in ALT. However, as mentioned in the ALT discussion, the descriptive statistics and post-hoc results still reveal apparent differences between the EOL/IE groups and the NIE group, which supports the argument that the test-takers perform differently according to language groups in DCT (Tables 10 & 13).

Table 12

Tukey HSD Comparison for Level (DCT)

					97.5% Confidence Interval	
(I) Level	(J) Level	Mean Diff (I-J)	Std. Error	*p*	Lower Bound	Upper Bound
1	2	14.78*	.830	.000	12.61	16.96
	3	27.19*	.830	.000	25.01	29.36
2	1	-14.78*	.830	.000	-16.96	-12.61
	3	12.40*	.832	.000	10.22	14.59
3	1	-27.19*	.830	.000	-29.36	-25.01
	2	-12.40*	.832	.000	-14.59	-10.22

Note. Level 1= High, Level 2=Intermediate, Level 3= Low
* *p* < 0.025

Table 13

Tukey HSD Comparison for Group (DCT)

					97.5% Confidence Interval	
(I) Group	(J) Group	Mean Diff (I-J)	Std. Error	*p*	Lower Bound	Upper Bound
1	2	2.69	1.174	.059	-.39	5.76
	3	16.51*	.900	.000	14.15	18.87
2	1	-2.69	1.174	.059	-5.76	.39
	3	13.82*	.956	.000	11.31	16.33
3	1	-16.51*	.900	.000	-18.87	-14.15
	2	-13.82*	.956	.000	-16.33	-11.31

Note. Group 1=EOL, Group 2=IE, Group 3=NIE
* *p* < 0.025

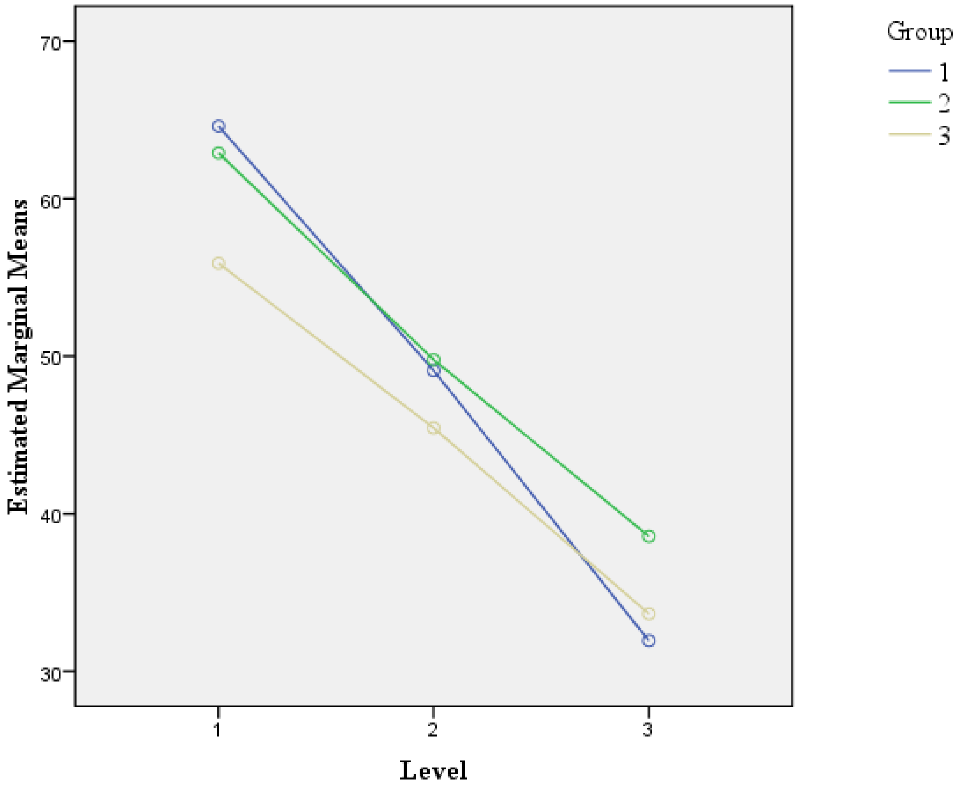

Figure 1. Significant differences of Language Proficiencies and Language Groups (DCT)

Note. Level 1= High, Level 2=Intermediate, Level 3= Low

Group 1=EOL, Group 2=IE, Group 3=NIE

In short, it can be inferred that differences in the examinees' test performances on the ALT and DCT are determined mostly by language proficiency levels rather than by language group. Even though the two-way ANOVA results in DCT show significant differences in the three factors, language group, language proficiency level and the interaction, the effect sizes for language group and the interaction were still low relative to that for language proficiency level. In spite of the small effect size for language group and the interaction of group and level, according to comparisons of mean scores between ALT and DCT, it is noticeable that test performances of high-level examinees in EOL and IE groups showed an apparent contrast with the test-takers in the NIE group in the same level. In addition, in terms of NIE group, the three levels from high to low also show better performances on the ALT than on the DCT. Thus, the different test performances of each language group for the ALT or DCT raised the possibility that the test format might affect the test-takers' performances even though the degree of impact remains small.

