A Speaking and Listening Achievement Test:

Assessing Advanced Learners in the Community English Program

Teachers College, Columbia University

A&HL 4088: Second Language Assessment

Dr. Kirby Grabowski

December 15, 2014

TABLE OF CONTENTS

I. INTRODUCTION

A. Motivation for the Test and Study

This test was developed for adult ESL learners studying at the Community English Program (CEP) at Teachers College, Columbia University. The CEP is a community language program offering communicative language classes to adult learners of diverse nationalities, proficiency levels, ages, and socio-economic backgrounds. It is also a language education lab where TESOL and Applied Linguistics faculty and students from Teachers College teach courses and conduct empirical research.

Overall, the CEP is divided into nineteen levels based on proficiency, with six beginner, intermediate, and advanced levels, as well as an Advanced Studies class. Each session is ten weeks, and classes meet three times per week for two hours at a time, for a total of 60 hours of instruction. The courses, which emphasize integrated listening, reading, writing, and speaking skills, are taught with the aid of a theme-based textbook. Grammar, pronunciation, and vocabulary are also emphasized throughout each unit.

Tests are routinely conducted in order for both teachers and students to assess whether students have met learning objectives. In order to maintain consistency across levels in the program, the CEP requires that teachers administer three unit tests and one final exam throughout the course of each teaching session. While the CEP establishes extremely broad course objectives, more specific, functional objectives are outlined at the beginning of each chapter. These functional objectives are tested through unit tests; the primary goal of these tests, then, is to see whether or not learners have achieved these objectives.

The achievement test was designed for students in the Advanced 4 (A4) English class at the CEP. As an achievement test, it aimed to assess student learning both summatively and

formatively. Summatively, this test provided the teachers and students with a general idea of what was learned and whether or not objectives were met. The assessment data was also used formatively by the teachers to plan further, tailored instruction to help learners notice and close learning gaps identified by the test. For students, this information is critical for self-awareness of learning progressions and to shape future learning practices.

The purpose of this paper is to evaluate and discuss this test, which was designed for Unit 2, "World of Dreams." Specifically, two sections of the test will be analyzed: the listening section, a discrete-point multiple-choice examination, and the speaking section, a discussion-based constructed-response task. First, the research questions will be posed. Then, methods used to create and administer the test will be discussed. After that, findings from the data will be presented and analyzed. Finally, recommendations for the creation and administration of similar achievement tests at the CEP will be offered.

B. Research Questions

The following research questions will be addressed in this paper:

1. What is the nature of listening ability and speaking ability in the unit test?

2. To what extent were the raters consistent when rating speaking ability in the unit test?

3. To what extent does listening ability relate to speaking in the unit test?

4. Is there a correlation between absences and tardies and student performance in the unit test?

II. METHOD

A. Research Design

Every CEP test is required to cover the four communicative skills (reading, writing, listening, and speaking). While this test included all of these components, the focus of this paper is on the development and administration of a multiple-choice, selected-response listening

section as well as a constructed-response speaking task. This study may be classified in terms of data collection method as nonexperimental, quantitative, and statistical (Grohtjahn, 1987). It was non-experimental as the participants consisted of a singular, intact group. The scoring and assessment of both sections was quantitative, and statistical analyses were used to interpret the results.

B. **Participants**

The participants comprised 22 adult, non-native speakers of English from a variety of L1 backgrounds: eight Japanese, four Korean, two Chinese, two Spanish, two German, and four French. The participants included six men and 16 women. While all were placed into the upper-advanced level at the CEP, proficiency levels differed on an individual basis. Time spent in the U.S. prior to enrollment ranged from three weeks to four years. Educational backgrounds varied, but all of the students had either graduated with or were in the process of completing an undergraduate degree. At least nine had completed advanced degrees in fields including medicine, law, literature (Korean and Austrian), engineering, and business.

A survey-based preliminary needs analysis conducted by the teachers at the start of the semester provided information about student motivation for taking the course; 18 of the 22 students indicated a desire to improve conversation skills, and 8 indicated a desire to learn English for the purposes of finding or improving their job prospects. Only one of the participants indicated a desire to improve English for testing purposes.

Based on the characteristics of this population, the findings of this study are best generalized to other adult, upper-advanced English classes comprised predominantly of highly-educated students who seek to improve their conversational English skills.

The raters used for the constructed-response task comprised two current graduate students at the Teachers College. One of the raters was both a teacher of the tested population and co-author of this paper. One is studying Teaching English as a Second Language (TESOL) and the other Applied Linguistics, and both have completed coursework relevant to second language learning. Additionally, the raters have had experience teaching and rating the placement tests in the CEP.

C. Instrument Design

1. Theoretical Model: Listening Ability

The textbook for the course, *In Charge 2* (Daise, 2003), defines "listening ability" in terms of a variety of different functional listening objectives in the "Scope and Sequence" section of the book. In Unit 2, the objective is "listening for personal interpretation."

In order to assess "listening ability", then, a theoretical model of this construct was needed. Initially, Buck's (2001) adaptation of Bachman and Palmer's (1996) theoretical framework for language ability appeared promising to use as a basis for the theoretical conceptualization of listening ability. The central constructs, *language competence* and *strategic competence*, provide a useful distinction between the knowledge one has about a language and the strategies (cognitive and metacognitive) that one would require to successfully manage, apply, and implement this knowledge. Both competences play an essential role in listening ability, and are highly interrelated. Strategic competence is necessary for a successful demonstration of language competence, and is, thus, indirectly assessed on any examination of proficiency.

However, while Bachman and Palmer's (1996) framework does point out that strategic competence and language competence are distinct factors that may influence decision-making on

a test, it does little to explain how these areas may be assessed and measured when used in combination. Thus, it was necessary to find an additional framework that would offer a clear bridge between strategic competence and language competence, one that would consider how the two might interact. Kim's (2009) framework was selected as it focuses on deciphering language meaning, a skill which requires both competences. However, as Kim's framework was designed for reading, *not* listening, it was necessary to adjust the specifics of the framework to better apply to our context.

As Kim (2009) notes, "reading should be seen as a cognitive activity, where the reader interacts with the text to derive meaning" (p. 3). Yet, this skill is not unique to reading; it is quite evident in listening as well, where listeners interact with spoken text to derive meaning. Bozorgian (2012) draws the connection between the two areas, arguing that "perceiving receptive input demands a pliable cognitive process to revise cognitive representations in that both listeners and readers construct while receiving input" (p. 3). Both reading listening and reading are receptive skills that share an end goal of deciphering meaning within discourse.

Listening ability, then, was defined in terms of Kim's (2009) constructs. Specifically, there are three core variables: *reading for endophoric-literal meaning*, *reading for endophoric-implied meaning*, and *reading for exophoric-implied meaning*. As Kim's (2009) framework is being adjusted to suit a listening context, these variables will be referred to as "listening for" rather than "reading for" various types of meaning. Listening for endophoric-literal meaning is based purely on the listener's ability to identify literal meaning from the passage based on information, which is clearly, or explicitly stated. The variables which comprise this domain, on our test, are "listening for main idea" and "listening for detail", when this information is very clearly incorporated into the listening passage. In contrast, listening for endophoric-implied

meaning requires the listener to infer the main idea or details based upon information which is implied within the passage, but not explicitly or directly stated. Finally, listening for exophoric implied meaning requires looking beyond the context of the passage when interpreting meaning. Kim includes five aspects of this domain: deriving contextual, psychological, or sociolinguistic, sociocultural, or rhetorical meaning from the hearer's background knowledge beyond the text. This test uses all but two of these variables - deriving sociolinguistic and sociocultural meaning.

Additionally, Kim (2009) notes that inferring meaning and deriving meaning are skills that tend to be more challenging than understanding literal meaning, provided that learners have a basic understanding of the grammatical and vocabulary structures used within a text. As a result, she posits that "incorporating various types of inference items can lead to tests that better differentiate among advanced readers" (p. 2). Drawing from this, questions eliciting understanding of implied meaning on our test were predicted to discriminate between those who are better able to decipher a more general notion of meaning from the listening passage and those who are not.

Figure 1 shows the construct of listening ability as it has been interpreted for this unit test, including *listening for endophoric-literal meaning*, *listening for endophoric-implied meaning*, and *listening for exophoric-implied meaning.*

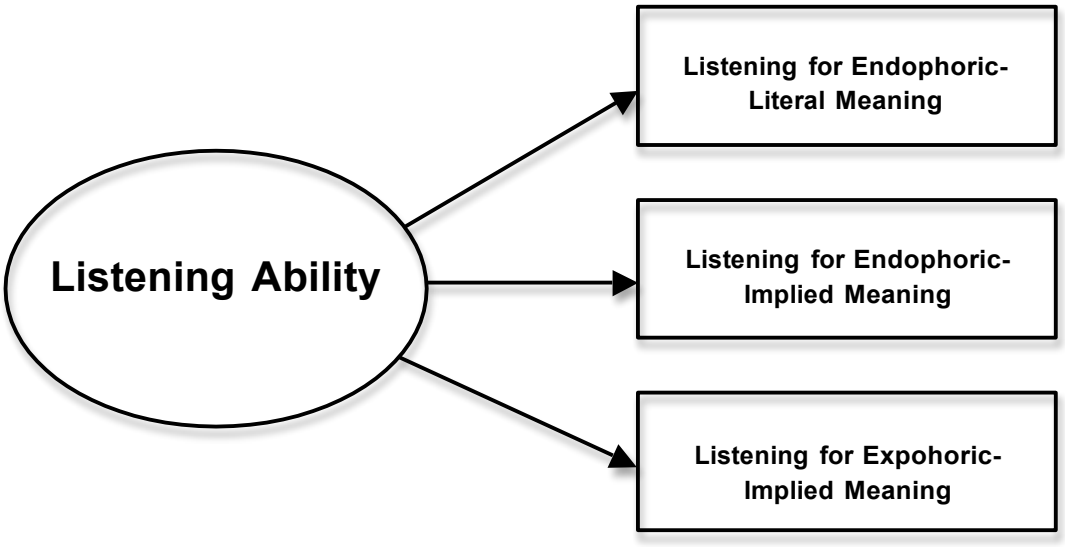

Figure 1. The construct of listening ability used in the current study.

2. Theoretical Model: Speaking Ability

Speaking ability was also defined as a target skill in the "Scope and Sequence" portion in the learners' textbook. In this particular unit of *In Charge 2* (Daise, 2003), the functional speaking objective was to improve students' discussion abilities in the areas of turn-taking, clarification of miscommunication, and staying on track. Because these language functions require group interaction to complete, it was clear that students' ability could not be assessed purely through a monologic form; at least one of our constructs would necessitate the inclusion of ability to perform interactional practices (openings, closing, turn-taking, etc.) within a group setting.

