Constructing and Piloting a Vocabulary Test

A DOUBLE PLAN B PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

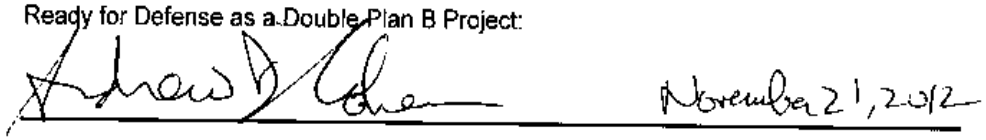
BY

Stephen Kis

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS IN ENGLISH AS A SECOND LANGUAGE

December 2012

Ready for Defense as a Double Plan B Project:



November 21, 2012

Project Supervisor

Date

Table of Contents

Constructing and Piloting a Vocabulary Test

1.0 Introduction

The purpose of this paper is to present and discuss the construction, implementation, and evaluation of a language assessment instrument that can, generally speaking, be called a cumulative vocabulary achievement test. The test was designed for students in a speaking and listening ESL class at the University of Minnesota. This paper will begin by considering the recommendations of experts on vocabulary in relation to the process of test construction. Next, it will describe the testing situation in detail and the process of constructing the test. After that, the paper will discuss the process of piloting and revisions to the pilot. Lastly, the paper will discuss the final administration and provide a statistical analysis of the test results.

2.0 Literature Review

Before discussing the vocabulary test that is the focus of this study, this paper will consider the following five questions that can serve as a useful guide for a vocabulary test construction project.

1. What is vocabulary knowledge?

2. What is the purpose of the test?

3. What words should be on the test?

4. What format should be used for the test items?

5. What recommendations have been made for piloting vocabulary tests?

2.1 What is Vocabulary Knowledge?

Before discussing the concept, "vocabulary knowledge," it is worth considering the definition of vocabulary itself. As Read (2012a) has pointed out, many peoples' understanding of the term *vocabulary* "is very much influenced by . . . experience with dictionaries, which present words as independent semantic units, each with its own entry" (p. 257). However, this definition of vocabulary should be more inclusive, in Read's opinion, in order to include multiple

words in a family (write, writes, writing, wrote) and not just the headword (write), lexical units consisting of more than a single word, like compound nouns (bank account, speed limit), and idioms (kick the bucket, barking up the wrong tree). The tendency to focus on vocabulary as only single word units is especially evident in commercial vocabulary tests and activities in textbooks, and may be due to the fact that creating multi-unit vocabulary items is more difficult.

In addition to an oversimplified definition of vocabulary, most classroom instruction and tests on vocabulary reveal a limited understanding of *word knowledge*. Meaning is often the aspect of vocabulary that receives the most attention and seems to be valued the most highly. However, the definition of vocabulary knowledge can be quite extensive. In her book *Word knowledge: A vocabulary teacher's handbook*, author Cheryl Boyd Zimmerman maintains that meaning is only one of five aspects of word knowledge alongside knowledge of collocations, grammatical features, word parts, and appropriateness. These five aspects are broken down further into layers. Layers of meaning include knowledge of "positive and negative connotation" and "degree ('strength' of a word)"; collocations include the layers "fixed phrases" and "preposition use"; grammatical knowledge contains four layers, "passive and active verbs," "verb complements," "count and uncountable nouns," and "parts of speech"; word parts include the layers "right meaning, wrong suffix" and "word building gone awry"; and, lastly, register and appropriate forms include the layers "formal/informal," "polite or people sensitive/impolite," and "direct/euphemistic" (Zimmerman, 2009, pp. 5-6).

How a test constructer perceives the terms vocabulary and word knowledge will ultimately contribute to the vocabulary assessment constructed. If the definitions of these terms are as inclusive as those by Zimmerman and Read, then it should be clear to the test constructor that a great many vocabulary items could potentially be created and many aspects of vocabulary assessed.

2.2 What Is the Purpose of the Test?

Construction of a vocabulary assessment measure should begin by first identifying the purpose of the assessment. Three common kinds of language tests, which are not exclusive to vocabulary assessment, are the most likely ones to be developed for classroom purposes and are all probably familiar to most teachers to some extent. These three purposes for tests are diagnostic, placement, and achievement. The purpose of a vocabulary diagnostic test is to find the gaps in students' vocabulary knowledge so that the necessary instruction can be provided. The purpose of a placement test is to place students in a specific course level according to their vocabulary knowledge. Two kinds of achievement tests have been identified by Nation, both of which share similar purposes. The first is a short-term achievement test, where the aim is to see

whether a recently studied group of words has been learned. The second is a long-term achievement test, which aims to see whether a course has been successful in teaching particular words. Additional purposes of both kinds of achievement tests may also be to test how well learners can use vocabulary they have learned in the classroom in conjunction with any of the four language skills (reading, writing, speaking, and listening), as well as purposes related to their washback effect, such as motivating students to study, and making selected words more clear through testing (Nation, 2001).

In addition to being familiar with the common testing purposes above, differentiating between two kinds of vocabulary knowledge, *depth* of vocabulary knowledge and *breadth* of vocabulary knowledge, would also be useful before beginning the construction of a test. Looking at how well a particular word is known is called measuring depth of knowledge. This is contrasted with measuring breadth of vocabulary knowledge, which would be more concerned with measuring how many words are known (Nation, 2001).

If a test constructor is interested in measuring depth of knowledge to any extent, then it would be wise to consider the facets of vocabulary knowledge presented by Read and Zimmerman, which are all aspects of a word that can be learned and would therefore be measures of depth. The extent to which a test constructer wishes to measure depth of vocabulary knowledge is a challenging decision that will inevitably have to be made, and as Read mentions "there is no agreement on what the key components of depth of vocabulary are and no standard test of depth" (Read, 2012a, p. 259). While the quotation from Read makes clear that this can be a daunting decision, a considerable number of recommendations *have* been presented in the literature which will be discussed later in this paper. What is evident is that the extent to which an assessment measures depth of vocabulary knowledge must be limited to some extent because testing all aspects of each word on a test is likely to be impractical.

Testing breadth of vocabulary knowledge may be a purpose in and of itself or, in relation to classroom teaching, could be used for placement decisions. Recommendations for constructing tests whose primary aim is to assess breadth of vocabulary knowledge are discussed in Nation (2001), but are beyond the scope of this paper. However, even if measuring breadth of vocabulary knowledge is not the focus of a test, the test constructor will still likely need to consider it to some extent since it is often the case that a tester would like to test more words than would be practical. Therefore, some recommendations with regard to breadth of knowledge will be considered later in this paper regarding achievement tests such as the one constructed in this study.

After considering the purpose of a test, the test constructer will be in a better position to answer the questions that follow.

2.3 What Words Should Be on the Test?

Since the primary focus of this paper is to discuss the construction and implementation of a vocabulary achievement test and since achievement tests are so closely connected to classroom instruction, this section will begin with a few recommendations for teaching vocabulary, that when followed will logically lead to an assessment that is more beneficial in terms of the washback it yields to the students.

Nation (2011), stated that "it is unusual to find teachers who have a well-supported idea of what vocabulary their learners already know. As a result, language courses often contain a mixture of useful vocabulary and vocabulary that by no means represents the best choice for those learners" (p. 530). Nation's observation highlights the necessity to use needs analysis as a first step in selecting the vocabulary that should be covered in a language course in order for it to be useful for the students. Once the teacher has some of idea of what vocabulary is likely to be useful to students, an additional step would be to give students a vocabulary pretest in order to determine what knowledge students possess about a given pool of words which might be generated by the teacher, or more likely taken from a textbook, and determine whether or not all of those words require further instruction, and how much instruction, based on the results of the pretest.

Additionally, *frequency* has been identified as a key criterion for selecting words for teaching, as well as testing. For frequency, Read groups words into three categories: "high-frequency," "low-frequency," and "specialized vocabulary" (2002) (which includes academic vocabulary). The type of vocabulary chosen for both instruction and assessment will inevitably be closely related to the purpose of the class or test, but generally high-frequency vocabulary will have a greater value for learners. Not surprisingly, for classroom instruction, Zimmerman recommends that teachers "target high-frequency words because they will naturally occur inside and outside the classroom, providing the repeated exposures that learners need" (Zimmerman, 2009, p. 7). However, she also emphasizes the importance of teaching academic words to learners who intend to pursue academic study in English. Targeting academic words has become more simplified since the creation of the Academic Word List (AWL). The AWL has been a well-applied resource for textbook designers in recent years and should also be used as a resource by teachers who wish to target academic vocabulary.

The words selected for a vocabulary assessment are also likely to require consideration of *sample rate*, because as mentioned earlier, any test is only going to include a small subsample of the words that a student could or should potentially be tested on. Therefore, a test constructor needs to select a manageable number of words from a pool. Schmitt defines sample rate as the number of vocabulary words chosen for a test compared to the total number of possible words (Schmitt, 2000). Schmitt also raises issues of how sample rate influences the *validity, reliability,* and *practicality* of a test. Higher sample rates and greater numbers of words increase the quality of a test because the test will be more representative of a learner's total knowledge of the vocabulary used in the assessment. This raises the issue of

validity, which should potentially be greater with a higher sample rate. Reliability, the consistency of a test's behavior over time, can also be improved with a higher sampling rate. However, one disadvantage of a higher sample rate is a decrease in the practicality of a test, because a higher sample rate means more words on the test, which also means more time to administer the test, more time to score it, and so on. To compensate for the loss of practicality that follows from a higher sample rate, however, Schmitt recommends choosing "an item format that can be answered quickly, allowing for the maximum number of items to be placed on a test" (Schmitt, 2000, p. 167). Additional advice regarding sample rate and achievement testing is also offered by Nation (2001), who recommends taking a certain number of words from each week of lessons or each unit so that a test will represent all the words studied in a reasonable way (pp. 375-378).