Survey of Test Performance-Affecting Factors

The survey conducted in this study investigated which factors affect the examinees' test performances on the ALT and DCT, and evaluated whether these factors are construct-relevant. The results of the survey have also provided insights into the different characteristics of each language group on the two tests.

Table 14

Comparisons of Mean Scores and Standard Deviations of the Whole Group and Each Language Group

	N	*M*	*SD*
Whole Group	59	3.39	1.08
EOL	34	3.90	1.08
IE	19	3.55	.97
NIE	6	3.20	1.08

Note. Group 1=EOL, Group 2=IE, Group 3=NIE
Likert Scales range from 1(Not at all) to 5(Very).

First, the general overview of answers to the survey turned out to be positive. According to Table 14, the mean score of the whole group was 3.39, above the midpoint of the five-point scale (from one to five) by .30. This means that the factors described in the survey did not significantly intrude upon the examinees' test performances on the ELI LPTs. However, Table 13 presents the slight differences in responses to the survey according to language groups. The mean score for the EOL group was the highest among the three groups, 3.90, followed by the IE and NIE groups with mean scores of 3.55 and 3.20, respectively. These results indicate that the EOL group may be the most resistant to factors affecting test performance, while the NIE is the most vulnerable to these factors among the three language groups. However, despite the relatively low mean score for the NIE group compared to the EOL and IE groups, the NIE mean score is still above three, which means that they were not profoundly affected by the factors. When it comes to standard deviations, the three groups show very similar results (Table 14). The standard deviations of the EOL and NIE were almost the same, 1.08. The standard deviations were 1.077

and 1.079, respectively. The IE group's standard deviation was .97 which is slightly lower than the other two groups. This means that the answers of IE group show more similarity than those of the other groups, EOL and NIE.

Table 15

Mean score comparisons between ALT and DCT on Test Performance Affecting Factors

Factors	ALT			DCT		
	EOL	IE	NIE	EOL	IE	NIE
Construct-relevant factors						
1. Comfortable with a lecture listening in class* (Text familiarity)	4.00 (EOL)/ 4.23(IE)/ 3.56(NIE)					
2. Recognizing the words	x	x	x	4.33	3.53	3.50
3. Understanding the content	x	x	x	4.33	4.26	3.32
4. Note-taking	4.17	3.56	3.18	x	x	X
5. New vocabulary	3.83	3.47	2.61	3.83	3.47	2.70
Construct-irrelevant factors						
6. Reading overriding effect on listening test	3.17	3.28	3.35	x	x	x
7. Test strategy (Reading questions and options)	4.67	4.05	3.85	x	x	x
8. Familiar with a test format	4.83	3.41	3.86	4.00	3.11	2.79
9. Anxiety	2.60	3.26	2.77	3.00	3.21	2.75
10. Comfortable with a lecture listening in a test (Text familiarity)	3.33	3.61	3.45	3.67	3.33	2.87
11. Memory effect	4.17	3.79	3.38	4.00	3.37	2.79
12. Fast speech delivery rate	3.33	2.63	2.66	3.00	2.42	2.48
13. Writing effect on listening test	x	x	x	3.17	3.11	2.53
14. Easiness of Topic (Background Knowledge)	4.17	4.06	3.63	4.17	3.84	2.94
15. Clear test instruction	5.00	4.26	4.28	5.00	4.42	4.24

Note. 1. The figures of answers to the five factors, including Reading overriding effect on listening test, Writing effect on listening test, New vocabulary, Fast speech delivery rate, and Anxiety were reversed.