In defining speaking ability theoretically, Bachman and Palmer's (1996) framework was again utilized; this time, the adaptation presented by Luoma (2004) was used. Three of the language competence components of Bachman and Palmer's (1996) framework, highlighted by Luoma (2004), were integrated: *grammatical knowledge*, *textual knowledge*, and *functional knowledge*. In adapting these constructs to suit the needs of the test, terminology was adjusted so

as to broaden the categories and make them more accessible in terms of common terminology used in the classroom.

First, *grammatical knowledge,* defined by Luoma as "how individual utterances or sentences are organized" (p. 100), became *language control,* which encompasses the general grammatical accuracy of statements as well as the overall complexity and variety of sentence structures used. Second, *textual knowledge*, which Luoma designates as "how utterances or sentences are organised to form texts" (p. 100) became *organizational control*, which is defined more broadly as focus on overall cohesiveness of speech. Fluency was also added into this dimension of organizational control, and was defined as an absence of excessive pausing or hesitation, and general well-connectedness in speech. The third category is *functional knowledge,* which Luoma designates as, "how utterances or sentences and texts are related to the communicative goals of language users," (p. 100). This notion is derived from concepts presented by Halliday and Hasan (1976), and includes ideational, manipulative, heuristic, and imaginative functions of language. As these functions are intended to help language users "exchange ideas, exert an impact on the world around us, extend our knowledge of the world, and create an imaginary world characterized by humorous or aesthetic expression," (Cummins, 2000, p. 124) they essentially indicate appropriacy of language used in particular situations. Thus, the concept of *functional knowledge* was modified to *conversational control.* This construct represents the students' ability to appropriately use the discussion markers explicitly taught in this chapter of the textbook in order to facilitate their conversations.

In justifying the use of conversational control as a speaking component, it is useful to examine the deep underlying connection between listening and speaking skills, which are largely inextricable. Historically, listening has been characterized as a receptive skill and speaking as

productive; however, this is not always an accurate distinction, particularly when looking at interactions, which require the integration of both areas. Instead, listening is an "interactive, interpretive process in which listeners engage in a dynamic construction of meaning" (Murphy, 1991, p. 56). Therefore, speaking and listening ability should both be seen as important factors when assessing group performance. As Clark and Hecht (1983) state, "language use demands that two distinct processes -- production and comprehension -- be coordinated. And that in turn suggests that one part of acquisition consists of coordinating what one can produce with what one can understand" (p. 326). Thus to "coordinate" these skills, the constructed-response speaking task utilized the construct of *conversational control*. Although the test takers were responsible for accurate, complex, and organized utterances, they also needed to facilitate the conversation using appropriate discussion markers. The only way for them to correctly use these markers would be through first attending to the input produced by the other members of their group and then use the correct discussion maker to clarify miscommunications, stay on track, or invite other speakers to give their opinions. In using *conversational control* as a construct, speaking and listening abilities were conceptualized as highly interrelated.

Figure 2 shows the construct of speaking ability as it has been interpreted for this unit test, consisting of *language control, organizational control,* and *conversational control.* Specifically, these constructs were also blended with and parallel to those presented on a rubric currently in use by the CEP, which reflects the TLU domain of the CEP.

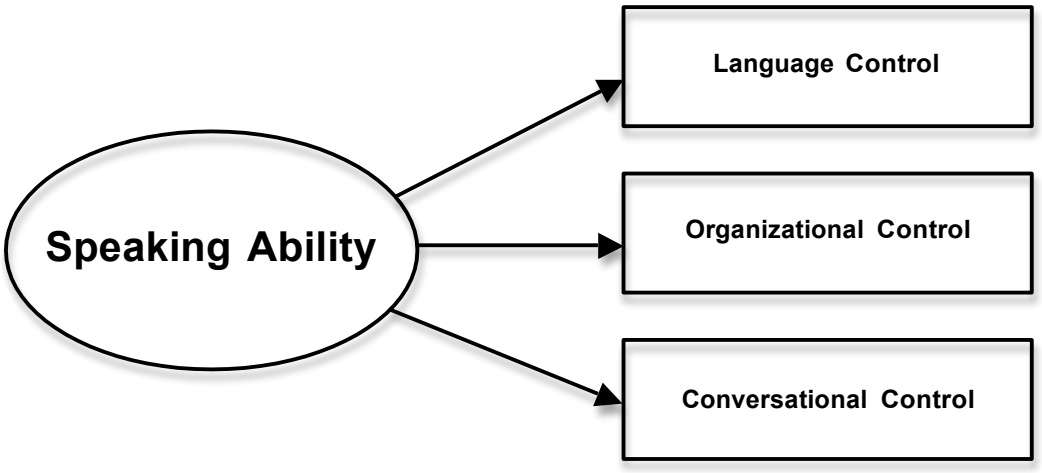

Figure 2. The construct of speaking ability used in the current study.

3. Theoretical Model: Connection between Listening & Speaking Ability

Physically speaking, there is much debate as to how much, if at all, speech comprehension and production processes interact in the brain. This debate prompted Menenti, Gierhan, Segaert, and Hagoort's (2011) study on the overlap of speaking and listening processes in the brain. The researchers used functional MRI (fMRI) scans to measure the brain's activity when participants produced and comprehended active and passive sentences. Findings revealed that the neuronal infrastructure responsible for semantic, lexical, and syntactic processing is shared, meaning that "[l]anguage production and comprehension are two facets of one language system in the brain" (p. 1179). In other words, listening and speaking are connected processes. Because of the interrelatedness of the processes in the brain, there is a likely correlation between listening and speaking ability.

In addition to having a physical, observable link in the brain, second language acquisition (SLA) research situates listening and speaking on Van Patten's (1996) model of L2 acquisition for oral communication; listening is associated with input, intake, and uptake processes, while speaking is a form of output. In order for a learner to successfully produce an utterance, its

phonological, morphological, and lexical components at some point were processed as input by the learner. This relationship between input and output is most evident in interactions, where interlocutors must attend to both listening and responding in real time.

Because of the primacy of input in SLA, which claims that listening lays the foundation for speaking (Murphy, 1991), it was predicted that listening ability and speaking ability would share a strong correlation in the test. Bozorgian (2012) also describes listening as the primary channel of learning a language, proceeding speaking, as exemplified by a learner's reflection: "I understand everything you say, but I can't repeat it" (p. 3). One reason why listening improves speaking is evidenced in Swain's (1995) Output Hypothesis, which maintains that output (in this case speaking) can facilitate noticing input (listening) when learners recognize gaps in their linguistic knowledge through output. Once learners notice their gaps, they are more likely to attend to input that will fill these gaps and make their utterances more target-like. Thus, learners who have a high proficiency in speaking are likely to have employed noticing strategies in their listening, an indicator of the necessity of listening in improving speaking ability.

Finally, the construct of *conversational control* assessed in the speaking task assumes a connection between listening and speaking abilities. Without simultaneous attention to both listening and speaking, learners would be unable to use the discourse markers taught in Unit 2 accurately. Figure 3 shows the construct of speaking ability as it is related to listening ability in this unit test.

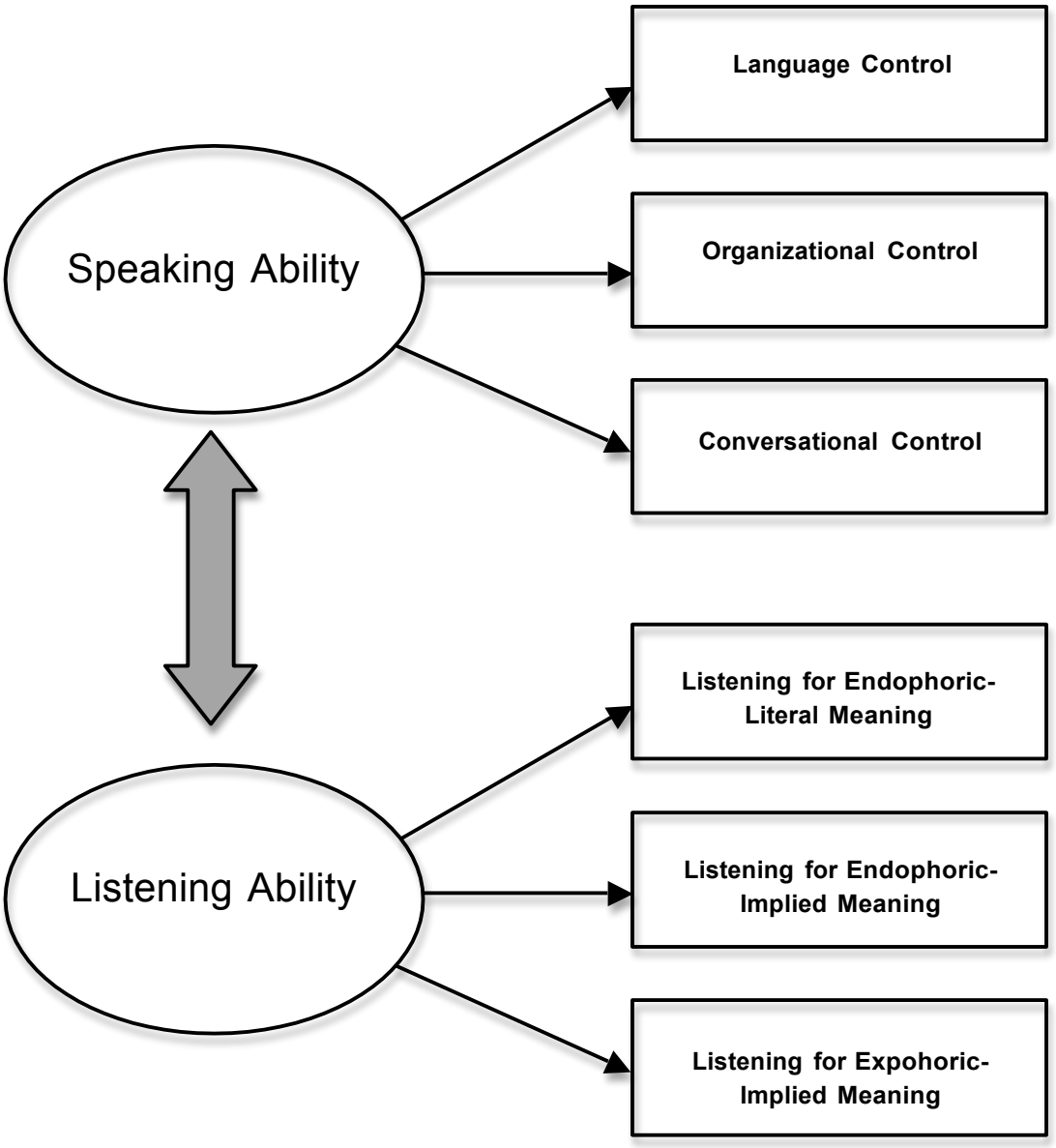

Figure 3. The construct of listening ability and speaking ability used in the current study.