2.4 What Format Should Be Used for the Test Items?

In order to determine which test items should be used for a vocabulary test, the test constructor should first be familiar with a few dichotomous terms that have been frequently used to describe items in the testing literature. This section will begin with a discussion of the *receptive/productive*, *recognition/recall*, *partial/imprecise*, and *embedded/discrete* distinctions, and conclude with some general recommendations on choosing item types and how to construct common item types.

According to Schmitt (2000), test constructors need to consider the mode of a test (listening, reading, speaking, writing). Schmitt also notes that most vocabulary tests are in the written mode. The mode or modes in which the vocabulary is presented will ultimately affect the validity of a test and in some cases it may be desirable to test vocabulary along with other skills. However, upon evaluation the assessor must consider *all* the skills being utilized for completing an item when interpreting the test results (Schmitt, 2000). The consideration of "mode" for test items is closely related to the receptive/productive distinction. According to Nation, "receptive knowledge is that used in listening and reading, and involves going from the form of a word to its meaning," while "productive knowledge is that used in speaking and writing, and involves going from the meaning to the word form" (Nation, 2001, pp. 358-359). Decisions to use either receptive or productive items for an achievement test could thus logically be made according to the language skills that were focused on during a course. In addition to being associated with different language skills, receptive items have often been interpreted as testing the ability to *comprehend* a word, while productive items have been used to determine the ability to *use* a word. If we consider the following receptive item cited in Read (2000), it is easy to see how one might associate it with testing comprehension.

loathe means A. dislike intensely
B. become seriously ill
C. search carefully
D. look very angry
(Read, 2000, p. 156)

For this item the learner is presented with a word and is required to select the correct meaning as evidence that he understands it.
The following production item, also from Read (2000), reveals an obvious association with the ability to use the word:

> Because of the snow, the football match was _____ until the following week.
> (p. 156)

For this item the learner is presented with a sentence that is meant to elicit a target word to show that the learner can use it.

Despite the fact that the terms receptive and productive may seem straightforward at a glance, these terms have been problematized by Read (2000) with regard to their application for defining a vocabulary assessment because there have been some inconsistencies in how they have been used in the testing literature. For example, if we consider the production item above, Read mentions that while researchers have equated the ability to correctly answer this type of item with use, this inference is "questionable" because this item only requires production "in a very restricted sense" (p. 156). In other words, the ability to answer the above item may show that the learner can write the word in this very controlled sentence, but we should not assume that this is evidence that the learner can use this word in other contexts or in other modes. In this sense, it seems that some researchers did not heed the warnings with regard to validity mentioned above by Schmitt. For this reason, Read, instead, prefers the terms recognition and recall. The former, represented by the first example, according to Read, "means that the test-takers are presented with the target word and are asked to show that they understand its meaning" (Read, 2000, p. 155). The latter term, represented by the second example, is described as follows: learners "are provided with some stimulus designed to elicit the target word from their memory" (Read, 2000, p. 155).

Both the receptive/productive and recognition/recall distinctions may prove useful for considering vocabulary items. However, the former should perhaps be used with caution. Additionally, these item types are closely connected with difficulty. Nation notes that when an item format is controlled receptive recall is easier than productive recall. Nation also mentions

that, "recognition items are easier because even with partial knowledge a test-taker may be able to make the right choice" (Nation, 2001, pp. 358-359)

Another distinction which can be made for vocabulary items is imprecise/precise. According to Nation, this distinction relates to the degree of accuracy required in the answer. This can be reflected in the similarity of the choices provided, the degree of prompting, and the degree of acceptance of an approximate answer (Nation, 2001). These two different item characteristics will also affect difficulty. Items allowing for imprecise knowledge are easier because credit is given for partial knowledge. Some items, such as the multiple-choice item example shown above, will logically assess precise knowledge, while others may be less precise. Let us also consider a less controlled sentence-writing item where the student is presented only with a word and asked to use it in a sentence. For this item the test scorer may have an idea of what kind of sentence shows complete understanding of the word, but may also decide to give partial credit to a student who still manages to write a sentence that shows understanding of the meaning or perhaps hints at an understanding of the meaning which is masked by grammatical mistakes.

Lastly, let us consider the embedded/discrete distinction for vocabulary items. An embedded item focuses on vocabulary as part of a broader construct, for example, when a speaking test uses vocabulary as one criterion from which the test-taker is judged. Discrete test items focus on a specific aspect of vocabulary knowledge based on a preselected set of words. For example, a traditional discrete-point item would be the multiple-choice item shown earlier where the focus is on meaning. The embedded/discrete distinction has also been referred to as context-dependent/context-independent. With regard to context, embedded items provide more context for the vocabulary word being tested while discrete items provide little if any context. Considerations for item construction regarding the discrete/embedded distinction will be clearly related to decisions about mode, and as noted by Schmitt, if a teacher desires to measure vocabulary knowledge as something separate from other language skills, then the items devised should be more discrete. As items approach the embedded end of the continuum, they will be testing other skills along with vocabulary and in turn will be assessing other skills as well (Schmitt, 2000). Additionally, Read maintains that it is unnecessary to provide context for vocabulary items, "unless the test-takers are required to engage with the contextual information in some meaningful way in making their response to the test task" (Read, 2000, pp. 162-163).

In summary, vocabulary items have often been classified along a continuum according to the four dichotomies mentioned above in literature on vocabulary assessment. These dichotomies reveal that the test constructor will need to make decisions about the mode/s used for responding to test items, how a response will be elicited, the degree of word knowledge that will be required to answer an item, and the degree to which an item should be contextualized. One major advantage of being familiar with these dichotomies is that they

should also encourage the test constructer to think more critically when interpreting the scores of the test since a more rigorous evaluation of the individual items and the constructs implicit in those items will ultimately lead to a better evaluation of the test as a whole.

In addition to the distinctions mentioned above, some more general recommendations for constructing vocabulary items appear in the literature (Nation, 2001):

1. The knowledge required to answer test items correctly should be similar to the knowledge you want to test. If the test is an achievement test, then it should reflect the knowledge taught in the course.

2. It should be easy to make items to test all of the vocabulary that the test constructer wants to test. Particularly in the case of classroom tests, teachers should not have to spend hours constructing test items that will appear on a test that will only take a short time to administer.

3. The test items should be planned carefully with consideration of how they will be scored so that when scoring takes place it is as easy as possible.

4. Test items should provide useful repetition of vocabulary and perhaps extend the learners' knowledge. Consideration should be given to the creative nature of language which involves understanding and using words in new contexts.

Additionally, some useful advice has been given for constructing two common vocabulary item types: multiple-choice and sentence-writing. Nation has noted that "there seems to be no major disadvantage in using multiple-choice except perhaps the amount of work required to make the items" (Nation, 2001, p. 350). In order to reduce the amount of work required for developing this useful item type, he suggests using the matching format. Using the matching format can significantly reduce the amount of distracters that need to be made and also allows a lot more items to be tested in the same amount of time. In addition, Read has suggested that definitions included in a matching section should be easy for learners to understand and should definitely not include words that are more difficult than the word which is to be defined. He recommends learners' dictionaries as a helpful resource for writing definitions, since they use high-frequency vocabulary. Like Nation, Read also emphasizes that just adding a few distracters to a list of possible choices for a matching section will significantly reduce the students' ability to guess the answer, and adds that distracters should be taken from the pool of words that students have studied. He also mentions that test constructors should use words which are all from the same class in order to reduce elimination of possible definitions through knowledge of parts of speech. Although not directly related to construction,

Read adds a practical suggestion for sentence-writing items by recommending that the scorer give two points for the item by allocating one mark for the meaning of the word and another for its grammar (Read, 2000, p. 176).

Lastly, when constructing items for a vocabulary assessment, as with any test construction project, the test constructer should be aware that a draft test should contain more items than are needed for the final version so that poor-performing items can be removed without having to pilot replacement items.

2.5 What Recommendations Have Been Made for Piloting Vocabulary Tests?

Several suggestions for piloting vocabulary tests are offered by Read (2012b). A brief discussion of Read's observations should be enough to provide a novice test constructor with a foundation for piloting a new vocabulary test. This section will begin by considering suggestions for developing two kinds of vocabulary test items and conclude with a brief summary of the process involved in piloting.

During the development of conventional vocabulary test items, like the multiple-choice and matching formats, the test constructor should assess the items based on the following criteria. The items should be developed so that only one clear answer is possible, and should be redesigned if another acceptable answer is found during piloting. The correct response should also be clearly expressed, and not easily guessed by being longer or noticeably more specific than the other options. Additionally, the incorrect answers (distracters) should be functioning effectively by attracting at least some of the lower-ability test-takers. Lastly, the items should be at an appropriate level of difficulty (Read, 2012b). With regard to this last point, Read does not specifically define the appropriate level of difficulty, however, an IF (item facility) between .60 and .80 has been recommended for all kinds of items on classroom tests (Cohen, 1994).