2. Answers to the *comfortable with a lecture listening in class* factor are for both ALT and DCT.

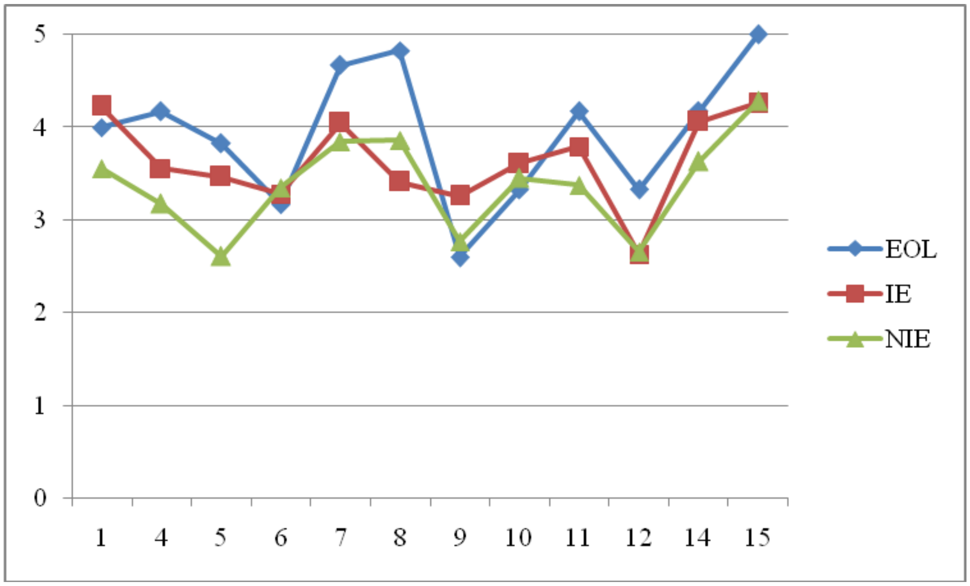

Figure 2. ALT Survey Responses of Three Language Groups, EOL, IE, and NIE

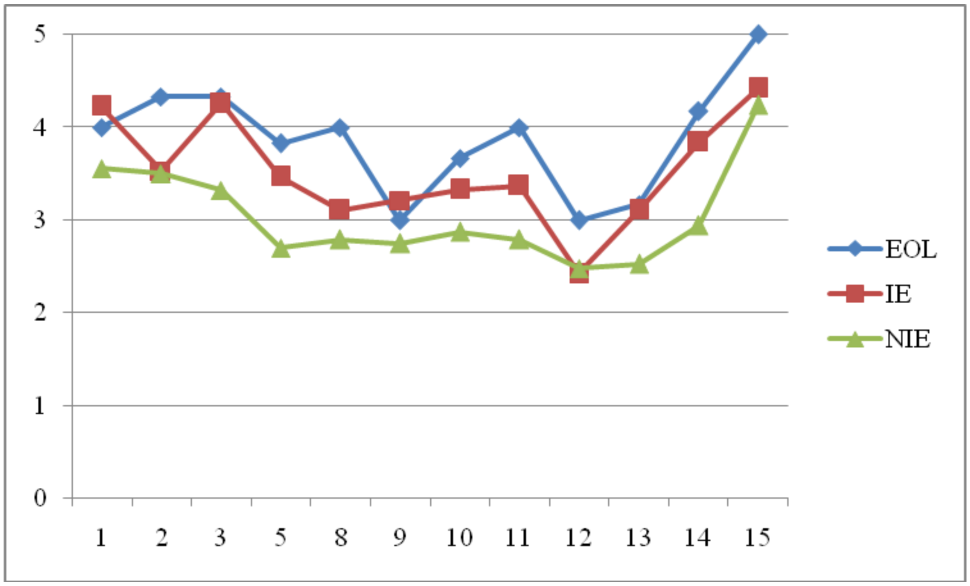

Figure 3. DCT Survey Responses of Three Language Groups, EOL, IE, and NIE

Table 15, Figure 2, and Figure 3 provide more detailed information about how differently each language group, EOL, IE, and NIE, responded to the two categories of factors, *construct-relevant* and *construct-irrelevant variables*. First, regarding the construct-relevant factors described in the survey, the three language groups did not show any particular difficulties with them, with the exception that *new vocabulary* appears to be an obstacle for the NIE group. The mean scores for their answers for the *new vocabulary* factor on the ALT and DCT, were 2.61 and 2.70, respectively—well below 3.00, which means that they found more new words than the other groups, EOL and IE, on the two tests. This result suggests that their deficiency in vocabulary might lower their listening comprehension ability on the listening tests. Next, in terms of *construct-irrelevant variables*, more factors from this category appear to affect the test-takers' performances. However, according to Table 15, Figure 2, and Figure 3, these factors influence the NIE group much more than either the EOL or IE group. The answers of the EOL and IE groups in this category presented only two factors, *anxiety* for EOL and *fast speech delivery rate* for IE. Regarding the EOL group, their responses show the highest anxiety among the three groups in the ALT with a mean score of 2.60 (The responses to anxiety were reverse scored). This figure, 2.60 is lower than the mean score for *anxiety* in the EOL group on the DCT by .30. Thus, the results reveal that the EOL group is more likely to feel anxious when taking the ALT than the other language groups. In addition, anxiety might affect the test performance of the EOL group on the ALT more than on the DCT. Next, the IE group did not demonstrate any particular difficulties caused by the factors in the survey except for *speech delivery rate*. Their answers to all the factors mentioned in the survey were 3.00 or above, while the respondents of the IE group answered that *the speech rate* of the ALT and DCT are fast with mean scores of 2.63 for the ALT and 2.42 for the DCT. These figures are the lowest among the three language groups despite a slight difference from those of the NIE groups.