4. TLU Domain

The content for this test was selected both on the basis of thematic appropriacy as well as the desired target language-use domain: academic, business, and daily-life interactional English contexts. For the listening section, the test focused on English for daily-life interactional contexts by providing a number of authentic dialogues between varying interlocutors, which are

representative of the conversations that the students might encounter outside the classroom. The speaking section intended to target the academic and business domains by prompting the students to engage in a discussion using conversational markers that are commonly employed by expert language users in these domains.

5. Operationalization

Listening ability was operationalized in a selected-response task containing 15 total dichotomously-scored multiple-choice questions. Three listening passages were used, each in conjunction with five questions corresponding to the passage. The total length of the listening test was approximately 20 minutes; the listening passages were all around one minute and 45 seconds in length, followed by four minutes to complete the questions in each section.

Speaking ability was operationalized in an extended-production task consisting of one prompted discussion scored through the use of an analytic rubric, scaled from 1-5. The task was conducted in groups of three (and one group of four). Each individual was given three minutes to read the task sheet, and five minutes to discuss the prompt with their group-mates. The students recorded their responses. An overview of the test may be seen in Table 1.

Table 1: *Overview of Test Structure*

Task Component	Task Type	Length (Items)	Time (Minutes)	Topic	Scoring
Listening Ability - *Endophoric-Literal Meaning* - *Endophoric-Implied Meaning* - *Exophoric-Implied Meaning*	Selected-response	15 Items total 5 items per passage	20 minutes (including listening passages)	Dreams Sleepwalking, Dream Journals, Painting from Dreams	Dichotomous scoring 15 points total
Speaking Ability - *Language Control* - *Organizational Control* - *Conversational Control*	Extended-production discussion	1 question	3 minutes preparation 5 minutes speaking	Personal Dreams Strategies for Achieving Dreams	Analytic rubric, scaled from 1-5 for each of the following: - Language Control - Organizational Control - Conversational Control 15 points total, 2 raters, Average score

Following the test, two raters assigned scores to the test-takers. In order to establish a high level of concordance among the two different raters, a norming session was held. In this session, the raters analyzed one of the student responses together in order to ensure mutual comprehension and application of the rubric.

6. The Test

A copy of the test may be found in Appendix A

7. Item Coding

The multiple-choice listening items were then coded by category and answer key, as shown in Table 2. There are three broader variable categories for the listening passage, each containing five items. These items can be further categorized by type.

Table 2: *Listening Item Coding*

Observed Variable	Focus	Key	Item
Endophoric-Literal	Identifying Main Idea	A	6
	Identifying Detail	C	1
		D	2
		A	11
		C	12
Endophoric- Implied	Inferring Main Idea	A	3
		D	13
		B	14
	Inferring Detail	C	7
		C	8
Exophoric- Implied	Deriving Contextual Meaning	B	5
		A	9
	Deriving Psychological Meaning	B	10
		C	15
	Deriving Rhetorical Meaning	B	4

D. Administration Procedures

The test was administered on Thursday, October 16, 2014 to 19 students in the CEP. As three students were absent, a make-up examination was proctored on Sunday, October 26th, following the same procedures. The order of the test sections was as follows, in chronological order: speaking, listening, reading, and writing. To begin the speaking test, students were put into speaking groups by the teachers. Then the speaking prompt was distributed and the teacher read through the prompt and directions. Students were given three minutes to finish reading and prepare their responses for speaking. The groups recorded the speaking section using handheld voice recorders – one per group. For the listening section, answers sheets were given out and instructions were read aloud. For each listening passage, the audio file was played and students

were given four minutes to answer the corresponding questions before the next section began. The students were not given additional time to answer the questions.

III. TEST ANALYSES AND RESULTS

A. Results for Listening Task

1. Descriptive Statistics

22 students (n=22) participated in the listening exam. The total number of listening items was 15 (k=15). Overall, the mean was 10.09, and the mode and median were both 11. The distribution indicated negative skewness at -1.38 with positive kurtosis at 1.83. The maximum possible score was 15. The minimum score received by the test-takers was 5 and the maximum was 13, with an overall range of 9 points. The standard deviation was 1.93. A summary of these results may be found in Table 3.

Table 3: *Listening task descriptive statistics.*

			Central Tendency			**Distribution**		**Dispersion**			
	N	K	Mean	Mode	Median	Skewness	Kurtosis	Min	Max	Range	SD
Listening Tot	22	15	10.09	11	11	-1.38	1.83	5	13	9	1.93
Endophoric-Literal Tot	22	5	4.45	5	5	-1.68	3.04	2	5	4	.80
Endophoric-Implied Tot	22	5	2.91	3	3	.15	-1.11	2	4	3	.75
Exophoric-Implied Tot	22	5	2.73	3	3	-.71	-.33	1	4	4	.98

Overall, the average score of 11 indicates that the students performed at or above "average," which we will quantify as approximately 70%. This is reflected in the slight negative skewness, as shown in Table 3 and Figure 3. As the test is an achievement test, it is expected that the students perform at or above the average, considering that they are all motivated to do well

and the teacher spent a sufficient amount of time addressing the target listening skills in class. The lack of significant outliers is also an indication of the general propensity of the group to perform well. Additionally, as each listening passage directly addressed aspects of the TLU domain with which the students are familiar, such as a conversation between friends, a conflict between siblings, and a radio interview, the passages did not include a great deal of complex vocabulary or conceptual content that may have been more challenging. These types of conversation do not require that the listener have specific topical knowledge, as they are fairly common types of interactions to which test-takers have been exposed. Furthermore, the passages were relatively short in order to lower cognitive processing demands. Thus, the passage may have been relatively easy for some students to process, hence the fairly high central tendencies.

Figure 4 depicts the leptokurtic distribution of the test. This distribution, as indicated by the positive kurtosis total of 1.83, indicates there was little variability within the group. This could be due to accurate level placement of students in terms of their listening comprehension ability; in other words, the majority of students seem to possess comparable skills in this area.

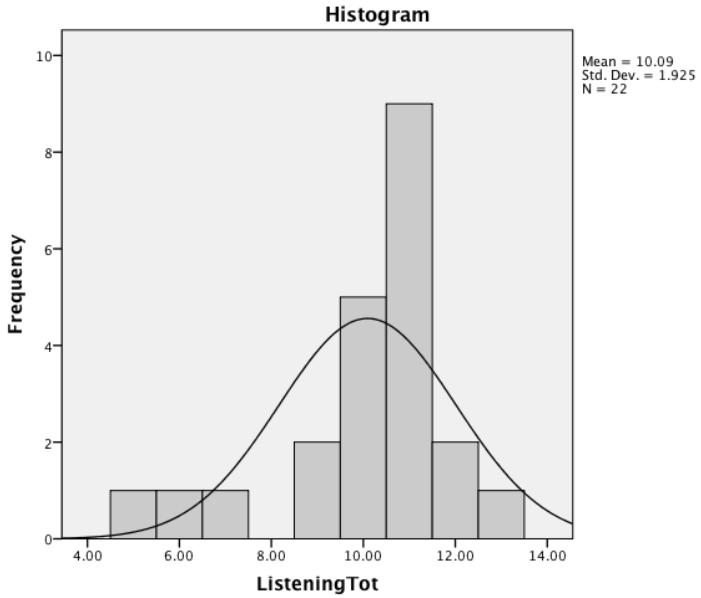

Figure 4. Histogram of listening scores.

Although the overall central tendency was to do well on the test, a breakdown of the composite variables for each section of the listening test (*listening for endophoric-literal meaning*, *listening for endophoric-implied meaning*, and *listening for exophoric-implied meaning*) revealed a number of discrepancies. These discrepancies confirmed some of the predictions about the difficulty of certain types of items. Prior to administration of the test, it was hypothesized that the endophoric-literal items would be the easiest for students to correctly answer, as they test information which is directly provided in the passage and requires no interpretation. It was also predicted that endophoric-implied items would be the most challenging, as they require the listener to infer connections between multiple disparate parts of the listening passage, which is heard only once. This skill is even more complex than that of deriving exophoric-implied meaning, which requires the application of the passage to prior knowledge, a process that naturally occurs when input is converted into intake. The difference in difficulty of these item types was evidenced in the data; the endophoric-literal items displayed negative skewness at -1.68, even higher than the overall listening skewness at -1.38, indicating that the majority of test takers scored at or above average. Moreover, the exophoric-implied items also had a negative skewness, indicating a slight tendency to score at or above average. In contrast, the endophoric-implied items were the only ones that had a positive skewness at .15, reflecting the challenging nature of these items.

Another anomalistic tendency was reflected in the highly leptokurtotic distribution of the endophoric-literal items; not only was the tendency positive at 3.04, but it also was markedly higher than the overall listening kurtosis at 1.83. This indicates that there was very little variability in the group at performing these types of items. On the other hand, the other two item categories, endophoric-implied and exophoric-implied displayed a negative kurtosis at -1.11 and

-.33 respectively. This clearly shows variability among scores in these areas. One explanation for this result is that the lower-scoring students needed more time to focus on practicing inferencing skills in class rather than spending time practicing identifying skills. This is excellent formative information for the teachers, who know to instead focus more on inferencing skills in future classes.

2. Internal-consistency Reliability and Standard Error of Measurement

The internal reliability of a test is based upon the extent to which different items on a test are able to measure the same overall construct. In order to calculate this figure, Cronbach's alpha was used, as it is the most widely-accepted statistical calculation of internal-consistency reliability. Cronbach's alpha for the listening test, which contained 14 items and was given to 22 test takers, was .45. Originally, 15 items were included in the test. However, as one item was answered correctly by all test takers, it was removed from analysis, as it provided no meaningful information to discriminate among test takers' performance. This information is summarized in Table 4.

Table 4: *Reliability of the Listening Task*

(n=22)

Cronbach's alpha	k (# of items)
.45	14

The reliability of this section of the test is moderately low. It is preferable that classroom tests have minimum reliability in the range of .67-.80, depending on the impact of the test (Carr, 2011). The level of reliability displayed here indicates that at least 55% of the test was attributable to construct-irrelevant variance, or error variance, while only 45% is due to true score variance. This means that there is more error than actual discriminating test content on this

section of the exam. This means that the majority of the test ought to be revised if used again in the future.

The low reliability could have been caused by the heterogeneous nature of the questions. Because there were, within the main construct of listening ability, three constructs being measured, further divided into two, two, and three additional sub-constructs respectively, there were a total of seven different sub-constructs being measured on a fourteen-item test. To address this issue in the future, it might be better to focus on one of these subconstructs or to include additional questions in order to create a more holistic indication of listening ability.

Furthermore, the low reliability is also an indicator that some high-achieving test takers missed questions that were correctly answered by low-achieving test takers, and vice versa. This could, on the one hand, point to individual differences between test takers; but, more likely, it is an indication of poorly written items. This line of inquiry will be further pursued in the following section on item analysis.