Read also makes a few recommendations for assessing the quality of gap-filling items. First, the context given for these items should be sufficient to allow the test-takers to supply an acceptable response. Second, if more than one acceptable answer can be produced for a given item, then the test-constructor should have a list of acceptable answers before the test is scored. Third, if the intention is for the sentence to elicit only one target word, then only one acceptable answer should be possible. This can be simplified by providing only the first letter of the target word. Lastly, if the words supplied by test-takers are to be judged based on grammatical correctness and spelling, then clear guidelines should be established before scoring the test (Read, 2012b).

In addition to following Read's recommendations for the above item types, having a clear idea of what the piloting of a test often entails would be useful information for any test developer. A common first step after writing items is to get feedback on the items from native

speakers or proficient users of the target language so that any clear problems with the way the items have been written can be identified early on in the process. After correcting any obvious flaws in the items, the next step of piloting is to administer the test to a group of learners who have backgrounds (proficiency, native language, etc.) which are as similar as possible to the population of test-takers who will take the final test. It should be noted that arriving at a high-quality test typically involves piloting it more than once. If the test constructor is going to look at the reliability of a pilot test, then it is also suggested that the number of learners involved in the study be as high as possible. Also, before being administered the test, students should be briefed on the instructions of the test, especially if it is an unfamiliar format, and they should be given some practice items. Additionally, it may be desirable to ask some of the students from the pilot group to engage in verbal report as they work through the items in order to gain insights about how they respond to the items. A final step that could be used in evaluating the pilot-test results would be to have some independent measure of the test-takers' knowledge of the target words, which could be accomplished by giving a separate written test in a different format or by interviewing the learners individually (Read, 2012b).

By reviewing the literature discussed in this section I learned that vocabulary and vocabulary knowledge are more complex concepts than I initially expected. I also learned that a vocabulary test should be closely connected to its purpose, and that many purposes exist for making decisions about learners based on their vocabulary knowledge. Additionally, the literature revealed that words must be selected carefully and that many factors must be considered when selecting words in order to create a high-quality vocabulary test. I also became more aware of the close relationship between the knowledge assessed by a test and the format of test items. Lastly, the recommendations for piloting a vocabulary test helped me to gain a better understanding of what this process entails and revealed additional steps for future research.

In light of these considerations, we will now look at the vocabulary test construction project.

3.0 A Description of the Test Construction Project

3.1 The Nature of the Course

As indicated above, the test designed for this project was intended for students in a speaking and listening ESL class that I was teaching at the University of Minnesota. A large amount of the material that these students studied came from a required textbook, *Contemporary Topics 3*

(Beglar & Murray, 2009). The textbook was by no means the only resource used for instruction in this class, but it contained the vocabulary that students studied throughout the semester.

The students in this course were at level 3 out of 4 in the Minnesota English Language Program (MELP). The TOEFL is typically the proficiency measure used to place the students into the MELP program, and since these students were at level 3 their TOEFL scores should have been between 477 and 520. The MELP program is an intensive one in which students typically take 5 courses per semester, which amounts to approximately 25 hours of classroom study per week. Most students in MELP intend to pursue academic study at an American university. Nearly all of the students in my class intended to pursue their studies at the University of Minnesota. The backgrounds of the students in my class were quite varied. Of the 11 students whose test data were included in the final version of the assessment, 5 were Chinese, 1 Thai, 1 Korean, 1 Belarusian, 1 French, and 2 Emiratis. The semester for these students was 14 weeks long.

Ten vocabulary words, which were related to an academic subject, were presented at the beginning of each unit in the students' textbook, and within nearly eleven weeks of study the students had covered 7 units. As a result, my students had studied 70 vocabulary words before I began creating the test. These vocabulary words were presented in the textbook in meaning-focused activities, typically multiple-choice questions, that featured the words in a sentence that contextualized the meaning of the target word before asking the students to select the correct definition. Students completed these activities outside of class before beginning each unit and then received feedback about their performance via classroom discussions of the correct answers.

Additionally, an important activity which shaped the design of this assessment was the completion of a vocabulary chart that included each of the 10 words from units studied during the course. Vocabulary charts for each unit studied included the target vocabulary and required students to: 1) indicate the part of speech for the target word along with synonyms, 2) note the syllable that received primary stress, 3) provide examples of other forms of each word, and 4) write a sentence using each word. An example from a chart students were asked to complete is shown below.

Vocabulary for Unit X				
Word and Part of Speech	Synonym	Stress	Other Forms	Example Sentence
Ex. mundane adj.	commonplace	mundane	mundanely adv.; mundaneness n.	Although many people love the movie Avatar, I thought the movie's plot was quite **mundane**.
*1. adulthood				

Note: The * indicates an example of what students would see for this activity.

Class time was used for students to complete these vocabulary charts. Before embarking on the task, students were put in groups. Each group was assigned a number of words from the chart and given a time limit to find all of the required information. After a while, they were given dictionaries and allowed to use other resources to find the necessary information. Once each group had found the information for their assigned words, they were put into new groups (using a jigsaw model for creating groups) and told to teach their words to their new group members. If class time was running short, instead of asking the students to teach the words, a chart was drawn on the whiteboard and students were called on to provide the missing information. In both cases feedback was provided by the teacher and mistakes in the charts were corrected.

Students were instructed to keep their vocabulary charts for self-study and were told that they would be tested on this information near the end of the course. Vocabulary learning in the course was not an end in itself, but was necessary to enable the students to understand the academic lectures and other listening texts presented in the textbook and also to assist the learners in discussions related to the lectures. Since the vocabulary students studied was frequently recycled in other activities in the textbook, students encountered the vocabulary several times in various contexts and through both receptive modes. The students' grades for this course were based largely on tests which required them to listen to and answer comprehension questions about the lectures for each unit, as well as three presentations related to the themes in the units studied.

3.2 A Description of the Pilot Test

From the description of the way vocabulary was covered in the course, especially the vocabulary chart activity, it should be clear that vocabulary instruction was greatly influenced by Zimmerman's definition of word knowledge in the sense that multiple facets of words were

studied (see 2.1). Although I was uncertain as to the kind of assessment our classroom activities would lead to when I began teaching the course, I was very interested in developing students' depth of vocabulary knowledge. In order to provide some evidence of students' progress in the course and to give them feedback on how well and to what extent they learned the vocabulary that was taught, I decided to develop the vocabulary test that is the subject of this paper.

This vocabulary test can be defined according to Nation's definition as a *long-term achievement test* (see 2.2). An additional purpose of this test was that it would also be used for promotion, in the sense that it would be one factor in determining the students' total grade for the course. The vocabulary test would count for 10% of the students' total grade, and since they would need to get at least a 70% in the course to pass the class and move on to the following level, the students should have been very motivated to study for it. Lastly, I considered that the test might also be used to help me evaluate how well I taught the vocabulary for this course, particularly through use of the vocabulary chart activity.

Choosing the words for this vocabulary test was simplified to some extent because the words that the students studied in class appeared in a required textbook and the primary purpose of the test was to measure achievement. It should also be noted that all of the words in the textbook were presented in academic contexts (many of these words appeared in the AWL), and since nearly all of the students in my class planned to pursue university study a positive washback effect from studying these words and taking the test was expected.

During the construction of the test, it was soon realized that using all of the words students had studied would not be practical, especially since I was interested in testing several aspects of vocabulary knowledge (The initial pool of words appears with definitions in Appendix A). For this reason, a subset of words from the units studied was chosen following Nation's advice (see 2.3).

Since the vocabulary chart activity had required students to record information about several aspects of the target vocabulary it seemed logical to also include many of these aspects on the achievement test they would take so the test would be closely connected with the instruction they received (see 3.1). Although the vocabulary chart did not require the students to record the definitions for target vocabulary, this was still considered an aspect of vocabulary that was worthy of being tested because meaning was covered to a great extent through activities which appeared in the required textbook.

A basic outline of the pilot test is shown in the table below. The complete pilot test appears in Appendix B.

Basic Outline of the Pilot Test

Subtest	Task	Number of items
Part 1	Identifying stressed syllable	20
Part 2	Matching words to definitions	30
Part 3	Identifying parts of speech	15
Part 4	Writing definitions	15
Part 5	Writing sentences	15

As the table above shows the pilot test was composed of 5 subtests and included 95 items. The general construct that was intended to be measured by this assessment was vocabulary knowledge. However, the specific aspects of vocabulary knowledge that the test was meant to assess will be further clarified below through a discussion of the individual subtests. Additionally, this section will discuss decisions that were made during the development of individual items which are related to the general recommendations for constructing items mentioned earlier by Nation (see 2.4).