However, contrary to the generally positive answers of the EOL and the IE groups, the answers of the NIE group presented a number of factors that might negatively affect their test performances, especially on the DCT. Table 15, Figure 2, and Figure 3 show that the three factors of *anxiety, speech rate*, and *vocabulary* are reported as influencing the test performances of the NIE group on both the ALT and DCT. The mean scores for *anxiety, speech rate*, and *vocabulary* are 2.77, 2.66, and 2.61, respectively, for ALT, and 2.75, 2.48, and 2.70, respectively, for the DCT. In addition to these common factors between the two tests, Table 15 and Figure 3 present additional difficulties that this group appears to be having when taking the DCT. They are: *test-format and text unfamiliarity; topic difficulty; writing effect;* and *memory effect*. Thus, it can be inferred that a higher number of factors in the ELI listening tests (with regard to the DCT) challenge the NIE group examinees more than the other language groups (EOL and IE).

These dissimilarities among the language groups may make more sense when analyzed in more depth with a focus on the relationships among the factors described in the survey. First, the two factors, *recognizing the words* and *understanding the content*, that belong to the category of *construct-relevant variables*, have to do with cognitive processes in listening. So perhaps these language groups are going through different cognitive processes when they are taking the DCT. According to Table 15 and Figure 3, the NIE group focuses on *word recognitions* (3.50) more than *content comprehension* (3.32), while the IE group relies much more on *understanding the content* (4.26) than on *recognizing the words* (3.53). The EOL group appears to use both cognitive processes actively with mean scores of 4.33. This result is in accordance with Buck's (2001) claim that higher level listeners tend to focus on understanding the listening inputs by storing the information in chunks, while the lower level listeners attempt to recognize the words and often fail to retain the information they receive. Thus, one possible interpretation of these dissimilarities is that the different language proficiencies of the three language groups might lead the examinees to utilize different approaches to taking the dictation test. According to Table 9, the NIE group had the lowest overall mean DCT score with 42.07, followed by the IE group (55.89), and the EOL group (58.58).

Next, the *reading overriding effect* and *the test strategy of reading questions and options as clues for finding the answers* presented contradictory results (Table 15 & Figure 2). The *reading overriding effect* in multiple-choice questions means that excessive reading in listening tests might intrude on the test-takers' listening performances. However, as one of the test strategies, *reading questions and options in listening test*, provides additional clues for listeners to compensate for the deficiency of their listening skills as well. Unfortunately, in the survey, the answers related to these *reading effects* in ALT turned out to be conflicting. According to Table 15 and Figure 2, in terms of the *reading overriding effect*, the EOL group showed a lower mean score (3.17) than the other two groups, the IE group (3.28), and the NIE group (3.35). This result indicates that NIE group examinees felt the least bothered by reading questions and options in the ALT among the three groups while the reading effect appears to have intruded the most on the EOL group. However, these results do not match with their answers to the next question about *using reading questions as one of their compensatory strategies*. Interestingly, the EOL group showed the highest mean score with 4.67 among the three language groups so reading questions and options appears to help them somewhat more in answering the questions. These mismatched responses between *reading overriding effect* and *test strategy* make it hard to predict any tendency in reading factors for the ALT.

Following the reading effect, the three factors, *test format familiarity, text familiarity based on different situations*, and *anxiety* were analyzed together to interpret different features of each language group (Table 15, Figure 2, & Figure 3). As mentioned above, the EOL group showed the highest anxiety towards the ALT among the three groups. However, it is notable that the EOL group answered that they are highly familiar with the ALT test format with a mean score of 4.83. This value was even higher than the mean score for *test format familiarity* towards the DCT at 4.00. Thus, one possible interpretation is that a high level of anxiety has a negative influence on the EOL group's test performances on the ALT regardless of familiarity with test format. This interpretation may be supported by their different answers for *text familiarity according to situations, in class and in tests*. The EOL group responded that they feel generally comfortable with lecture listening in class (4.00). The mean scores for *lecture-listening familiarity* were lower for the ALT (3.33) than for the DCT (3.67). This result increases the credibility of the interpretation that anxiety towards ALT lowered lecture-listening familiarity on the ALT more than on the DCT. In contrast, with regard to anxiety levels, the IE and NIE groups showed similar patterns to each other. Both of these groups appear to experience higher anxiety towards the DCT than the ALT. The mean scores for *anxiety* on the ALT reported by the IE and NIE groups are 3.26 and 2.77, respectively, and the corresponding values are only 3.21 and 2.75 for the DCT. Regarding their *test format* and *text familiarities*, the IE and NIE groups answered that they are less comfortable with the DCT than the ALT and their answers to *lecture listening familiarity* was lower for the DCT than for the ALT (see Table 15, Figure 2, & Figure 3).