One way that reliability directly impacts classroom test score reporting is in its use as part of the equation for determining the acceptable cut scores for students within the given classroom context. In the CEP, a passing score is a 70%. This means that on a fourteen-item test, 10 items must be answered correctly in order for a student to obtain a passing grade.

Because of the low reliability of the test, it would be unethical for the teachers to report those raw scores to their students. In order to figure out a true cut score for this exam, the standard error of measurement (SEM) was sought. The SEM was calculated using the following formula: $SEM = S\sqrt{1 - r_{xx'}}$, where S=1.93 (the standard deviation) and $r_{xx'} = .45$ (the test's reliability). The resulting SEM was 1.66.

A true cut score takes into consideration the standard error of measurement, so the SEM was subtracted from the cut score of 10 once, and twice, to determine the acceptable passing range within different confidence levels. At a 68% confidence interval (±1 SEM), the cut score would be 8.34 (rounded up to 9); at a 95% confidence interval (±2 SEMs) the cut score would be 6.68 (rounded up to 7). These numbers were rounded up to the next integer as this was a dichotomously-scored test and no partial credit was given. If this were a high-stakes testing environment, it would be necessary to consider the 95% confidence interval as the standard. However, in a classroom achievement test context, the 68% confidence interval is generally considered acceptable. Therefore, test takers with scores at or above 9 on the test received a passing grade.

3. Item Analyses

Item analyses were conducted in order to determine whether the items on the test were of the appropriate difficulty level for students and whether they adequately discriminated between high-achieving and low-achieving test-takers.

Item difficulty was measured using *p-value*, which indicates the proportion of test-takers that answered the item correctly compared to all test-takers who answered the item. For classroom achievement tests, p-values are ideally between .6 and .95, meaning that between 60 and 95 percent of the test-takers were able to answer the item correctly. Values below .6 would indicate some type of gap between what was taught and what was assessed or an inadequacy of the question itself.

The *D-index*, or discrimination index, measures how well the item discriminates between high performers who answered correctly and low performers who answer correctly. The D-index range is from -1 to 1. If, for instance, every test-taker were to answer an item either correctly or

incorrectly, then the D-index would be 0, indicating that item is unable to discriminate. If every low-performing test-taker were to answer an item correctly that other high-performing test-takers were unable to answer correctly, the D-index would likely become negative. This would indicate that this item is problematic as it fails to separate and rank students based on their performance. Ideally, the D-index of each item should be above .4 in order for the item to be considered of superior quality; however, a D-index of .3 is also considered acceptable in a classroom assessment context.

Table 5 depicts the p-values (difficulty) and D-index (discrimination) of the 15 test items. It also contains an indication of what the Cronbach's alpha, or reliability, would be for the test if it were removed. Finally, a decision is indicated for each item, which expresses what decision was made regarding the item upon further analysis: to keep, to revise, or to remove the item. In order to determine whether items needed revision, they were evaluated based on the p-value and D-index ranges. If the difficulty was too high or the item failed to discriminate, removal of the item overall was also considered.

Table 5: *Item Analysis of Listening Task*

Item	Difficulty p-value	Discrimination D-index	Alpha if Deleted	Decision
IdDet1	.95	.51	.38	Keep
IdDet2	.95	-.32	.50	Remove
InMain3	.95	-.10	.47	Remove
DeRhet4	.82	.50	.33	Keep
DeCon5	.55	.08	.46	Revise
IdMain6	.77	.36	.37	Keep
InDet7	.86	.12	.44	Revise
InDet8	.14	.24	.41	Revise
DeCon9	.27	.18	.42	Revise
DePsy10	.36	-.09	.51	Remove
IdDet11	1.00	0.00	--*	Remove*
IdDet12	.77	.22	.41	Revise
InMain13	.77	.22	.41	Revise
InMain14	.18	.09	.48	Revise
DePsy15	.73	.19	.42	Revise

*Note: Item 11 was deleted from statistical analysis by SPSS; all test-takers answered this item correctly, therefore it was unable to discriminate between high and low achieving test-takers.

Overall, the difficulty of the items ranged from .14 to 1, indicating a wide range of variability of p-values. Nine of the items were in the acceptable .6 to .95 range, indicating a generally acceptable difficulty level of the test overall. Some items clearly were too difficult or too easy and need further revision in order to ensure that the test is fair and measures what it should.

The overall discrimination of items, however, was more problematic. The D-index ranged from -.32 all the way to .51, which indicates that items were, generally speaking, insufficient at discriminating among bands of performance. Three of the items even had negative D-indices, and nine had a D-index below the acceptable .3 margin. The highly disproportionate number of

items might have been one of the causes for a low reliability; thus it was necessary to remove items with low D-indices in order to repair the test and elevate reliability. The low D-indices were perhaps due to the complexity of the constructs, individual variation within the class, and the challenges associated with measuring and applying this particular theoretical framework of listening ability. As this was intended to be a model for reading ability, there were many aspects of the framework that were difficult to apply – for instance, inference questions are quite difficult in listening as the hearer is unable to go back into the text as they would in a reading passage.

Three items were deemed wholly acceptable in terms of both difficulty and discrimination: items 1, 4, and 6. The best performing items were 4 and 6; both had p-values within the acceptable range, with 77 and 82 percent of test takers answering the item correctly. The D-index for item 4, however, was much better than item 6 with a D-index of .5 whereas item 6's D-index was only acceptable at .36.

A number of items were selected for revision, as they did not fall within the acceptable ranges for both p-value and D-index. Items numbered 5, 8, 9, and 14 are problematic, as both their p-values and D-indices were deemed unacceptable and require revision because the questions were too difficult and did not discriminate well. The p-value (.18) and D-index (.09) of Item 14 exemplify the poor performance of this group. Other items, such as 7, 12, 13, and 15, although having acceptable p-values, still require revision as their D-indices are too low. For example, number 7 has a great p-value at .86, however its D-index of .12 was a result of being answered correctly by low-performing test-takers.

Finally, items were selected for removal not only based on their p-values and D-indices, but also the potential improvement that their removal could bring to enhance the overall test reliability. Without removal of any questions, the test reliability was low at .45. It was decided to

first remove items that possessed the lowest D-indices, items 2, 3, and 10; Table 8 reveals that three items have a negative D-index. These items also have the highest increase in Alpha if deleted, thus they were removed. Additionally, item 11 was automatically removed as it was answered correctly by all test-takers, thus providing no information about differences in test-taker ability. Although it is clearly not preferable to have a test with only 11 items, the removal of these items was beneficial as it increased the reliability to .58, a 13% increase, as indicated in Table 6. These negatively discriminating items were also unfair to test takers, as they did not discriminate well with this group of students.

Table 6: *Reliability of test post-deletion of problematic items.*

(n=22)

Cronbach's Alpha	K (# of items)
.58	11

4. Distractor Analyses

Distractor analyses were conducted, primarily to gain insight as to why test-takers or varying levels selected certain responses. Additionally, this type of analysis helps to determine how items may be revised for improvement.

Two items, 5 and 8, were chosen for analysis based on their low p-values and low D-indices, as these numbers indicated that the questions were both too difficult and did not discriminate effectively. Item 5 was an exophoric-implied item, coded as "deriving contextual meaning" and Item 8 was an endophoric-implied item, coded as "inferring detail." These items were also selected for their dissimilarity; although both items test ability to understand non-literal meaning, they ultimately measure different constructs: one measures the ability to infer meaning while the other measures the ability to derive meaning by using the test-taker's outside knowledge.

In order to determine the p-value and D-index for each item by hand, it was necessary to determine which test takers fell under the bands of "high-performing" and "low-performing" groups. The 22 test-takers were split into thirds in order to create a high group (n_h=7) and a low group (n_l=7) that were sufficiently large enough to use for distractor analysis. The high-performing group consists of students who scored at least 11 and the low-performing scored below 10.

Item 8 is shown in Figure 5. The item responses were accidentally jumbled on the test itself; note that the answer options read sequentially "A-D-B-C" and not "A-B-C-D." Though it is possible that this may have caused confusion among the test-takers, all of them circled their answers as was requested by the directions, and did not write out the letter of their response. Thus it is believed that any error that may have resulted due to this mislabeling is unlikely, though possible.

8. Liz believes that dreams represent _____.

a. future possibilities

d. pieces of your memory

b. your greatest hopes and fears

c. a connection between people.

Key = c

Figure 5. Item 8 from the Listening Task.

Although the key was C, it was revealing to discover that 11 out of the 22 students had selected B, seven of whom were in both the high and low performing groups. A greater proportion of high-performing test-takers selected B over C, the key. A small but fair proportion of all test-takers selected A and D (4 and 4 respectively), showing that these distractors performed adequately. However, the key, C, was the least selected option test-takers overall. This is problematic for an achievement test, where it is desirable that the larger proportion of

students, particularly the high-achieving students, would select the correct answer. A summary of these frequencies of student responses can be found in Table 8.

Table 8: *Distractor Analysis for Item 8 from the Listening Task*

Selected-response	Frequency	High (n=7)	Low (n=7)
A	4	2	0
B	11	3	4
C (key)	3	2	0
D	4	0	3

The key, "a connection between people," was intended to evoke an inference from the following statement in the listening passage: "I think if a lot of people have experienced this same thing, it must mean that there's some kind of deeper, more universal meaning to dreams…" The test-taker would have had to infer that "universal" indicates something that connects people as it is mutually shared among them. However, a number of students selected B, "your greatest hopes and fears." This was intended to distract as it does not refer specifically to universality. However, if one steps back and analyzes the listening passage on a broader level, this option could be viewed as plausible. This is because the speaker, Liz, explains that she started writing in her dream journal after she had had a nightmare about spiders, which prompted her to believe that the dream was actually about a fight she had recently had with her mother. Because the dream Liz had was actually a nightmare, one could say that she believed that dreams and fears are connected. It would not be too great a leap to think that if one can dream about fears, they can also dream about hopes. Additionally, the "and" conjunction uniting the two components of the response, "hopes *and* fears" may have been confusing for students; if the answer is one part but not the other (hopes or fears), is the item still plausible? Students may still be likely to pick the item regardless, and leaving in both options makes the item unfair. Because more high-

achieving test takers chose B (this distractor) instead of C (the key), it called into question the validity of this response and prompted the researchers to consider how to adjust the results. On future tests, this response would be replaced with a better response; for the purposes of this examination, it was decided that the option B should be included as another possible key.

Upon rekeying the item, the reliability of the test went slightly down, as shown in Table 9. However, in the interest of fairness, and since the change was so minimal (.01), it was decided that this updated reliability would be utilized.