3.2.1 Subtest 1: Identifying the stressed syllable.

The first subtest was composed of 20 items. This subtest was designed to assess the students' abilities to listen for the stressed syllable in target vocabulary words. An example of this type of item appears below:

1. attitudes a—tti—ttudes

As shown above, for each item in this subtest the word being tested was written normally, next to the same word broken down into syllables with the insertion of a dash between each syllable. To indicate the “syllable with the most stress,” students were asked to circle the stressed syllable in a word after hearing it. It was assumed that students would be able to recognize the correct stress for these words through listening since they had heard them many times in class. Students had also received instruction on listening for word stress and were informally tested on this for the target words by their instructor during the vocabulary chart activity. In this way the item followed the first of Nation's four general recommendations for development because it reflected knowledge that was taught during the course. The construction of this type of item was also fairly easy. I began by looking at the complete pool of vocabulary that could potentially be included on the test and then tried to choose about three words from each unit that, when possible, had a different number of syllables. By using this method, it was hoped that the words chosen for the complete subtest would have various patterns for primary stress. An online dictionary was found to be a useful resource for

identifying the syllable boundaries and writing the words with divisions between each syllable as shown above.

An alternative format to this item, having the words presented without marking their syllable boundaries, was also considered, but from experience in the classroom it was found that students did not always underline the stressed syllables clearly in their vocabulary charts. It was predicted that presenting the words in this way would cause difficulties with scoring this subtest because the test rater would have more trouble determining whether or not a student had correctly identified the stressed syllable.

Having the students record these words in the language lab was also considered as another possibility for assessing their knowledge of stress for the target vocabulary words. This would have been beneficial since the test was being created for a speaking and listening course. The students were already familiar with using a recording application called Wimba Voice Board for other assignments in the, which would have facilitated the administration of this subtest. In this case the students would have been asked to read the words from a list and record them with headsets. However, it was also anticipated that this method would be problematic because it might not be possible to prevent the students from hearing their classmates' responses to the test items. Thus, it was expected that some weaker students would model the pronunciation of more proficient peers and that this would compromise the reliability of the test results. Also, administering this kind of subtest would either require administering the whole test in the language lab, where the setting made it more difficult to prevent cheating, or to administer this one section in the lab and the other subtests in the regular classroom, which would make this subtest more impractical. Thus, the chosen format was considered more favorable than the alternatives.

3.2.2 Subtest 2: Matching the words to definitions.

The second subtest required students to match vocabulary words to correct definitions. It was intended to assess the students' ability to recognize the meaning of target vocabulary words without the use of context clues. An example of this type of item is shown below:

1. nondairy	___	A. not very interesting
		B. not made with milk or cream

This subtest featured 30 vocabulary words, and thus, included 30 items. Answers were indicated by writing the letter for the correct definition from a list next to the corresponding word. As the following example from *Contemporary Topics 3* illustrates, the items for this subtest clearly reflected knowledge that was taught in the course.

Match each word to the correct definition.

a. accent	d. flexible	g. nonverbal behavior	i. random
b. discrete	e. generation	h. precise	j. ultimately
c. distinct	f. impressive		

_____1. able to change or be changed easily

_____2. when ideas or things are separate from each other

(Beglar & Murray, 2009, p. 83)

It should be mentioned, however, that vocabulary activities in the units studied in their textbook did not *always* involve the exact format presented above. However, the students did receive instruction regarding the meaning of words for each unit, and as mentioned earlier, each unit always included some activity that required the students to identify the meaning of target vocabulary. What the above example additionally shows is that the students could also be expected to be familiar with the matching item format and that this might lead to higher reliability for the achievement test.

Although the standard multiple-choice format was also considered as an option for testing meaning, matching was chosen because it was easier to create distracters for this format and required the creation of fewer distracters. Creating this subtest was also relatively easy because definitions for each of these were taken directly the students' textbook. These definitions were chosen because they were readily available and prevented the testing of other definitions of the words studied. The list of vocabulary words from each unit, shown in Appendix A, simplified the creation of this subtest even more because it organized the vocabulary and definitions according to units and included the parts of speech for each of the target words. Having the parts of speech clearly laid out made it easier to follow Reads advice to group words along with others which are the same part of speech (see 2.4). Another advantage of choosing the matching format was that scoring it would be very easy and could be done quickly and objectively.

Subtest 2 was ultimately broken down into three parts (A, B, C), with Part A featuring only nouns, Part B verbs, and Part C adjectives. One problem encountered when organizing these sections according to parts of speech, however, was that it became more difficult to choose a similar number of words from each unit. Upon inspection of the complete vocabulary list, it was discovered that the parts of speech for words in each unit were not evenly

distributed. Thus, it became necessary to test the meanings of more words from some units than others.

3.2.3 Subtest 3: Identifying parts of speech.

The purpose of the third subtest was to assess the students' ability to identify the part of speech of a word from the context. An example of this type of item appears below:

1. The proposed cuts have caused considerable **controversy**.

_____

This subtest included 15 items and featured a sentence for each word requiring identification. The target words were underlined, and the test taker was simply asked to write the correct part of speech next to the underlined word. This subtest employed different forms of target vocabulary than the ones presented in the students' textbook. Students were expected to have had some prior knowledge of these forms since they were common forms that were presented during feedback on the vocabulary charts, and thus this subtest was closely related to instruction during the course. Creating this type of item was not especially difficult. By collecting and making copies of the students' vocabulary sheets, a record of alternative forms for target vocabulary was already available. Choosing words for this subtest was arrived at simply by selecting nearly an equal amount of words from each unit studied. In order to ensure that the sentences for this subtest were comprehensible to the learners, and to avoid testing the students' understanding of words other than the target words, most sentences were taking from the *Oxford American dictionary for learners of English* so they would be composed of high-frequency vocabulary.

It was soon decided that the sentences in this subtest could be used again to test a different aspect of word knowledge in the following subtest. Because the sentences that appeared for the target words in the learner's dictionary did not always seem to provide a clear enough context for the word so that students could be expected to define the word later in the following subtest, some sentences had to be created originally. One advantage of this item type was that scoring could be done easily using the list of the correct parts of speech that had already been created.

3.2.4 Subtest 4: Writing definitions.

The fourth subtest required the students to write definitions for the words that appeared in the sentences in subtest three. This subtest was designed to assess the students' ability to write a definition for a word based on the context. Therefore, this was another

subtest that focused on the meaning aspect of the words. An example of this type of item follows:

1. controversy:
Means_____

Like subtest three, subtest four included 15 items. The words from subtest three were listed and juxtaposed with “[the word] means” and followed by a blank. An explanation using an example from the students' textbook, shown below, illustrates how this subtest was connected to classroom instruction.

Listen to each sentence. Then guess the meaning of the boldfaced words. Work with a partner.

1. Gerhard speaks English with a slight German **accent**. His pronunciation is a little different from someone who grew up in an English-speaking country.

(Beglar & Murray, 2009, p. 83)

The example above shows how vocabulary was most often initially presented to the students in the units they studied during the course. From this example, it should be clear that one vocabulary skill that was a component of the course was inferring meaning from context. As a first step in inferring the meaning of target vocabulary students were often prompted to think about the part of speech for the words presented in sentences. This is a reason why a decision was made to connect this subtest with the task of identifying parts of speech.

The main difficulty in constructing this item type was already mentioned in the previous section and these items could be constructed very quickly.

One concern with this subtest, however, was with how it would be scored. It was anticipated that a large number of the responses students would provide for this subtest would be questionable because students could differ in the wording of their definitions. In order to provide more certainty that students understood the word they were asked to define, some limitations were put on acceptable answers. For example, students were told that they would “not receive points for answers that use another form of the word.” A key reason why this limitation was put in place was to prevent students from using knowledge of affixes along with part of speech knowledge in order to avoid providing a meaningful definition. For this reason an example of an unacceptable answer was given using the word “unreliable.” The example spelled out that “not reliable” would not receive credit as a response. Use of “not” along with a synonym for “reliable,” however, was considered a demonstration of knowing the word’s meaning. For example, "not dependable" would be an acceptable answer.

3.2.5 Subtest 5: Writing sentences.

The fifth subtest featured a list of words with instructions to “write a sentence . . . to show that you know what the word means and how it is used.” This subtest was intended to assess the students' ability to use a word grammatically in addition to using it meaningfully. An example of this type of item is given below:

1. attitude:

_____

This final subtest included 15 items that targeted words from the textbook units. This item type was directly related to a task that the students performed while completing their vocabulary charts and was thus directly related to instruction and vocabulary learning during the course. An advantage of this item type was that it was extremely easy to construct. As with most of the other items, choosing the words for this subtest was accomplished by trying to select a nearly equal number of words from each chapter studied. One difficult consideration that had to be made for this item type, however, was how the item would be scored. After looking at the vocabulary charts that the students had completed in class, it was obvious that many of the learners were unable to construct sentences without any grammatical errors. Because of this, it seemed unfair to expect the students to do this on the achievement test. On the other hand, the sentence-writing section of the vocabulary chart also revealed that it was difficult because of the grammatical mistakes to determine whether or not some of the students could use the target vocabulary words. It was clear that some guidelines for scoring needed to be developed to address this issue.

Read's recommendation for scoring this type of item (allotting one point for grammar and one point for use) could be expected to promote reliability and practicality for this subtest. However, considerations for scoring this subtest were further complicated by the fact that not all grammatical errors from the vocabulary chart seemed to affect the students' overall intelligibility. These students were, after all, not yet fluent speakers of English and therefore should be expected to make some grammatical mistakes. Consider for example, the following sentence that one student wrote for the word *interracial*:

My friend is from US and his wife is from Phillipine, so they are interracial couple.