As the interpretations of the three *test format familiarity, text familiarity*, and *anxiety* factors can be approached together, so can the answers for the *memory effect, note-taking*, and *fast speech rate* factors, especially for comparison across the language groups (Table 15, Figure 2 & Figure 3). When it comes to *the memory effect*, the NIE group reported that their test performances on the DCT are more vulnerable to memory effects than for the ALT in that their mean score of 3.38 was higher for the ALT than for the DCT (2.79). These results are interesting in that the lectures used in the ALT are much longer than the script of the DCT, and the examinees are allowed to listen to the ALT lectures only once while the short script of DCT is played three times for the listeners. This may mean that the NIE examinees have more difficulty remembering what they have listened to in the DCT, despite the short script length and repeated listening, than they have on the ALT. These confusing results may make sense when compared with the answers for two factors, *note-taking* and *speech delivery rate*. Table 15 shows that the mean score of the NIE group for *note-taking* was 3.18, above the midpoint 3.00, while that for *speech delivery rate* was 2.48, which may mean that the NIE group may tend to be more vulnerable to fast speech delivery than to note-taking skills. Hence, these results appear to predict that a fast speech rate might work as an obstacle to remembering what they have listened to, yet, their note-taking skills can help them retain information from lectures.

Remarkably, the IE group answers seem to indicate that they tend to feel more sensitive to speech delivery rate than the NIE group. According to Table 15, the mean scores of the IE group on *speech delivery rate* are 2.63 for the ALT and 2.42 for the DCT. These figures are slightly lower than the mean scores of the NIE group for the ALT (2.66) and the DCT (2.48). However, the answers of the IE group for *note-taking* (3.56) and *memory effect* (3.79 for ALT, 3.37 for DCT) were higher than those of the NIE group. The mean score of the NIE for *note-taking* was 3.18, while the mean scores of *memory effects* are 3.38 for the ALT and 2.79 for the DCT. Thus, it may be fair to say that, even though the IE group feels the speech delivery rates in both the ALT and the DCT are fast, they can generally process the information quickly enough to avoid the influence of memory effects. Unlike the other two groups, the EOL group did not demonstrate any particular problems with these three factors.

Finally, it was noted above that the rest of the factors described in the survey—*vocabulary*, *writing effect*, and *topic*—may have presented difficulties for the NIE groups when taking the ALT and DCT, unlike the EOL and IE groups. Each group reported that *test instructions* were very clear with means above 4.00 (Table 15, Figure 2, & Figure 3).

In brief, the findings from this survey designed to investigate potential factors affecting the listening test performances reveal that the examinees were not negatively influenced by either of the categories, *construct-relevant* and *construct-irrelevant variables*. The answers from the whole group appeared to be generally positive, and the mean scores of the three language groups were all above 3.00. However, apparent differences between EOL/IE groups and NIE group were identified. Among the three groups, the EOL group had the most positive answers while the NIE had the most difficulties among the three groups. This probably means that the EOL group was the most resistant to various factors related to the test performance followed by the IE group. The

NIE was discovered to have test performances that could be more easily influenced by the factors examined here as compared to the other groups, especially by the construct-irrelevant factors. This in-depth analysis of the relationships of these factors to proficiency also revealed a number of intriguing features of each group such as differences in cognitive processes for listening, the reading effect, anxiety, and the memory effect.

DISCUSSION

1. How Does the ELI Define the Construct Validity of Listening Comprehension?

The ELI listening comprehension construct was defined based on listening tasks and strategies described in ELI curriculum. The major tasks and strategies include lecture listening, presentation, discussion and listening comprehension, critical listening, and interactive listening. Based on these results, the listening skills which are required to complete the three main tasks turned out to exceed the scope of understanding literal meanings to include metacognitive skills. Moreover, the ELI curriculum distinguishes critical listening from general listening comprehension, which implies that the ELI emphasizes this critical listening skill. However, the relative importance of each skill, listening comprehension, critical listening, and interactive listening was not mentioned specifically in the curriculum (see the Goals and Objectives in ELI Speaking and Listening Curriculum). Thus, it can be said that the ELI attempts to define the listening comprehension construct through a hierarchy of cognitive skills from lower to higher order. However, the ELI does not specify the priority and degree of the importance of each listening skill. Whether or not the proportions of test items measuring each listening skill in ELI LPTs are appropriately distributed cannot be examined since the ELI does not specify the relative importance of each skill in their curriculum. In addition to these testing concerns, there is a possibility that the ELI instructors might teach their courses with different perceptions towards these skills unless they get clear guidelines. This possibility could affect the degree to which each course achieves the student outcomes for the three listening skills because instructors might emphasize each listening skill differently.