Table 9: *Reliability of Test Post-deletion of Problematic Items and Rekeying of Item 8.*

Cronbach's Alpha	K (# of items)
.57	11

The next item that was selected for analysis was an exophoric-implied question, Item 5 on the listening section. It was selected because of the low D-index (.08), indicating that the item hardly discriminated among test-takers. In addition, the question was somewhat difficult with a low p-value (.55). The item is displayed in Figure 6.

5. This conversation probably took place at _____.
a. work
b. school
c. an apartment
d. a coffee shop
Key = b

Figure 6. Item 5 from the Listening Task

Although the correct answer was B, "school," a large number of high-performing test-takers selected C, "an apartment." It was expected that the test-takers would pick B because of the following context within the listening passage: "That psychology class was so interesting!" and the response, "I know, sleep disorders are so fascinating!" From this, it was assumed that

students would understand that the two speakers were talking at some point in time following a class. The use of the referent, "that" implies an exophoric connection to a class which had previously occurred at a point in the recent past. This could have been made slightly clearer if there had been additional clues within the transcript to indicate that the speakers had just finished class or were having this conversation in passing in the hallway. For example, one of the speakers could have more explicitly stated, "That psychology class we were just in was so fascinating!" But obviously, providing such explicit context makes the conversation far more unnatural, and also transforms the question from an exophoric-inferential question to more of a literal-detail question. Instead, it was decided that a better conclusion to the passage that indicated that the two speakers were then leaving and going to other classes could have improved the passage. For instance, saying, "Oh, sorry, I have to run to Dr. Fuch's class now," or, "Oh shoot, I think I am going to be late for assessment class, gotta run!" would have made this context a little clearer. The addition of background noises - perhaps a school bell, the sound of shuffling students, or closing doors - could have better indicated that this conversation took place in the hallway of a school.

Despite the issues raised above, twelve students (approximately half of the class) were able to select the correct key. One of those students had originally selected D, "a coffee shop," but then changed her answer to the key, B, and even wrote a note on the item: "because they are talking about class." This indicates that she *was* using the skill of deriving contextual meaning by inferring that because the speakers in the conversation were talking about school, it was likely that they were in fact at school. The frequencies of test-taker response selections for Item 5 are indicated in Table 10.

Table 10: *Distractor Analysis for Item 5 from the Listening Task*

Selected-response	Frequency	High (n=7)	Low (n=7)
A	0	0	0
B (key)	12	3	3
C	8	4	3
D	2	0	1

When looking to see why test-takers chose other distractors, it is important to note that A, B, and D are all public spaces whereas C is a private space. Because there is no background noise on the soundtrack, the test-takers might have selected C, "an apartment," because it is most likely the quietest location. And in fact, the recording *did* actually take place in an apartment. The test-takers were likely drawing their inferences from factors such as the fact that this took place between two friends, it was a personal anecdote, and the lack of a clear conclusion indicating that the two speakers were going their separate ways. In order to make all of the distractors more equal, C should be changed to another public space, such as a park or shopping center if the test were to be administered again.

One of the distractors, A, "work," was not selected by any students. This has been referred to by an assessment expert as a "potato" (Grabowski, 2014, Personal Communication). It was most likely that this option was not selected because the directions for the task state: "The following is a conversation about sleepwalking between two friends, Katie and Emily." By labeling the speakers "friends," test-takers may have been wary to add the additional label of "co-workers," to the two speakers. Instead, they might look to the other options because they are places that friends are more likely to be together. Therefore, the directions should be changed to

say, “between two speakers,” so that the test-takers do not have any preconceived notions about the speakers’ relationship prior to hearing the passage.

Overall, the item is successful in that it causes test-takers to use inferencing skills to derive contextual meaning for the listening passage. However, the basis of this inference is a quick, short exchange at the beginning of the listening passage. This item needs revision to create a listening passage which offers more clues upon which the inferences can be based. These clues should not only occur at the beginning of the passage, but throughout. Hopefully, this would raise the p-value of the item, making the item easier. By changing the distractors and the task directions, it is also hoped that the D-index would increase by causing higher-achieving test-takers to select the correct answer.

5. Evidence of Construct Validity within the MC Task

Construct validity refers to the extent to which a test measures its intended underlying theoretical constructs. One way construct validity may be established is through the provision of correlational evidence. To do so, correlations between the variables comprising the construct of listening ability were calculated through the use of a Pearson product-moment correlation, which measures the magnitude of the inter-relatedness of these variables on a scale ranging from -1.00 to 1.00. Correlations may be high (>.75), moderate (.5 to .74), low (.25 to .49) or uncorrelated (<.25). Additionally, correlations may also be statistically significant, indicating the generalizability of the correlation and the extent to which it might be said that this correlation was not due purely to chance factors.

The results are shown in Table 11. Overall, there was no correlation between endophoric-literal and endophoric-implied items (.21), a low correlation between endophoric-literal and exophoric-implied items (.36), and a highly moderate as well as statistically significant (to the

.01 level) correlation between endophoric-implied and exophoric-implied items (.62**).

Table 11: *Correlation Matrix for the Listening Task*

	Endophoric-Literal Total	**Endophoric-Implied Total**	**Exophoric-Implied Total**
Endophoric-Literal Total	(1.00)		
Endophoric-Implied Total	.21	(1.00)	
Exophoric-Implied Total	.36	.62**	(1.00)

** Correlation is significant at the 0.01 level

Because both endophoric- and exophoric-implied items require inferencing skills, it is unsurprising that the correlation was both high and statistically significant. Regardless of whether the items sought out information that was inside (endophoric) or outside (exophoric) the listening passage, both types of items employ a type of cognitive processing that requires the test-taker to go beyond simple repetition of literal, stated information. The test-taker must use skills such as synthesizing information, looking for underlying concepts, decontextualizing meaning, and other such cognitive strategies in order to correctly answer an inference question. On the other hand, literal questions require more basic listening skills such as basic auditory processing and retrieval. Hence, it is logical that the endophoric-literal items had little or no correlation with the endophoric-implied and exophoric-implied items.

B. Results for Speaking Task

1. Descriptive Statistics

22 students participated in the speaking exam. Students responded to one prompt, which was scaled out of five points. The speaking average was calculated by combining rater averages

and dividing the it to place the scores back onto a 1-5 point scale.. Overall, the mean for the speaking test was 3.70, the mode was 3.33, and the median was 3.67. There was positive skewness at .057 and negative kurtosis at -1.01. The minimum score was 2.83 while the maximum was 4.5, with a total range of 2.67 points. The standard deviation was .50. These results are summarized in Table 12.

Table 12: *Speaking task descriptive statistics.*

			Central Tendency			**Distribution**		**Dispersion**			
	N	K	Mean	Mode	Median	Skewness	Kurtosis	Min	Max	Range	SD
Speaking Avg	22	1	3.70	3.33	3.67	.057	-1.01	2.83	4.5	2.67	.50
Gram Control Avg	22	1	3.48	3.50	3.50	0.97	-.13	2.5	4.5	3	.48
Org Control Avg	22	1	3.66	3.50	3.50	0.530	-.54	2.5	5	3.5	.70
Convo Control Avg	22	1	3.98	4.00	4.00	-.291	-.81	2.5	5	3.5	.75

When looking at the individual composite variables, some unique features were discovered. First, there was a negative skewness of -.29 in conversational control, as compared to the positive skewness of organizational control at .53 and .97 of grammatical control. This difference is perhaps related to the nature of the prompt that was used; in the task, the students were asked to incorporate specific conversational tokens within the conversation. This required the students to have some familiarity and practice with these tokens. Because they were covered in class and the students were aware prior to the test that they would need to study these tokens for the exam, they were all prepared to use them. Additionally, because the class is so large at 23 students, it is expected that the students work in groups on a daily basis, so they are exceptionally well-acquainted with each other and feel comfortable discussing in a way that encourages individual participation during group interactions. The positive skewness of

grammatical accuracy and organizational control shows that the teachers should focus more on these areas when practicing speaking in small groups, rather than focusing solely on the aspect of conversational control. However, as the curriculum from the textbook requires that the teacher teach conversational discussion skills, it is understandable that the teacher chose this focus. Nevertheless, the students would benefit greatly from further practice in the areas of grammatical control and organizational control in speaking.

Furthermore, the means of both grammatical control and organizational control are significantly lower, which brought down the central tendency in these areas, leading to a positive skewness. The reason for these lower means may be a result of the raters' mutual decision not to assign many perfect scores in these categories.

All of the composite variables, including the overall speaking average, indicated a platykurtic distribution, as visible in Figure 7. This reveals that the students are widely distributed in terms of their speaking abilities. This may be a reflection of the lack of a speaking placement test at the CEP; students are placed into levels based on their reading, writing, and listening skills, and an oral proficiency test is currently not conducted. Thus students in this class have a wide range of speaking abilities.

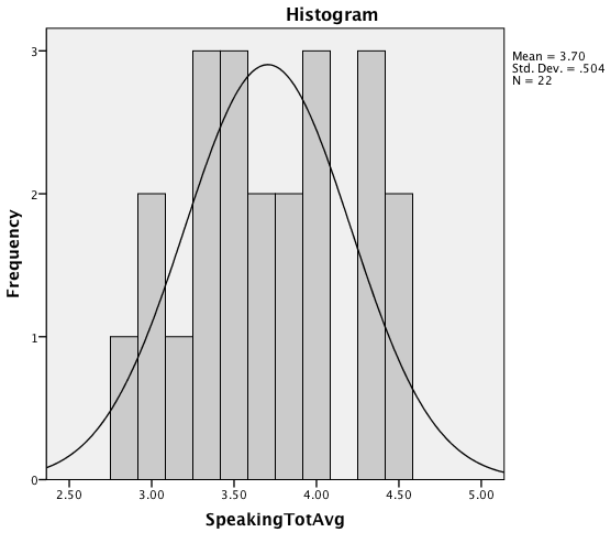

Figure 7. *Histogram of total speaking scores.*

2. Internal-consistency Reliability and Standard Error of Measurement

The internal-consistency reliability for the speaking section was also calculated according to Cronbach's alpha. This was measured according to three separate averaged variables for grammatical control, organizational control, and conversational control. To determine these averages, the scores of both raters in these categories were added and divided by the number of raters (2). For example, if a test taker received a 3 in grammatical control from Rater 1 and a 4 from Rater 2, their averaged score for that section would be 3.5. Taking into account all of the averaged scores in the three domains, the result of Cronbach's alpha was .66, as displayed in Table 13.

Table 13: *Reliability of the Speaking Task*

(n=22)

Cronbach's alpha	**k (# of items)**
.66	1

This internal-consistency reliability result was nearly acceptable, though somewhat problematic. According to this calculation, 44% of the resulting test scores were attributable to construct-irrelevant variance and not to true score variance, which comprised 66% of the results. Compared to the multiple-choice section of the test, these results are significantly more reliable; nonetheless there is certainly room for improvement.