There are some obvious grammatical mistakes in this sentence, such as the omission of articles. However, the above sentence also seems to show a clear understanding of the target word. Errors dealing with articles are very common for learners even at higher levels than the ones in

my own class. The question was, should the learners be penalized for making grammatical mistakes such as this one? Read's suggestions for scoring sentence-writing items made no mention of this issue.

Consequently, it became clear that allowing some grammatical errors in this subtest would require guidelines for acceptable grammatical errors. The problem was that it was still unclear what grammatical errors *should* be considered acceptable if any were to be accepted at all. Allowing some grammatical errors would require a rubric to deal with the acceptability of errors in this subtest if the scoring was to be done objectively. Creating such a rubric would be time-consuming and difficult, and was also likely to make scoring this subtest more challenging. Therefore, the scoring criteria for this subtest ultimately followed Read's recommendation. This subtest also included an example to demonstrate that a full-credit answer would require a sentence that revealed an understanding of the meaning of the target word using correct grammar, that a half-credit response would show an understanding of meaning but involve incorrect grammar, and that a no-credit response would not show an understanding of the meaning of the target word.

As the description above of the test and the individual subtests has shown, the test as a whole and the individual subtests all included a fairly large number of items. This was done intentionally because having more items than what is necessary for a final test has been recommended in the literature on piloting (see 2.5).

3.3 Administration of the Pilot Test

Arrangements were made for the test to be piloted in a context that was similar to the one in which the final test was to be to administered in order to increase the possibility that the final test would yield reliable results after it underwent revisions. The main purposes of this initial pilot session were to determine if a 50-minute time limit was appropriate for the administration of the test, and if the test items were at an appropriate level of difficulty.

The pilot was administered to students in a different section of level-3 speaking and listening at MELP. This course was taught by a colleague whom I shared an office with, so I had considerable insights into the details of the pilot group's course. For this course the learners were using the same required textbook as my own students. This was optimal since the test that was being piloted was meant to assess knowledge of words in this textbook. Through conversations with the instructor of this course it was revealed that these students were receiving instruction on all of the aspects of vocabulary that were meant to be assessed by the test. In fact, this instructor was also well aware of the vocabulary chart activity that was being used by my own students because she had assisted me in developing it. However, this instructor decided not to have her students complete the vocabulary charts for her own class,

and instead preferred to convey this information through other methods of instruction. Additionally, since the placement procedures for this course also required the students to have a TOEFL score along the same range as the students who would take the final test, the pilot group was similar in proficiency. Although perhaps not as significant, the language backgrounds of many of the students who took the pilot were also similar to the backgrounds of students who took the final test. Most students in the pilot group were native speakers of Chinese, Arabic, or Korean.

While the pilot group and final group shared many similarities, it is essential to note that the final group studied two units in their textbook that the pilot group did not (units 8 and 9). For this reason, students in the pilot group had not received formal instruction in their course on some of the words in the pilot test.

The pilot test was administered by my colleague on December 5, 2011 to her entire class, although only ten students signed a release so that their data could be used for this study. The constructor of the test was not present during the administration of the pilot, however, details regarding how the pilot was presented were conveyed by the colleague who administered it. As should be clear from the title on the pilot, it was presented as a "practice test" (see Appendix B). The students were told that this test was being administered for research purposes and would not count toward their grade in the course. However, the students were also reminded that they would eventually be tested on this vocabulary as part of their course grade, and that the results of this test would give them feedback that would help them study for it. Presenting the test in this way was expected to sufficiently motivate the students to take it seriously without causing them too much anxiety. The administrator of this test also explained the directions for each section before the test began and students were encouraged to request clarifications. Students in the pilot group had 50 minutes to complete all of the items once the test began. Therefore, the time allotted for the test was the same as it was expected to be during the final administration.

Once the 50-minute time limit was up, the test administrator collected the students' tests. The tests were then delivered to the test constructor for grading.

3.4 Scoring and Revising the Pilot

When the results of the pilot were received, it was found that students had skipped many items, especially in subtests 4 and 5. Observations of the test administrator revealed that this was likely due to the time limit. For this reason, the primary revision that was deemed necessary for the pilot was to eliminate some of the items.

Decisions about which items to eliminate were made primarily by examining the number of correct answers for each item. The test items were analyzed by constructing a tally

sheet for each of the subtests (see Appendix C). These tally sheets categorized the students' responses to the first three subtests below the headings “correct” and “incorrect." For the last two subtests, two additional headings were included to categorize the students’ responses: “partial credit” and “no attempt.” The rationale for distinguishing between incorrect answers and "no attempt" responses in subtests 4 and 5 was that these subtests were the only ones in which students skipped items and it was thought that this distinction would potentially have value during the revision process. In essence these tally sheets were used to "eyeball" the item facility (IF) for each of the test items.

IF is a statistic used to examine the number of students who correctly answered a given item. An IF index of .27 indicates that 27 percent of the students correctly answered the item. An IF of .96 indicates that 96 percent of the students correctly answered the item. Therefore, lower IFs indicate more difficult items, while higher IFs indicate easier items. In this discussion of revisions to the five subtests the IFs of items are reported in this way instead of the way the data appeared on the tally sheets in order to make this analysis more reader-friendly. Also, only the IFs of the relevant items are mentioned in this section. A complete list of IFs for all of the items in the pilot is shown in Appendix B alongside the actual pilot test. Appendix B also shows which items were omitted from the final test.

Additionally, after the pilot test had been administered, the results of a “peer critique” of the pilot test were received. This critique was done by my colleague, an MATESOL student and ESL instructor at the University of Minnesota, and appears in Appendix D of this paper. Although, as mentioned in the literature review of this paper, it is generally recommended to have a pilot test critiqued before it is administered to the pilot group (see 2.5), time constraints did not make this possible. In order to enable the test to be piloted under conditions that were similar to that of the final group, the pilot had to be administered before it was critiqued.

The remainder of this section will discuss how a closer examination of the pilot, which included inspection of the IFs of each item, and feedback presented in the peer critique affected the revision of each of the 5 subtests before the final administration.

As expected, scoring subtest 1 (which required students to identify primary stress) was very easy and could be done quickly because the format basically made it multiple-choice. The primary considerations for revising this subtest came from feedback given in the peer critique and an analysis of the IFs for the items. One recommendation in the critique was not to break down the syllables of the words in this subtest since students at this level of proficiency might be able to do this on their own and that perhaps this oversimplified the task. This suggestion did not make it to the final version, however, because the vocabulary sheet activity had revealed many examples of students circling or underlining in confusing ways that made it difficult to judge whether or not they could correctly identify the stressed syllable when the words were written normally. Thus, the main factor that inevitably influenced the revision of subtest 1 was the IFs of the items.

Since the students from the pilot group did not complete the vocabulary chart activity in their course, it was inferred that the final group might be able to perform better on these items. For this reason, it was decided that the items which were answered correctly by most of the test-takers were good candidates for removal from the test since they might be answered correctly more easily for reasons other than ability. The chart below shows the IFs for the items in subtest one that were omitted from the final test. By removing only eight of the twenty original items, it was also thought that this subtest would still be comprehensive enough so that it would provide a valid measure of the students' abilities to recognize stress without becoming excessively lengthy.

IF Indices for Items Omitted from Subtest 1

Item	IF
2	0.70
6	0.80
8	0.70
10	0.80
11	0.70
12	0.70
14	0.90
15	0.80

Scoring subtest 2 (matching) was also accomplished very quickly and easily, as expected, because of the format of the items. Revision of this subtest was primarily motivated by the belief that the items for this subtest should be at a reasonable level of difficulty, but that revision should also result in the preservation of enough items so that the subtest would be a valid measure of the students' ability to identify the meaning of the target words. For this reason the IFs for the items in each section were considered.

As the table below shows, two items were omitted from each of the first two sections and three from the third section. These items were chosen because there were expected to be too easy for the final group of test-takers since their IFs were very high.

IF Indices for Items Omitted from Subtest 2

Item	IF
Part A *28	0.90
*29	1.00
Part B*31	0.70
*37	0.90
Part C *44	1.00
*49	0.80
*50	1.00

An analysis of the answer choices for the first section of the subtest also revealed that two of the choices might be acceptable for one of the items. Choice N ("eagerness or willingness to do something") and choice I ("readiness to take action") both seemed to be acceptable choices for the word *motivation*. For this reason, choice I was omitted from the final version of the test. Additionally, a thorough examination of the target words for Part B revealed that items 32 (*evolve*) and 40 (*mature*) could both be answered with the same definition ("develop") presented in the choices. Choice F was obviously the best answer choice. For this reason, *mature* and its definition should have been omitted from the final test. However, an error was made during revision because the choice ("develop physically or emotionally") was omitted from the final test, but *mature* was not. Therefore, the final test included two items for Part B which could only be answered with one choice from the options. This error will be revisited later in an analysis of the data for the final test.