2. How Well Does the ELIPT Measure What the ELI Defines As Academic Listening?

The ambiguity of the relative importance of each listening skill prevents immediate answers to the second research question in this study. According to the results of the test item analysis, the ELI ALT turns out to include seven global questions, which are measuring high-order cognitive skills, out of 35 items. However, among these seven global questions, four questions (6, 12, 14, and 35) had low item discrimination (.11, .24, .15, and .16, respectively, as shown in Table 1). This means that only three global questions out of 35 items, which is 8.57%, function well enough to measure comprehension beyond literal meanings. This leads to the conclusion that the ALT mostly measures literal meanings. In addition, according to the current literature, the DCT, the other ELI Listening Placement Test, is believed to examine literal meanings as well, although there is little consensus about which specific listening skills a DCT measures. Thus, it can be argued that test performances on the ALT and DCT depend highly on test-takers' abilities to understanding literal meanings. This finding leaves the ELI with a decision to make as to the degree to which the higher-order cognitive skills of listening comprehension and critical listening should be examined through the ELI LPTs. The literature is also ambivalent about whether it is necessary to measure test-takers' actual listening skills, their understanding of literal meaning, or skills beyond literal meaning. Some researchers (De Jong & Glas, 1987) claim that literal meaning represents test-takers' actual listening skills, and that understanding beyond literal meanings involves their cognitive skills. Others researchers (Suen, 1994; Burger & Doherty, 1992; Thompson 1995) suggest that the excessive narrowing of listening comprehension to only literal meanings leads to threats to construct validity for listening tests. Accordingly, based on this literature review, the ELI should make a decision and specify the scope of listening comprehension in thinking about the validity of the ELI LPTs.

One of the pieces of evidence supporting the construct validity of the ELI LPTs, the Teacher Confidence Level survey conducted in 2009 indicates that the teacher confidence level about students' placements was above 90%, which means that instructors felt that what they taught in class was matched well with students' levels. The lack of teacher confidence level data for 2010 and 2011 students creates a gap in the present study. However, additional interviews with administrators and instructors in spring and fall 2010 and spring 2011 did indicate similar findings to those found in 2009. Administrators and instructors in ELI listening and speaking courses indicated high confidence in their current students' placements. These findings support the argument for moderately high construct validity for the ELI LPTs.

In addition, the results of internal correlations for the ELI LPTs added additional evidence to strengthen the validity argument for the ELI LPTs. As already mentioned in the results section, the ELI listening placement tests, the ALT and DCT scores correlated moderately with the whole test scores (minus each test). In addition to solid correlations between ALT/ DCT scores and the total scores, the finding that the correlation coefficients between ALT and DCT scores are higher than those between ALT and RCT scores may indicate that test-takers' performances on ALT and DCT were more highly related to listening skills than to test methods.

Despite evidence in favor of the construct validity of the ELI LPTs, some concerns about the ELI LPTs were raised as well. First, the ALT reliability, which is a precondition for validity, was only moderately high (at .75). There are several possible explanations for this mediocre reliability. According to Ebel and Frisbie (1986), low reliability coefficients might be caused by: (a) a short length of test, (b) a test comprised of heterogeneous items, (c) a test with less discriminating items or with too easy or too difficult items, (d) test-takers that formed a homogeneous group in terms of their language proficiency, and/or (e) a speeded test. Among these possibilities, item discrimination and facility values indicate that item performance may be one of the factors affecting the current reliability. As expected, the item analysis indicated that a number of items had discrimination and facility values that might lead to lower reliability. Almost 29% of the entire ALT items fell into the categories of marginal (six items) or poor items (four items) (Brown, 2005). This means that the ELI needs to evaluate these items and consider revising them in order to make the listening tests more reliable. However, it still holds true that other factors such as homogeneity in language proficiencies of the test-takers and diverse language backgrounds of test-takers might be also be factors affecting the reliability.

3. How Differently Do Test-Takers Perform on the Dictation Test and the Multiple-Choice Test According To Language Group?

The results of the descriptive statistics and two-way ANOVA addressed the third research question in this study. As mentioned in the Results section, it turns out that, overall, language group membership does not affect test performances significantly for the ALT and DCT. The test scores are more likely to be determined by different language proficiency levels. Even though significant differences between language groups on the DCT were found, the effect size was not large enough to have a profound influence on test performances. Although the overall effect sizes of language group on the ALT and DCT scores were much lower than those of language proficiency levels, it was discovered that the NIE group performed better on the ALT in the three levels, while only the high-level proficiency group in the EOL group and high and intermediate levels of the IE group performed better on the ALT that on the DCT. This leads to the conclusion that test format may have affected the examinees' test performances on the ELI LPTs even though its impact on their overall test performances might be relatively small. Thus, this study did not indicate that test-takers' official languages played a significant role as one of the factors affecting the examinees' test performances on the ALT and DCT. However, the findings for different test performances according to language group do appear to be related to the issue of test bias which was discussed in literature review.

The ELI should be aware that the ELI LPTs appear to give advantages or disadvantages to the examinees according to language group, particularly for high level learners, depending on the test format. For example, it can be predicted that high level Japanese students might get some advantages from taking the ALT.