The standard error of measurement (SEM) on this test was calculated using the following formula: $SEM = S\sqrt{1 - r_{xx'}}$, where S=0.5 (the standard deviation) and $r_{xx'} = .66$ (the standard deviation).The resulting SEM was .29, which is relatively low.

The SEM is used to inform the adjusted cut-score, which can be used to accurately report the results to the students. Although there was only 1 prompt on the test, it was scored out of 15 points, divided into three construct categories, each receiving 1-5 points overall. According to CEP guidelines, which mandate that a passable cut score be set at 70%, the lowest possible passing score would be an 11 out of 15 (73%). A 68% confidence interval (±1 SEM) was determined, and it was found that the lowest possible score would be 10.21 (rounded up to 11). A 95% confidence interval (±2 SEMs) was also determined, and it was found that the lowest possible score would be 9.92 (rounded up to 10). Interestingly, this low standard of error, if used in conjunction with a 68% confidence interval, would produce the same cut score that is expected by the CEP. This means that even though Cronbach's alpha predicted a relatively moderate to low reliability, the test was still sufficient in terms of mirroring the passable score requirements outlined by the CEP.

3. Inter-Rater Reliability

Inter-rater reliability was calculated to identify the agreement between two raters on the speaking task. A Spearman rank-order correlation was used for the areas of grammar control,

organizational control, and conversational control, as these variables are all on an ordinal scale. The finding, displayed in Table 14, showed that the inter-rater reliability was .38 for grammatical control, .55** for organizational control, and .79** for conversational control. Of these variables, all but grammatical control were statistically significant to the .01 level, indicating that the correlations were most likely not due to chance.

Table 14: *Inter-rater Reliability for Individual Constructs*

	Rater 2 Grammatical Control	**Rater 2 Organizational Control**	**Rater 2 Conversational Control**
Rater 1 Grammatical Control	.38		
Rater 1 Organizational Control		.55**	
Rater 1 Conversational Control			.79**

** Correlation is significant at the .01 level.

The high correlation between the raters in the area of conversational control is easily understood. During the norming session, the raters talked extensively about the expectations for this area, particularly because there was uncertainty about what to do in the event that a student did not use one of the conversational discourse markers in their discussion group. Furthermore, the use of these markers can be scored in a fairly objective manner, because the speakers tend to fall into clear categories: not using the marker, using the marker incorrectly, or using the marker correctly. This made conversational control fairly easy to score in a straightforward manner for both of the raters.

One reason why grammatical control may have had such a contrastingly low correlation between the raters is that this area was divided into two sub-constructs: accuracy and complexity. It was later discovered that one rater was focusing more on accuracy while the other was focusing more on complexity when determining the rankings. Rater 2 had a tendency to score higher because she focused on the relative fewness of errors in students' oral production, while Rater 1 was more interested in the use of complex sentences rather than simple sentences. This could be due to the fact that Rater 2 is also the teacher for this class, and therefore frequently employs error correction while teaching her students; she believes this tendency transferred from the classroom to her rating practices. This could have caused her to be more apt to award higher points to students who make fewer errors, regardless of the complexity of their sentences.

A moderate correlation occurred with organizational control, which was also broken into three sub-constructs: logical development of ideas, use of logical connectors and cohesive devices, and expression of fluency. This is likely due to the subjectivity of the rubric in this area, and heightened by the number of sub-constructs. More time could have been spent norming the rubric in this area, and clearer explanations of these sub-constructs could have been developed.

Finally, a Pearson product-moment correlation was used to determine the overall inter-rater reliability for speaking, as the variables used were interval variables. This was statistically significant at .64**, as shown in Table 15. This highly moderate result is due to having differences in agreement across the three categories, especially in the area of grammatical control.

Table 15: *Inter-rater Reliability for Total Speaking Scores*

	Rater 1 Avg	**Rater 2 Avg**
Rater 1 Avg	1.00	
Rater 2 Avg	.64**	1.00

** Correlation is significant at the .01 level.

4. Evidence of Construct Validity within the Extended-production Task

Evidence of construct validity for the extended-production task was determined through the production of a correlational matrix, which examined the extent to which the variables comprising speaking ability as defined in this test (grammatical control, organizational control, and conversational control) were related. This was performed through a Pearson product-moment correlation. Results are displayed in Table 16.

Table 16: *Correlation Matrix for the Speaking Task*

	Grammatical Control Avg	**Organizational Control Avg**	**Conversational Control Avg**
Grammatical Control Avg	(1.00)		
Organizational Control Avg	.70**	(1.00)	
Conversational Control Avg	.33	.30	(1.00)

** Correlation is significant at the 0.01 level

It was found that conversational control has a low relationship with the other two variables, organizational control and conversational control. On the other hand, there is a highly moderate, as well as statistically significant relationship between grammatical control and organizational control. This is likely because of the strong underlying theoretical connection between grammar and organization; an increase in one of these skills might be reflected through an increase in the other. Coherence and cohesion are oftentimes difficult to tease apart; this is evidenced by the fact that grammatical structures often add organizational connection to speech. Furthermore, grammatical errors may disrupt the fluency or flow of the speaker. Conversely, conversational control, as defined through the rubric as the ability to cooperatively engage in the

conversation as well as to appropriately use a discussion marker, is a skill that is more easily separated, or independent of the other two variables. This is because one can be a competent-sounding speaker but still lack the interactional competence required to be a successful conversationalist.

A perfect example of this can be illustrated with the speaking score of test taker 9. This student scored highly in both grammatical and organizational control due to his fluency, complexity, and well-organized responses. However, his score in conversational control was quite low in comparison because the raters found his interruptions of other group members and the content of some of his clarification requests (i.e. "What do you mean?") as well as his overall tone of speech to be rude pragmatically. Thus the two raters were easily able to separate conversational control from grammatical and organizational control, awarding low scores in the former and high scores in the latter.

C. Other Evidence of Validity

1. Relationships between the Two Parts of the Test

It was predicted that there would be a high and strong correlation between listening and speaking abilities on the test due to the theoretical interdependency of these skills. However, after running a Pearson product-moment correlation, a weak correlation of .16 was found between the two tasks. This is unexpected, as listening and speaking should theoretically relate. This finding could be due to the differing abilities of the test-takers in these areas, or to the differing operationalization of these tasks on the test, such as only having one speaking task, but 15 MC questions for the listening section. However, this result probably is due to the low reliability and dubious construction of the listening test.

Another potential cause for this finding is that the construct of speaking ability included a

pragmatic component – conversational control – which was initially not explored when hypothesizing about the relationships between the two areas of the test. On the test, conversational control was rather narrowly defined, and only examined the use of one discourse marker. This allowed for construct under-representation with regard to pragmatics. From a theoretical perspective, it is quite possible that a definition of speaking ability that includes a pragmatic component could be correlated with listening ability, especially because both listening ability and pragmatic skills require the use of inference. That is, when processing information, listeners have to deduce the underlying implicatures of utterances; similarly, when speaking in a pragmatically appropriate way, speakers have to produce utterances with implicatures that are contextually, socioculturally and sociolinguistically acceptable. Ideally, on future iterations of the test, the operationalization of this construct would be expanded to look at more instances of pragmatically appropriate utterances made by the test-takers in order to address these issues.

Table 17: *Correlation Matrix for the Speaking Task and Listening Task*

	Listening Total	**Speaking Total**
Listening Total	(1.00)	
Speaking Total	.16	(1.00)

2. Relationships between a background variable and performance

By the end of the session, the teachers of this class had noticed an increasing number of tardies and absences from their students, which appeared to coincide with lower grades on tests. A question was raised: is there a connection between the total number of tardies and absences and the performance of students on the test overall?

Nominal variables were created for both total number of absences and total number of tardies for each student throughout the first four weeks of class prior to the test. It was decided

that the tardies and absences should be assessed as separate variables due to the fact that they might have differing impacts on learning. The teacher's assumption was that if a student actually missed an entire two hours of instruction, they would perform worse on the test than those students who were rarely or never tardy.

The number of tardies and absences were correlated with the test takers' total scores (listening and speaking tasks) using a Pearson product-moment correlation. Findings can be found in Table 18. Unsurprisingly, there is a negative, weak correlation (-.28) between the number of absences and the total score. This is likely because the students missed important material that was taught in preparation for the test. For example, one student was absent the week prior to the test while she was on vacation, and therefore missed the majority of the unit's content. She was one of the lowest scorers on the test, with a score of seven out of 15 overall. This reveals the importance of attendance for students. On the other hand, because this correlation is not statistically significant it is entirely possible that these results are due to chance.

Table 18: *Correlation Matrix for Tardies, Absences, and Total Scores*

	Tardies	**Absences**	**Total Score**
Tardies	(1.00)		
Absences	-.30	(1.00)	
Total Score	.12	-.28	(1.00)

There was also a negative, weak correlation (-.30) between tardies and absences overall, which may indicate that students are typically either absent or late, but typically not both. For example, one student was late three times but never absent; another student was absent three times but never late. This is likely due to both the specific causes for tardiness and absence as well as the students' attitudes about attendance in general. A chronically tardy student might live

far away, have other responsibilities outside of class, or may not place large importance on class attendance, to name a few of the potential, but limitless possibilities. However, because this correlation is also not statistically significant, the relationship could be due to chance.

The weak and low positive correlation between tardies and total score (.12) was initially puzzling. To determine the cause of this result, a report of the mean scores based on the background variable of tardiness was produced, and is shown in Table 19. These results show that typically students who had zero tardies performed better (a score of 11.7) compared with those who had one or two tardies (10.6 and 11.5); however, the remaining three students who were tardy three or four times had a higher average test score (13 and 12). First, the low number of students overall in these categories contributes to their exceptionalism. Second, when examining the students in question specifically, it was found that all three of these late students are visiting scholars at Columbia. This would indicate that not only are they more likely to have a higher initial language ability prior to taking this class, but also that they may have been tardy because they were not placing as much emphasis on this English program as they were onto their other academic responsibilities. Of course, this analysis is limited in that other visiting scholars in the class were not habitually tardy, revealing that perhaps making such sweeping claims about students is problematic. Furthermore, it is impossible to know the strength of other confounding variables in producing this type of analysis.

Table 19: *Mean scores and total number of tardies*

Tardies	Mean	N
0	11.7	10
1	10.6	5
2	11.5	4
3	13	2
4	12	1
Total	**11.5**	**22**

Ultimately, it is important to note that these correlations are non-predictive, meaning that a student could not expect to have a higher score simply because they were tardy 3 times. However, the general tendency for absences and score to correlate negatively does draw attention to the fact that instruction and being in class could make a difference. It could be that the teachers did their job; the material being tested was the same material covered in class.