For subtest 3 (identifying parts of speech), the reviewer who wrote the peer critique found that this subtest had many more nouns than adverbs, adjectives, or verbs. Based on this observation, she questioned the validity of the subtest since it was intended to measure ability to identify parts of speech in general, but seemed to prioritize one part of speech over the others. To lend to this subtest's validity, she suggested including a more balanced number of parts of speech. As with the previous two subtests, looking at the IFs of these items was considered a logical way to choose which items to omit. However, it immediately became clear that omitting only the items with the highest IFs would not solve the problem identified in the peer critique. Since the critique revealed that this subtest included 7 nouns, 4 adverbs, 2 adjectives, and 2 verbs, it was decided that the omission of items should focus on nouns and adverbs because they received more attention than the other parts of speech. Items 51, 57, 59, and 64 were all omitted from the subsequent test because they were noun items with high IFs (as shown in the chart below).

IF Indices for Items Omitted from Subtest 3

Item	Part of Speech	IF
51 (*commerce*)	noun	0.90
54 (*distinctly*)	adverb	0.80
57 (*enabler*)	noun	0.80
59 (*guarantee*)	noun	0.80
64 (*confirmation*)	noun	0.90

The decision to omit item 54 was inevitably done for reasons other than its IF. Instead, it was thought that this item might be problematic because the sentence for this item may not have provided a rich enough context for students to successfully define it for the corresponding item in subtest 4.

4. His government is looking distinctly shaky. (Item 54)

Indeed, an inspection of the IF for the corresponding item (item 69) in the table for subtest 4, shown below, reveals that this item was answered correctly by only 2 out of 10 test-takers. Additionally, the answers of these two test-takers (*definite*, and *remarkable*) for item 69 only received half credit because they did not use the correct part of speech in their definitions. Also, awarding half credit for *definite* might also be considered a bit of a stretch, since even *definitely* (which presented the correct form) would have arguably been inappropriate in the sentence above.

An alternative to omitting this item would have been to create a new sentence for it in the subsequent version of the test. However, since there was no time to re-pilot this item before the final administration, eliminating it was considered a more reliable choice for improving the quality of the test.

Scoring subtest 4 (definition writing) proved more difficult than the previous 3 subtests. In the peer critique, the reviewer noted that the directions for this subtest ("write a definition for each word from Part 4 based on how it is used in the sentence above") seemed questionable and wondered if the definitions for words in subtest 4 needed to match the part of speech of the words from subtest 3. This was exactly what the directions were indicating from the viewpoint of the test constructor. However, the reviewer raised the possibility of accepting an answer like "assure" for the word *guarantee* even though it was used as a noun in the previous subtest. Initially, responses like this, however, were treated as incorrect answers during scoring. However, when it was observed that a noticeable number of answers defined the word according to a different part of speech, the items were rescored and students with this type of response were awarded partial credit. Because of this, it seemed reasonable to revise the final test so that students would be aware of the possibility of a partial-credit response, but since students in the final group had been encouraged to write definitions which used the same part of speech for the vocabulary chart activity, the ultimate decision was not to include this change on the final test.

The reviewer had also implied that it might be unreasonable to count definitions that used the same form of the word, such as "not reliable" for the target word *unreliable* as incorrect because words were commonly defined this way in dictionaries. In spite of this noteworthy observation, it was decided that this modification should not be made to the final test, because from the viewpoint of the test constructor, using the target word itself as part of the definition generally does not reveal an understanding of the word's meaning. However, one modification that was suggested by the reviewer which did impact the final test, was to label the examples for this subtest "acceptable answer" and "unacceptable answer," respectively.

The fact that items for this subtest were connected to items in the previous subtest made it unnecessary to consider the IFs of these items in order to determine which ones should be omitted. Since it was necessary to use the information presented in subtest 3 to answer the items in subtest 4, items were logically omitted from subtest 4 if the corresponding items were omitted from subtest 3. Thus, items 66, 69, 72, 74, and 79 were all omitted from the final test because they featured words used in the items omitted from subtest 3. However, the IFs of the first four items shown below also reveal that they might have been good candidates for omission based on their high level of difficulty.

IF Indices for Items Omitted from Subtest 4

Item	IF
66 (*commerce*)	0.30
69 (*distinctly*)	0.20
72 (*enabler*)	0.10
74 (*guarantee*)	0.40
79 (*confirmation*)	0.60

Scoring subtest 5 (sentence writing) required more time and effort than scoring the first three subtests, however, the time required for scoring was probably minimized due to the fact that it did not involve decisions about acceptable grammar mistakes, since a sentence with any grammar mistake was reduced by one point. This method of scoring, however, was challenged by the reviewer in the peer critique, who thought that reducing a respondent's grade for all grammatical mistakes in this subtest was unfair and that it might be reasonable to award full credit to an answer that included incorrect grammar but demonstrated knowledge of use. Her argument was reasonable and had been given consideration prior to constructing the pilot. The ideal way to construct criteria for scoring this subtest would have been to determine what grammatical mistakes should be counted as meaningful based on the responses of the pilot group. However, since many students had not attempted to answer the items for this subtest, there was an inadequate number of example sentences from which to determine such criteria. Therefore, this revision could not be made to the pilot and had no impact on how items were scored in this subtest for this administration. As with subtest 4, in response to the feedback of the reviewer, examples for subtest 5 were changed to "acceptable answer" and "unacceptable answer" on the final test in an effort to improve the clarity of the directions, which has been noted to have a positive effect on reliability.

The IFs of the items for this subtest were also considered in the revision of the pilot. Three items were omitted because they were considered too easy and two items were omitted because they were considered too difficult. This information is presented in the table below.

IF indices for Subtest 5

Item	IF
81	0.80
82	0.30
*83	0.80
84	0.80
85	0.60
86	0.70
*87	1.00
88	0.90
89	0.40
90	0.60
*91	0.00
*92	0.40
93	0.20
94	0.50
*95	1.00

Note: Items with an * were omitted from the final version. Calculation of the IFs above grouped partial-credit responses along with full-credit responses.

Items 83, 87, and 95 were omitted because they were answered correctly by most of the students, and it was therefore expected that these items might not discriminate as well as others. Upon a later inspection of the exact IFs for these items, however, it was discovered that item 88 might have been a better choice for elimination than item 83 since it had an even higher item facility. Items 91 and 92 were omitted from the subsequent test because they were answered correctly by few of the students and were thus thought to be at an unreasonable level of difficulty. A later inspection of the exact IFs for these items again revealed that another item would have been a better choice for omission statistically. In this case, items 82 and 93 both had lower difficulties than item 92. However, this was apparently an oversight.

Because the tally sheet for this subtest initially categorized the items into four categories (full credit, partial credit, no credit, and no attempt) instead of two categories, as in subtests 1-3, it seems that "eyeballing" the IFs from the tally sheet for this subtest was not done with the same degree of accuracy as for the first three subtests.

Ultimately, the revisions of the five subtests mentioned above greatly reduced the number of items on the final test. The final test is shown in Appendix E. As the test shows the number of items was reduced to 65 items total, leaving 12 , 23, 10, 10, and 10 items, on subtests 1, 2, 3, 4, and 5, respectively.

3.5 Administering the Final Test

The final test was administered in my own Advanced Speaking and Listening class at the University of Minnesota on December 12, 2011. A few days prior to taking the test, the students had received the study guide in Appendix F. As the study guide shows, students were provided with the same examples and directions for completing each subtest that appeared on the final test. Additionally, the study guide included a list of all of the words the students could potentially be tested on. The students were also advised to review their vocabulary charts and the vocabulary sections that were covered in class from their textbook. It was also emphasized to students that while these resources would help them to prepare for the test, they would not be able to use them during the test.

On the day of the test, students were again offered the opportunity to ask questions regarding the directions for each section of the test. Once students had confirmed that the directions were clear, they were told that they would have 50 minutes to complete the test and that they should quietly hand in their papers if they finished early. They were also told that even if they finished early, they would still have to stay for the remainder of class. This was done to encourage the students to work diligently for the entire time. Each of the words for the first subtest was read according to the directions and the students worked at their own pace for the remaining subtests.

It took most of the students 45 minutes to finish the entire test. A couple of students wanted some additional time. However, their tests were still collected promptly once the 50-minute time limit was up.

3.6 Scoring the Final Test

In contrast to the pilot test, students who took the final test attempted nearly all of the items. Scoring the first three subtests on the final test took about the same amount of effort and time as it had for the pilot version and the items were scored according to the directions that appeared on the test.

When scoring the fourth subtest, the consideration of partial-credit responses was once again revisited. Upon analysis of the students' responses, it was found that some answers revealed partial understanding of target words and therefore might deserve some credit. In order to score partial-credit responses objectively, the following criteria were created to define partial-credit answers: "half credit for answers that show a degree of understanding of the meaning, or express the meaning in a way that does not clearly show that the student can define the word with the same part of speech." Examples of half-credit responses are shown in the table below.