The ELI also should take into consideration that they need to account for potential biases due to test format in making proper administrative decisions about students' placement. Indeed, the results of this study tend to legitimize the current ELI placement policy of taking the highest score from either the ALT or DCT for test-takers' placement.

The survey was administered to investigate what factors affect the test performances of different language groups and whether those factors threaten the construct validity of the ALT and DCT tests. The survey indicated that the test-takers were not influenced, to any great extent, those factors mentioned in the survey. These results tend to strengthen the arguments that the construct validity of the ELI listening placements is not threatened by construct-relevant and construct-irrelevant variables. Moreover, the survey findings provide the ELI with more in-depth understanding of the characteristics of the different language groups. In particular, the difficulties identified in this study that each language group might potentially have with listening comprehension and test-taking may prove useful for the ELI for revising instruction in ELI listening courses. For example, the EOL group showed higher anxiety toward the ALT than the IE and NIE groups. In addition, NIE students were reported as having a myriad of difficulties including vocabulary, topics, memory effect, and speech delivery rate. In particular, the NIE students answered that they had some trouble remembering what they had listened to even though they understood the content. This result agrees with Goh (2000), who found in her study that low-level test takers had the same problems as the NIE students in this study. In both cases these test-takers indicated that they understood the content, but easily forgot what they had listened to. Goh (2000) suggested that their problems might be related to working memory in the parsing stage which is usually were meanings are constructed. Their limited working memory, which can also be called a short-term memory, allows only for shallow processing of information. This raises the possibility that the difficulties reported by the NIE group might be caused by their relatively low language proficiency compared to the other groups, the EOL and IE, as shown in mean scores comparisons of the ALT and DCT tests (Table 6 & 10). If so the ELI might consider conducting further research on how such differences between language groups in listening comprehension occur and how each language groups' difficulties should be handled in teaching ELI classes, in test development, and in administrative decision-making based on test score interpretations.

CONCLUSION

In this paper, the construct validity of the ELI LPTs, specifically the ALT and DCT, was investigated with three foci: (a) how the ELI defines listening comprehension, (b) how well the ELI LPTs measure their construct, and (c) whether their construct operationalization might be limited by test-bias due to factors such as different language backgrounds. These research questions were answered to some extent and the findings provide the ELI with constructive suggestions for improving the ELI LPT itself, its implementations, and course instruction. However, the listening comprehension construct and the factors affecting test performances are very complex issues. Hence, discussions of the findings among ELI administrators and instructors, and further research on the issues faced in the process of investigating construct validity, should continue long after this study is finished. The following topics might prove useful for discussion and further research:

1. The ELI needs to prioritize the different listening skills described in their curriculum and decide on the relative importance of each skill. Agreement on these issues should lead to another decision about what proportions of the listening test items should be global and local. These decisions could in turn guide instructors in terms of what listening skills they should teach in their listening courses and in what proportions.

2. Follow-up research on some of the interesting findings from the survey such as the EOL group’s high anxiety towards the ALT and the memory effects on test performances on the DCT for the NIE group would also be useful and interesting. These further findings might then help the ELI apply them to their teaching and test development.

Most importantly, follow-up studies on the above topics would provide additional information to help solidify the arguments presented in the present study on construct validity of ELI LPTs.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language testing construction and evaluation*. New York: Cambridge University.

Bachman, L. F., & Palmer, A. S. (1979). A multitrait-mutlimethod investigation into the construct validity of six tests of speaking and reading. In A. Palmer, P. Groot, & G. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149-165). Washington, DC: Georgetown University.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Briere, E. J. (1973). Cross cultural bias in language testing. In J. W. Oller, Jr., & J. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher* (pp. 214-227). Rowley, MA: Newbury House.

Briere, E. J. (1968). Testing ESL among Navajo children. *Language Learning, 18*(3), 11-21.

Briere, E. J., & Brown, R. H. (1971). Norming tests of ESL among Amerindian children. *TESOL Quaterly, 5*(4), 327-334.

Brown, H. D. (2006). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Brown, J. D. (2001). *Using surveys in language programs*. New York: Cambridge University.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

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

Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning, 42*(3), 313-357.

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

Burger, S., & J. Doherty. (1992). Testing receptive skills within a comprehension-based approach. In R. J. Courchene, J. I. Glidden, J. St. John, & C. Therien (Eds.), *Comprehension-based second language teaching* (pp. 299-318). Ottawa: University of Ottawa.

Cartledge, H. A. (1968). A defense of dictation. *ELT Journal, 22*(3), 226-231.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test.

Language Testing, 20(4), 409-439.

Chen, Z., & G. Henning. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-63.

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544-553.

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

De Jong, J. H. A. L., & Glas, C. (1987). Validation of listening comprehension tests using item response theory. *Language Testing, 4*(2), 170-192.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, *77*(2), 180-191.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quaterly, 13*(3), 437-58.