4. DISCUSSION AND CONCLUSIONS

The purpose of this paper was to create, administer, and evaluate the efficacy of both listening and speaking tasks for a classroom achievement test at the CEP. In order to understand the nature of listening and speaking ability in the test, theoretical models were selected and modified to suit the objectives and goals of the classroom context, as defined by the CLP, the textbook, the curriculum, and the demands of the TLU domain. These theoretical models were operationalized by defining variables and measuring them through three dichotomously-scored, multiple-choice listening tasks, and one extend-production speaking task that was scored through the use of a rubric. In order to determine the extent to which raters were consistent when rating

the unit test, inter-rater reliability was measured. Next, correlations were determined between variables within sections of the test as well as across the listening and speaking sections of the test overall. Finally, correlations between background variables – in this case, absences and tardies – and the total test score were examined in order to shed light on their potential impact on test performance.

The goal of the test was to find out whether or not the students had mastered the listening and speaking skills taught in Unit 2. It is hard to determine whether or not the students mastered the listening skills, as the reliability of the listening test was below an acceptable level for classroom assessments, meaning there was a fair amount of construct-irrelevant variance. Additionally, this reliability does not allow for the researchers to generalize findings to other comparable audiences. That is, if an entirely different group of students of the same target population were to take the test, the results would most likely be different. Even after removing four items and re-keying one answer, the increase in reliability was minimal.

Despite the reliability of the listening section, the multiple-choice items appeared to exhibit some evidence of construct validity and worked in a relatively predictable fashion. The highest degree of correlation occurred between the two inferencing item types: endophoric-implied and exophoric-implied items. Given that both require inferencing skills, this was an expected and desirable outcome. However, ideally the magnitude of the correlations would have been higher and would have exhibited statistical significance in all three categories – and not just one – in future tests. Because this is a pilot test, these items would ideally be revised and improved for future use in order to better exhibit evidence of validity.

Nonetheless, from an instructor's point of view, the results of the listening task provided useful formative information. The descriptive statistics revealed that there was little variance in

the answering of endophoric-literal items, which the test-takers tended to answer correctly; this is evidenced by a strong leptokurtic distribution and negative skewness for this item type. This means that the skill required to answer these questions – basic identification of literal details – is one that is relatively easy for learners of this level in comparison to their inferencing skills. The inferencing questions, contrarily, in particular the endophoric-implied items, were considerably more challenging for the test-takers. The implication for teachers of advanced learners is clear: inferencing skills, especially for listening, should be given instructional priority. Another implication, from the perspective of test construction, is that the test-takers ought to make the endophoric-literal items more challenging in order to match the level of difficulty of the inferencing items.

The speaking test also seemed fairly aligned with the construct; however, one anomalistic tendency was found with regard to conversational control, which seemed at odds with the other two constructs in terms of defining speaking ability. While grammatical control and organizational control exhibited a strong and statistically significant correlation, conversational control seemed to be measuring another, unrelated construct. The original intent was to try to construct a theoretical model of speaking ability that included a pragmatic component, as pragmatics is an essential component of being a competent conversationalist. Furthermore, the textbook itself included the objective of ―managing discussion skills,‖ which calls for an integration of both speaking and pragmatic skills. However, the way that conversational control was operationalized on the rubric was extremely narrow, only requiring students to utilize one discourse marker in an appropriate fashion, a mere fragment of the entity comprising pragmatic competence. One solution could have been to expand the definition of the construct of conversational control in order to make it more general. This would have required looking at the

pragmatic ability of the learners as demonstrated through all of their spoken responses, rather than the use of one marker. Alternatively, because of the difference in the way the speaking constructs were operationalized, it might have been better to construct a separate test for pragmatic competence, rather than try to fit it into a theoretical model for speaking ability.

Furthermore, while it was difficult to determine whether the listening section of the test accurately measured student abilities due to low reliability, the speaking section, with a higher reliability of .66, was ripe for interpretation. The scores genuinely seemed to reflect the variance among levels of speakers within the classroom; the teachers had expressed that there was a wide range of proficiency within speaking ability in the classroom, and this test seemed quite apt at capturing these differences. Inter-rater reliability of the speaking section, although generally quite acceptable, revealed a lower degree of concordance with regard to grammatical control. In addition to adjusting the operationalization of conversational control, this area could also be improved through a clearer rubric that explicitly details the components of this subconstruct. With some minor adjustment to the rubric – a clearer definition of grammatical control and a more general description of conversational control – this task could be quite successful in future iterations.

Unfortunately, there was no evidence that listening and speaking ability was related on this test, though the researchers strongly believe that there is a theoretical connection between these two skills. This was an unexpected outcome that may be attributable to the differences in the ways that these tasks were operationalized. Additionally, it is quite possible that the inclusion of the pragmatic component of conversational control could have led to murkiness in terms of deciphering relationships between these two areas. Because pragmatic skills are so distinct – that is, it is quite possible for a learner to have both high speaking and listening skills and a low

pragmatic competence – it is quite possible that it prevented the researchers from finding a correlation.

In addition to measuring and analyzing listening and speaking skills, the researchers were also interested in understanding how two background variables – the accumulation of tardies and absences in class – affected student performance on the test overall. Teachers may find themselves wondering: to what extent does the material I cover in the classroom matter when compared to the inherent abilities of the students in the class? Teachers may see absences and tardies as a lack of motivation; yet, no significant correlation was found on this test to indicate that these factors had an impact on performance. However, there was some evidence that the more absences a student acquired, the lower their performance was on the test. This could be due to error, of course, and ultimately the impact of this variable ought to be considered on a case-by-case basis.

While much was learned during the process of creating, administering, and evaluating this test, a variety of limitations mitigated its interpretability and generalizability. First, there was the issue of a low reliability of the test overall, which, as discussed, influences other areas such as validity and understanding of the background variable. Second, the multiple-choice test could have benefited by including more items. This is especially true because four items were removed from the test during analysis, leaving only eleven items with the tremendous undertaking of representing the construct of listening ability. Third, practical constraints also limited the ability include additional items, especially because the teachers are expected to test all four skills every time the students have a unit test. This means the test is already going to be somewhat lengthy, and teachers may not want to spend an entire class period administering a test. Fourth, the listening test in particular was affected by the researcher's inexperience in writing multiple-

choice items and constructing listening scripts. In the future, the researchers would like to implement more consistency in terms of the way that the listening scripts are produced. In this case, one of the three scripts was completely impromptu, while the other two were scripted, recorded, and re-recorded in order to include other speakers. Finally, the lack of a speaking test at the CEP also created a very imbalanced classroom in terms of skill level. The students in the class seemed to have incredibly disparate speaking abilities, which requires both the teacher and the test to have a wider range of content coverage. This limits the ability to see the fine discrepancies between the students.

As a pilot test, much was learned. In order to improve the reliability and validity of the test for future iterations, changes to the difficulty level and wording of the items, the clarity and coverage of the rubric, and the consistency and performance of the listening scripts would be made. These improvements would allow for a better operationalization of the constructs, which, ultimately, would shed more light onto the true nature of the listening and speaking ability of advanced L2 learners.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bozorgian, H. (2012). Listening skill requires a further look into second/foreign language learning. *ISRN Education*.

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

Clark, E., & Hecht, J. (1983). Comprehension, production, and language acquisition. *Annual Review of Psychology, 34*, 325-49.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Tonawanda, New York: Multilingual Matters.

Daise, D. (2003). *In Charge 2*. White Plains, NY: Pearson Education, Inc.

Feyten, C. M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal, 75*(2), 173-180.

Grabowski, K. (2014). "Personal Communication."

Grohtjahn, R. (1987). On the methodological basis of introspective methods. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (54-82). Clevedon, England: Multilingual Matters.

Halliday, M.A.K, & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.

Kim, A-Y. (2009). Investigating Second Language Reading Components: Reading for Different Types of Meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *9*(2).

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Menenti, L., Gierhan, S. M., Segaert, K., & Hagoort, P. (2011). Shared language overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological science, 22*(9), 1173-1182.

Murphy, J. M. (1991). Oral Communication in TESOL: Integrating Speaking, Listening, and Pronunciation. *TESOL Quarterly, 25*(1)*,* 51-75.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in Second Language Acquisition*, *15*, 165-179.

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

Appendix A

The Test

Listening Task 1: "The Sleepwalker"

A. The following is a conversation about sleepwalking between two friends, Katie and Emily. Read the questions. Then listen to the conversation. You will hear the passage one time.

B. Circle the letter of the best answer.

1. When Katie woke up while sleepwalking, she felt _____.
a. hurt
b. tired
c. scared (Endophoric Literal; Identifying Detail)
d. frustrated

2. The comedian _____ while sleepwalking.
a. sat in a chair
b. talked on the radio
c. went to the hospital
d. jumped out of a window (Endophoric Literal; Identifying Detail)

3. What was the main conclusion the speakers reached?
a. It can be dangerous to sleepwalk. (Endophoric Implied; Inferring Main Idea)
b. Sleepwalking is a fascinating topic.
c. Many people experience sleepwalking.
d. People who sleepwalk have unusual experiences.

4. Emily mentioned the story about the comedian to _____
a. contrast her own life to Katie's
b. provide a more serious example (Exophoric Implied; Deriving Rhetorical Meaning)
c. explain a different sleep disorder
d. show sleepwalking also happens to famous people

5. This conversation probably took place at _____.
a. work
b. school (Exophoric Implied; Deriving Contextual Meaning)
c. an apartment
d. a coffee shop

Listening 1 Script: "The Sleepwalker"

Katie: That psychology class was so interesting.

Emily: I know! Sleep disorders are so fascinating.

Katie: They really are. You know, when I was a kid, I actually used to sleepwalk all the time.

Emily: No way!

Katie: I know, right? Very crazy. Um, I have a lot of stories, but, honestly one of my favorites was, um. My family was visiting my grandparents' house and um, I was sleeping upstairs with my brothers and sisters. My parents were sleeping downstairs, right?

Emily: Mmmhm.

Katie: So, I actually sleptwalked all the way upstairs, from upstairs to downstairs. Uh, huh, down a flight of like 20 stairs or something, in my sleep.

Emily: Oh my gosh!

Katie: I somehow managed to sit down in this chair, and was just sitting there, until my mom finds me, like she heard some noise so she came out and was like, oh my gosh, my daughter is asleep in this chair. Um, and so she kind of gently, gently woke me up. And I remember waking up and being like, SO scared because I had no idea where I was, right? Can be pretty dangerous, you know?

Emily: Yeah, totally! Um, you're really lucky I think. I was listening to the radio a while back, and this comedian, he was really funny, but he told a story about how he jumped out of a window at a motel.

Katie: Oh my gosh, was he ok?

Emily: Yeah, he had to go the hospital, but, um, he had a sleepwalking problem.

Katie: Wow.

Emily: Yeah.

Katie: So, it seems like it can be really—it could have a really negative impact on people's lives.

Emily: Definitely.

Katie: Yeah.

Listening Task 2: "Spider Mother"

A. The following is a conversation between sisters. Read the questions. Then listen to the conversation. You will hear the passage one time.