Half-Credit Responses for Subtest 4

Half-Credit Answers	Example
Shows a degree of understanding of the definition	*finish* is given as a definition for *ultimate*
Definition does not take into consideration the part of speech for the target word	*changeable* is given as a definition for *flexibility*

Scoring the fifth subtest of the final, just as with the pilot test, also proved challenging and resulted in a reconsideration of the scoring criteria. Initially, the items were scored according to the directions so that all grammatical errors reduced a response by one point. Scoring in this way soon revealed that few students received full credit for items in this subtest, just like the pilot group. The penalty for incorrect grammar ultimately seemed to be too harsh considering that some grammatical mistakes the learners made did not significantly impair the meaning of the word used in the sentence. Additionally, it was realized at this point that the proficiency level of the learners (which was only intermediate) should be taken into consideration with regard to the grammatical errors made. Therefore, the students' errors were analyzed in order to devise a rubric for acceptable errors. It was decided that responses that made minor errors with prepositions, articles, and the third person singular were common and did not impair meaning. Additionally, some students put adverbs in positions that were ungrammatical or used a similar, but incorrect, modal instead of the right one. Mistakes such as these did not cloud the meaning of their sentences. The final decision was that cases like the above were still counted as full-credit responses. Examples of acceptable grammatical errors are shown in the table below.

Grammatical Errors That Were Accepted When Scoring Subtest 5

Acceptable Grammar Error	Example
Minor error involving absence or incorrect use of a preposition	*The US is an interracial society because there are many people around the world living there.
Minor error involving absence or incorrect use of an article	*The US is interracial society because there are many people from around the world living there.
Minor error involving third person singular	*She send pictures via email.
Ungrammatical positioning of an adverb	*A new car will work better constantly than an old car.
Substitution of a similar modal for the correct one	*You should follow the writing strategy so you could do well in your essay.

All other grammatical errors, however, still reduced the score by half. It was also decided that responses that used a word with a common collocation that did not indicate a clear understanding of meaning, such as "people can use non-verbal communication," for the target word *verbal* would also receive half credit. There were many other responses that clearly showed no understanding of a word and were very easy to grade without considering any adjustment in the scoring procedure.

The final test was graded according to the number of points a student received out of a total of 85 possible points.

3.7 The Results from Administering the Final Test

The indicators of central tendency and dispersion for the final test are shown in the table below. These statistics were calculated using Microsoft Excel. As the table shows, the mean for the test was almost 20 points below the total possible score. None of the students had the same score (indicated by the mode). This is not surprising though, because of the relatively small number of students who took the test. Additionally, the range of scores for this test was fairly wide. The lowest score on the test was 48 and the highest 83. Lastly, the standard deviation reveals that the test as a whole produced a fairly wide spread of scores for the test takers.

Indicators of Central Tendency and Dispersion

Statistics	Values for Total Test
Number of Students (*N)*	11
total possible points	85
mean (*M)*	65.55
mode	N/A
median	65
midpoint	65.5
low-high	48-83
range	36
standard deviation (*S)*	9.59

Additional statistics for this test were calculated using the fifth version of the Laboratory of Educational Research and Test Analysis Package (Lertap, 2003-2011). An examination of these statistics will be presented in the following four sections that will discuss the reliability of the tests and subtests, item difficulty and discrimination for the test items, correlations between the subtests and total test, and the validity of the test.

3.7.1 Reliability.

Test reliability is the extent to which the results of a test can be considered consistent or stable. In other words, a reliable test would yield the same or similar results upon additional administrations if all conditions for the testing environment remain constant (level of students, teaching, etc). Reliability was calculated by Lertap using the Cronbach Alpha coefficient. According to the rule of thumb for classroom tests presented by Douglas, a classroom test has good reliability if r = .70 or greater (Douglas, 2011, p. 107).

This section will present an analysis of the reliability for the entire test and discuss the factors that may have contributed to the test's reliability. The following section will present more information related to reliability with an analysis of the individual subtests.

The reliability for the test as a whole was very good (0.84). So why was the reliability for this test so high? We can begin investigating this by considering four factors that lead to higher reliability mentioned in Brown:

1. A longer test will tend to be more reliable than a shorter one.

2. A well-designed and carefully written test will tend to be more reliable than a shoddy one.

3. A test made up of items that test similar language material will tend to be more reliable than a test assessing a wide variety of material.

4. A test that is clearly related to the objectives of instruction will tend to be more reliable than a test that is not obviously related to what the students have learned.
(Brown, 2005, p. 215)

With regard to the first point, the test was fairly long for a classroom test. Recall that the test had a total of 65 items, which is quite a large amount. One of the reasons why it was possible to have so many items on the test and to administer it in 50 minutes was because the formats for many of the items (multiple-choice, matching, single-word response) allowed the test takers to complete items at a fairly fast pace.

Secondly, reliability of the test was probably enhanced due to the amount of effort spent on designing the test. Several possibilities for subtests and the format of the items were considered before even developing the pilot. Many of the item formats chosen for the test were not only chosen because they seemed appropriate in the mind of the test constructor, but also because they were considered appropriate measures of vocabulary knowledge by experts in vocabulary research. Additionally, many of the items that were not well-written were weeded out after piloting. Also, great care was taken to provide clear examples of how to

complete the test items, and the directions for completing items were revised further after the peer critique. The clear directions shown on the final test may have increased reliability by reducing errors that were a result of misinterpreting the directions instead of a lack of ability. Also, many of the items on the test could be scored easily and objectively since they had only one possible answer, and several methods were given careful consideration for scoring the more open-ended items before the final methods were adopted.

Third, the test may have been reliable because the items tested similar material. For one, each subtest was organized so that all of its items clearly focused on one objective (e.g., identifying the stressed syllable, identifying the definition, etc). Also, this was a vocabulary test and therefore, all of the subtests intended to focus on a particular objective related to vocabulary knowledge.

Fourth, the test was created with careful consideration given to the objectives of instruction and this in turn may have increased its reliability. For one, many of the subtests were based on aspects of word knowledge that were presented in the vocabulary chart activity, or in the students' textbook. Before taking the test, students were informed that the test would be based on information from these sources. Additionally, vocabulary learning was emphasized as an important objective in the course because the students were aware that knowledge of vocabulary would help them to perform challenging speaking and listening tasks.

3.7.2 Item discrimination and item facility.

Although the test as a whole had an excellent reliability coefficient, this was not true for all of the individual subtests. As the table below shows, two of the subtests fell below the optimal range for reliability (see 3.7.1).

Reliability for the Subtests and Total Test

Subtest	Reliability (Coefficient Alpha)
Subtest 1	0.83
Subtest 2	0.07
Subtest 3	0.84
Subtest 4	0.70
Subtest 5	0.64
Total Test	0.84

In addition to the factors mentioned above, test reliability is affected by the Item Discrimination (ID) and Item Facility (IF) of the individual test items. ID indicates the degree to which an item separates students who performed well from those who performed poorly on the test as a whole. It allows teachers to contrast performances of upper students with lower students. An item with a discrimination index of 1.00 is very high because it indicates the

maximum contrast between the upper and lower groups of students, that is, all the high students answered correctly and all the students in the lower group answered incorrectly. Since an item with an ID of 1.00 is indicating that the item separates the upper and lower groups in the same manner as the whole test scores, such an item is a good candidate for retention in any revised version of the test.

ID indexes can range from 1.00 (all the students in the upper group answered correctly and all the students in the lower group answered incorrectly) to -1.00 (if all of the students in the lower group answered correctly and all of the students in the upper group answered incorrectly) and can take on all values between 1.00 and -1.00. Items with a negative ID would indicate that the item is testing something quite different from the rest of the test since students in the lower group performed better than those in the higher group. Consideration should be given to such items to see why they are functioning in such a way.

An explanation of IF was already discussed earlier (see 3.4).

Optimal test items will have an ID that is 0.30 or greater and an IF of 0.60-0.80 (see Cohen, 1994). Therefore, to create a reliable test, test constructors will want to have as many items as possible that fall within these ranges. When selecting items for retention on a pilot version of a test, the test constructor will want to first look at the IFs of items within the desirable range and then from that group select the items with the highest ID possible. If a pilot version included well-constructed items from the beginning and included a large enough amount of items, the test constructor will likely find the process of arriving at a polished final test greatly simplified.

While the test as a whole had 14/65 items within the desirable range for both ID and IF, items with these ideal statistics were found more commonly in some subtests than others. The remainder of this section will consider the reliability of the individual subtests, especially in reference to the IDs and IFs of the items on each subtest.

The reliability coefficient for subtest 1 (word stress) was 0.83. Therefore, this subtest was extremely reliable. Four of the items were within the optimal statistical range for IF and ID. These items (3, 9, 10, and 11) had IFs ranging from 0.64-0.73 and IDs ranging from 0.60-0.74. Items 1 and 2, however, were of very low quality statistically, because they were answered correctly by all students (IF 1.00) and therefore did not discriminate. One item of concern is item 5 (shown below) since it had an ID of -0.01.

5. guaranteed guar—an—**teed** (Item 5)

A possible explanation for why this item discriminated poorly could be that this word included secondary stress. Since the first stressed syllable was the one that received the secondary stress, learners may have chosen this one simply because it was the first stressed syllable that they heard as the word was being pronounced. An examination of the students' responses

revealed that the two test takers who answered this item incorrectly did in fact choose the first syllable for their response. An examination of item 7 (*controversial*) also revealed this pattern for incorrect answers. For this item, the five students who answered incorrectly all chose the first syllable (the one that received secondary stress). Since the directions for subtest 1 stated that students should circle the syllable that receives the "*most*" stress, it seems unreasonable to infer that items with secondary stress were unfairly luring the students to the wrong answer because the directions account for only one clear correct answer.