Ferris, D., & Tagg, T. (1996). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly, 30*(1), 31-58.

Ferris, D. (1998). Students' views of academic aural/oral skills: A comprehensive needs analysis. *TESOL Quarterly, 32*(2), 289-318.

Field, J. (2008). *Listening in the language classroom.* Cambridge: Cambridge University.

Fouly, K., & Cziko, G. A. (1985). Determining the reliability, validity and scalability of the graduated dictation test. *Language Learning*, *35*(4), 555-566.

Glen, E. (1989). A content analysis of fifth definitions of listening. *Journal of the international listening association, 3*(1), 21-31.

Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55-75.

Hansen-Strain, L. (1987). Cognitive style and first language background in second language test performance. *Tesol Quarterly, 21*(3), 565-563

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241-268). Hong Kong: Cambridge

University.

Hildyard, A., & Olson, D. (1982). On the comprehension and memory of oral versus written discourse. In D. Tannen (Ed.), *Spoken and written language* (pp. 19-32). Norwood, NJ: Ablex.

Hio, Y. (1983). Dictation as a testing device-based on error analysis. *Language Laboratory, 20,* 2-16.

Kim, S. (2006). Academic oral communication needs of East Asian international graduate students in non-science and non-engineering fields. *English for Specific Purposes, 25*(4), 479-489.

Kunnan, A. J. (1990). Differential item functioning and native language and gender groups: The case of an ESL placement examination. *TESOL Quaterly, 24*(4), 741-746.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach*. Cambridge: Cambridge University.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University.

Oller, J. W., Jr. (1971). Dictation as a device for testing foreign language proficiency. *English Language Teaching*, *25*(3), 254-259.

Oller, J. W., Irvine, P., & Atai, P, (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24*(2), 245-252.

Oller, J. W., Jr., & V. Streiff. (1975). Dictation: A test of grammar-based expectancies. *ELT Journal, 30*(1), 25-36.

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

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

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

Rost, M. (1994). On-line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 93-128). Hong Kong: Cambridge University.

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

Sasaki, M. (1996). *Second Language proficiency, foreign language aptitude and intelligence: Quantitative and qualitative analyses.* New York: Peter Long.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing, 8*(1), 23-40.

Sugawara, Y. (1999). Dictation and listening comprehension: Does dictation promote listening comprehension? *Language Laboratory, 36*, 33-50.

Suen, D.L.K. (1994). Communicative considerations in a large-scale listening test. In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 32-55). Hong Kong: The Chinese University.

Thompson, I. (1996). Assessing foreign language skills: Data from Russian. *Modern Language Journal, 80*(1), 47-65.

Valette, R. M. (1964). The use of the dictee in the French language classroom. *Modern Language Journal*, *48*(1), 39-43.

Valette, R. M. (1967). *Modern language testing: A handbook*. New York: Harcourt Brace Jovanovich.

Vandergrift, L. (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal, 53*(3), 168-176.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Witkin, B. (1990). Listening theory and research: The state of the art. *Journal of the International Listening Association, 4*, 7-32.

Yi’an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice test. *Language Testing*, *15*(1), 21-44.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing, 3*(1), 80-98.

Zeidener, M. (1987). A comparison of language, sex and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*(1), 55-71.

APPENDIX:

SURVEY OF TEST PERFORMANCE-AFFECTING FACTORS

Part 1. Demographic Information

Male/ Female
Graduate/ Undergraduate students
Major: ______________________________
Age: __________
Native language: ______________

Part 2. Please read the questions and circle the number. (1: Not at all, 5: Very)

	Part A Dictation	**Part B** Multiple- choice
1. Generally, I am comfortable with listening to a lecture in class.	1 2 3 4 5	
2. I am comfortable with lecture-type listening on tests.	1 2 3 4 5	1 2 3 4 5
3. I could easily take notes while listening.		1 2 3 4 5
4. I did not have enough time to read the questions and options and answer the questions.		1 2 3 4 5
5. Reading questions and options helped me to answer the questions.		1 2 3 4 5
6. I did not have enough time to write down what I heard	1 2 3 4 5	
7. I could understand the content.	1 2 3 4 5	
8. I focused on recognizing the words.	1 2 3 4 5	
9. I was familiar with this test format.	1 2 3 4 5	1 2 3 4 5
10. Generally, topics were easy to understand.	1 2 3 4 5	1 2 3 4 5
11. I could remember what I listened to when I was answering the questions or writing down words.	1 2 3 4 5	1 2 3 4 5
12. I found many unknown vocabulary words.	1 2 3 4 5	1 2 3 4 5
13. I was anxious when taking a test.	1 2 3 4 5	1 2 3 4 5
14. The rate of speech was fast.	1 2 3 4 5	1 2 3 4 5
15. Test instructions were clear enough.	1 2 3 4 5	1 2 3 4 5

Thank you!