B. Circle the letter of the best answer.

6. Liz started writing in a dream journal because she _____

a. had a nightmare (Endophoric Literal; Identifying Main Idea)
b. was afraid of spiders
c. was feeling depressed
d. had a fight with her mom

7. Julia thinks that dream journals are _____.

a. ugly
b. unique
c. useless (Endophoric Implied; Inferring Detail)
d. universal

8. Liz believes that dreams represent _____**.**

a. future possibilities
d. pieces of your memory
b. your greatest hopes and fears
c. a connection between people (Endophoric Implied; Inferring Detail)

9. After this conversation, Liz will probably _____.

a. go to dinner (Exophoric Implied; Deriving Contextual Meaning)
b. fight with her mom
c. read more about spiders online
d. finish writing in her dream journal

10. Julia feels _____ about Liz' dream journal.

a. surprised
b. skeptical (Exophoric Implied; Deriving Psychological Meaning)
c. disinterested
d. disappointed

Listening 2 Script: "Spider Mother"

Julia: Come to dinner Liz, Mom's waiting!

Liz: Wait a minute Julia, I'm just finishing up my dream journal, I'll be down there in a second.

Julia: Umm, your what?

Liz: My dream journal.

Julia: Huh. *Interesting.* Umm… why are you doing that?

Liz: Well, it's because of this awful dream I had.

Julia: Uh huh… about what?

Liz: Well, you know how I had that big fight with mom last week?

Julia: Yeah.

Liz: Well… so that night I had this terrible dream where there were spiders crawling all over me. I could feel them, and when I woke up it really freaked me out, so I looked it up online and a lot of people said spiders crawling all over you indicate problems with your mother.

Julia: That's just a bunch of nonsense. You probably just saw a spider before you went to bed and then your brain thought about it while you were asleep.

Liz: No, I read about it and on tons of blogs and websites a lot of people mentioned this. I think if a lot of people have experienced this same thing, it must mean that there's some kind of deeper, more universal meaning to dreams…

Julia: No way. Whatever. You're still going to fight with mom anyway. You should be spending your time doing your homework or something more worth your time.

Liz: Well, I think if I write about it, I might actually be able to get more in touch with these feelings I'm having…

Julia: Ugh, why don't you just go read more about it online?

Liz: I'm done anyway. I'm coming, tell Mom I'll be there in a second.

Listening Task 3: "The Artist on the Radio"

A. The following is an interview heard on the radio. Read the questions. Then listen to the interview. You will hear the passage one time.

B. Circle the letter of the best answer.

11. The artist, Lucy, is trying this new technique because she _____

a. needed new inspiration (Endophoric Literal; Identifying Detail)

b. wanted to be more famous

c. wasn't getting enough sleep

d. hasn't sold enough paintings

12. When falling asleep, the artist holds a _____ in her hand.

a. stone

b. stick

c. spoon (Endophoric literal; Identifying Detail)

d. sponge

13. The main reason Lucy is being interviewed on the radio is because _____.

a. her paintings are dreamlike

b. her paintings have been selling well

c. she is trying a new technique

d. she has an art opening next week (Endophoric Implied; Inferring Main Idea)

14. Lucy uses this new technique because _____.

a. the loud sound inspires her

b. it allows her to paint her dreams (Endophoric Implied; Inferring Main Idea)

c. she paints better when she is tired

d. it is believed to increase creativity

15. The interviewer's attitude is _____

a. serious

b. persuasive

c. enthusiastic (Exophoric Implied; Deriving Psychological Meaning)

d. argumentative

Listening 3 Script: "The Artist on the Radio"

Paul: Welcome back. I'm Paul Baxter, and today we will be talking with a famous artist, Lucy Lewins, whose new painting series will be opening at the Museum of Modern Art next week.

Lucy: Hi, it's great to be here, thanks for having me, Paul.

Paul: So, Lucy, I've been wondering about the inspiration behind some of your latest and greatest works featuring a variety of unusual, very dreamlike imagery.

Lucy: Oh, you're too kind Paul.

Paul: Could you elaborate a little about your inspiration?

Lucy: Well, I've been trying out this new technique because I was feeling really uninspired and I wanted to bring some new life into my work. I know this sounds crazy, but I heard about it when reading about one of my favorite artists, the famous Salvadore Dali. So, here's what he would do. He would sit in a chair and place a tin plate on the floor next to him…

Paul: Oh!!!

Lucy: And he'd hold a spoon in his hand, above the plate… and he'd just sit there for a long time, and slowly, eventually he would start to fall asleep.

Paul: Okay, wow!!

Lucy: And then, when he would finally relax and fall asleep, the spoon would fall and clash with the plate, making this really, really loud sound, waking him up. And then… then there were these dreamlike images still fresh in his mind, so that he could begin painting right away.

Paul: Amazing, so have you been trying this?

Lucy: Yeah. It's amazing how you can really bring unconscious thoughts to conscious level, and if I get to work straight away, I'll be able to bring that into my painting too.

Speaking Prompt

With your group, discuss the following topic.

What are some strategies you could use to achieve your personal dreams? Do you agree or disagree with the other students in your group?

Every speaker in the group should make an effort to use **at least one** of these expressions during the discussion **at an appropriate time and in a natural way.**

Taking Turns
1. We still haven't heard your opinion.
2. You have the floor.
3. And now let's hear from _____
4. Perhaps _____ has something to add.

Staying on Track
1. Sorry to cut your story short, but we need to finish.
2. If we could all get back to the topic at hand…
3. Could we stick to the main point?

Clarifying Miscommunications
1. Let me rephrase what I think you said.
2. What I meant to say was...
3. If I understand you correctly, you're saying that _____ .
4. I think I missed something in our conversation here.
5. Are you telling us that _____ ?
6. Could we stop and recap? I'm a little confused.

You will have three minutes to think about the topic and prepare. Then, you will discuss the topic for five minutes with your group.

Please remember to take turns and share the floor

Speaking Rubric (adapted from CEP Placement rubric)

Name: _____

	Full Evidence 5	**Good Evidence 4**	**Adequate Evidence 3**	**Marginal Evidence 2**	**Limited Evidence 1**	**No Evidence 0**
Grammatical Knowledge Refers to the extent to which language displays **accuracy –** accurate use of grammar and vocabulary. **complexity** – use of a variety of sentence types (simple and complex, varied lengths and structures, coordination and subordination) to achieve desired effect.	Language provides **full** evidence of accuracy (except for rare minor errors, grammar and vocabulary are **precise**) Language provides **full** evidence of balanced sentence types.	Language provides **good** evidence of accuracy (some minor errors, meaning is seldom obscured) Language provides **good** evidence of balanced sentence types.	Language provides **adequate** evidence of accuracy (some errors, meaning is generally not obscured) Language provides **adequate** evidence of balanced sentence types (may contain more of one type).	Language provides **marginal** evidence of accuracy (many errors, meaning sometimes obscured) Language provides **marginal** evidence of balanced sentence types (predominance of some sentence types).	Language provides **limited** evidence of accuracy (errors frequently obscure meaning) Language provides **limited** evidence of balanced sentence patterns (relies on simple/short sentences).	Language provides **no** evidence of accuracy (sentences are often incomprehensible) Language provides **no** evidence of balanced sentence patterns
Organizational Control The extent to which the response is organized in a coherent manner, an cohesion between sentences and idea is successfully achieved through the use of logical connectors and cohesive devices. There is an absence of excessive pausing or hesitation, and speech is generally well-connected and fluent.	Demonstrates **full** evidence of logical development of ideas. Shows **full** evidence of sophisticated use of logical connectors and other cohesive devices. **Full** expression of fluency.	Demonstrates **good** evidence of logical development of ideas. Shows **good** evidence of sophisticated use of logical connectors and other cohesive devices. **Good** expression of fluency.	Demonstrates **adequate** evidence of logical development of ideas. Shows **adequate** evidence of sophisticated use of logical connectors and other cohesive devices. **Adequate** expression of fluency.	Demonstrates **marginal** evidence of logical development of ideas. Shows **marginal** evidence of sophisticated use of logical connectors and other cohesive devices. **Marginal** expression of fluency.	Demonstrates **limited** evidence of logical development of ideas. Shows **limited** evidence of sophisticated use of logical connectors and other cohesive devices. **Limited** expression of fluency.	Demonstrates **no** evidence of logical development of ideas. Displays virtually **no** organization. Does not contain enough evidence to evaluate.
Conversational Control Ability to manage conversation with regard to openings, closings, turn-taking, and other interactional practices. **appropriateness** – sociolinguistically and socioculturally appropriate use of given discussion markers from Unit 2 in the conversational setting.	Demonstrates **full** ability to cooperatively engage in the conversation. **Full** evidence of appropriate use of discussion marker in the conversation.	Demonstrates **good** ability to cooperatively engage in the conversation. **Good** evidence of appropriate use of discussion marker in the conversation.	Demonstrates **adequate** ability to cooperatively engage in the conversation. **Adequate** evidence of appropriate use of discussion marker in the conversation.	Demonstrates **marginal** ability to cooperatively engage in the conversation. **Marginal** evidence of appropriate use of discussion marker in the conversation.	Demonstrates **limited** ability to cooperatively engage in the conversation. **Limited** evidence of appropriate use of discussion marker in the conversation.	Demonstrates virtually **no** ability to cooperatively engage in the conversation. Does not contain enough evidence to evaluate.

ADDITIONAL COMMENTS:

Appendix B

Test Task Specifications

	Speaking Task	Listening Task
INPUT		
Format		
Channel	Visual	Aural, Visual
Form	Language	Language
Type	Prompt Short (sentences)	Passage Extended Discourse
Length		Passage 1: 282 words Passage 2: 278 words Passage 3: 281 words
Vehicle	Printed Prompt	Recorded Passages Printed Questions
Language Characteristics	n.a.	1. Endophoric Literal Meaning a. Main Idea b. Detail 2. Endophoric Implied Meaning c. Main Idea d. Detail 3. Exophoric Implied Meaning e. Contextual meaning f. Psychological meaning g. Rhetorical meaning
Topical characteristics	"World of Dreams" Theme What are some strategies you could use to achieve your personal dreams?	"World of Dreams" Theme 1. Sleepwalking 2. Dream Journals 3. Painting from Dreams
EXPECTED RESPONSE		
Format		
Channel	Visual	Visual
Form	Language	Non-Language
Type	Extended-production	Selected-response
Length	Extended Discourse	n.a.
Language Characteristics	1. Language Control a. Accuracy b. Complexity 2. Organizational Control a. Development of Ideas b. Cohesive Devices c. Fluency 3. Conversational Control a. Cooperative Engagement in Conversation b. Appropriate Use of Conversation Marker	n.a.
Reactivity	Reciprocal	Non-reciprocal
Scope of relationship	Broad	Narrow
Directness of relationship	Direct	Indirect