However, this pattern also leads to an insightful consideration regarding the instruction that the students received in class. When students completed the vocabulary chart activity, some of them asked questions about the secondary stress for some of the words that were studied. In response, the students were told that they would only need to learn the primary stress and that secondary stress would be covered later. Since there was not enough time during the semester to cover secondary stress, however, students did not receive instruction on it. Considering this last point, perhaps it *was* unfair to use words with secondary stress for this subtest since knowledge of secondary stress could be considered a prerequisite for answering these items correctly.

Ultimately, if this subtest were administered again, a logical revision would be to omit items 1 and 2 since they were statistically poor. This would preserve 10 items for this subtest, which should still provide a large enough amount of items to measure the intended construct. Although these 10 items would not all be within the desirable statistical range, this subtest was already reliable and is likely to be even more so upon revision. Lastly, the issue of the secondary stress items could be dealt with in another way that would not require the inclusion of new items. Instructional modifications could be made before a further administration so that the subsequent group of students could receive some instruction on secondary stress.

Some common characteristics of the best items in this subtest also imply one factor that might contribute to creating high-quality items of this kind. Two of the four items within the desirable statistical range (items 3 and 11) had five syllables, which is more than any other word used in this subtest. The results of these items imply that as the number of syllables increases for items of this type so will the reliability of the test. Test constructors should keep this in mind when creating this kind of subtest.

The reliability for subtest 2 (matching) was only 0.07 and was the lowest of all of the five subtests. For subtest 2, only one of the items, item 18 fell within the optimal range for IF and ID. The IF of this item was 0.73 and its ID was 0.33. Additionally, there were a lot of items that can be considered too easy on this subtest. In fact, 14 of the 23 items had an IF above 0.90 and many were answered correctly by all test-takers. Also, there is reason to be concerned about six of the items on this subtest (13, 14, 22, 23, 25, 30, and 32) because they had negative IDs. A logical reason for the poor discrimination of two of the items was mentioned earlier in the section about revising the pilot. Because of an error during the revision, only one correct

answer choice existed for items 22 and 23. To fix this error, item 23 should be deleted from the subsequent version of the test. This revision would likely improve the discrimination of item 22. Reasons for the poor discrimination of other items in this subtest are not as clear. However, the items could probably have been better constructed.

One revision that could potentially result in better performing items was identified in the work of Nation after the final version of the test was constructed. For matching, Nation recommends that using small blocks (fewer answers and answer choices) for a matching test can have a positive effect (Nation, 2001, p. 350). Nation's suggestion could be used in order to create an even better version of this subtest for future administrations.

Common characteristics which separated better from poorer items in this subtest could not be found. However, one explanation for the poor discrimination of many of the items in this subtest is that this format allows students to guess the answers of items more easily than others.

The reliability for subtest 3 was 0.84 and the highest of any subtest. The IFs for items in this subtest ranged from 0.55-1.00 and the IDs ranged from -0.14-0.48. Interestingly, although this subtest had the best reliability coefficient, none of the items had both an IF and ID that fell within the desirable range. One item for this subtest, item 38 (shown below), is potentially a concern because it has a negative discrimination index (-0.14).

3. He said it is still not possible to predict the ultimate outcome. **adjective** (item 38)

This item was investigated, and it was found that while other students correctly answered "adjective" for this item, one student answered "verb." Since the item itself does not seem misleading in any way and the performance of the test-taker who got this item wrong was in the high group for performance on the entire test, a reasonable explanation is that this test-taker was guessing. Given the fact that this subtest was highly reliable overall it could reasonably be maintained without a change before a subsequent administration even though the statistics of the individual items could be improved.

Interestingly, the two best discriminators, items 39 and 44 (shown below), were also the ones that provided the most context for the target words. These items both had a 0.48 ID.

4. The flexibility of the lens decreases with age; it is therefore common for our sight to worsen as we get older. **noun** (item 39)

9. Humans experience a delayed maturity; we arrive at all stages of life later than other mammals. **noun** (item 44)

How the context for this items would have influenced their ID, however, is unclear since the additional information for these items did not seem to provide any more clues related to parts of speech than the context in other items with lower IDs

The reliability coefficient for subtest 4 was 0.70, and therefore this subtest just met the recommended criteria for reliability. The reliability for this subtest, was enhanced due to the fact that it included five out of ten items (46, 48, 50, 52, and 54) in the optimal range for IF and ID. The IFs for these items ranged from 0.64-0.77 and their IDs ranged from 0.35-0.67. However, there were no common characteristics for these items which suggested why they might have performed better than items that were statistically poor.

Another factor that may have contributed to this subtest's reliability was the care taken in creating scoring criteria for the items in order to reduce test-scorer subjectivity. Additionally, the answers for items in this subtest were open-ended and did not allow students to employ test-wiseness strategies very easily. In other words, it was not so easy to get the items right without understanding the vocabulary. Therefore, these items were also more cognitively demanding, which is probably why many were consistently answered correctly by students who possessed a high degree of vocabulary knowledge as indicated by overall performance on the test.

While the statistics of some items in this subtest could be better in order to improve reliability, the fact that this subtest already meets the desirable level of reliability does not necessitate revision of this subtest before another administration.

Lastly, the reliability of subtest 5, 0.64, was just a bit below the desirable level. This subtest had two items (59 and 65) with both an IF and ID that fell within the desirable range. Item 59 had an IF of 0.73 and an ID of 0.65. Item 65 had an IF of 0.73 and an ID of 0.74. However, it is unclear why these items may have performed better than others.Two of the items (57 and 63), had exceptionally high IFs (0.90+) and were poor discriminators with IDs ranging from -0.12-0.5. Also, there is reason to be concerned about item 57 (shown below) because of its negative ID.

2. constantly: **When I was young, I hated my sister because she *constantly* bothered me.**

However, a clear explanation for why this item had negative discrimination could not be found after analyzing the students' responses.

Since the reliability for this subtest was not that far below the optimal level, a massive revision might not be necessary to create an acceptable subsequent version. A reasonable way of improving the reliability of this subtest might be to simply remove some of the poorest items. Item 57 would be a good candidate for omission because of its ID and IF as would item 61 since it had an IF of 0.86 and an ID of 0.00. Removal of these two items on a subsequent

version of the test could potentially yield the desired results for reliability while still allowing this subtest to include enough items to be potentially valid.

An additional consideration for a revision of this subtest would also be whether or not to change the directions to reflect the way in which the items were actually scored. Recall that directions on the final test read, "A sentence that uses . . . incorrect grammar is worth half credit," but that items were ultimately scored in a way that permitted some grammatical errors. While it might seem that it would be more fair to inform the learners that they will not be penalized for *all* grammatical errors in the directions for this subtest, doing so might also send them a message that correct grammar is unimportant. Exemplifying all of the permissible errors in the directions for this subtest would also be impractical, because the kinds of errors that were accepted were so numerous. Therefore, such circumstances could reasonably justify the test constructor to maintain the current directions for a subsequent test, use the scoring criteria presented earlier, and reveal to the students what errors were allowed after the test is handed back.

The IDs and IFs for all of the items on the final test are shown in Appendix E. After considering the possibilities presented in the analysis above for a further revision of the test, a subsequent version of the test was created to reflect these modifications. This test appears in Appendix G as the "Test for Future Administration."

3.7.3 Correlations.

This section will examine the correlation coefficients generated by Lertap and discuss any relationships existing between the total test and the subtests. In interpreting correlation coefficients, the first step is to check to see if any of the coefficients are statistically significant, then and only then, can one decide if a correlation is meaningful. The correlations between all subtests and the total test are presented in the table below.

Since the number of test-takers in this study was quite low, a correlation would have to be pretty high in order to be significant. How high? The literature would suggest that a significant correlation for 9 pairs of students would have to be at least .71 (Douglas, 2012, p. 98). In the case of this study, the number of pairs is only 5 and therefore a correlation would have to be even higher to be statistically significant. Of the individual subtests, we might expect a significant correlation between subtests 2 and 4 since they both were intended to assess knowledge of a word's meaning. However, the correlation between these two subtests is only 0.51, and therefore, not statistically significant. Indeed the only potentially significant correlations are those between performance on subtest 1 and the total test and subtest 5 and the total test. It is difficult to infer why these two correlations were the highest. However, subtest 5 certainly required more word knowledge than any other subtest and therefore might be reasoned to be related to a high level of general word knowledge.

The fact that few significant correlations were found for this test, however, is not surprising because although some of those aspects may have overlapped to a degree, each subtest was ultimately designed to test a different aspect of word knowledge. Similar to the findings in this study, Nation (2001) reports that Nist and Olejnik (1995) compared four different vocabulary tests of the same words: one requiring learners to write an illustrative sentence, another involving sentence completion, and others testing meanings and examples. An analysis of the correlations between each of these tests revealed no significant correlations. The lack of significant correlations found in this study suggested that even minor changes to the item format on a vocabulary test could substantially change what the items were measuring. (p. 365). Therefore, it is possible for two tests to focus on the same aspect of word knowledge and still be substantially different due to the way in which they test that knowledge.

Correlations between Subtests and Total Test

Correlations	Stress	Matching	Part of Speech	Definitions	Sentence	Total Test
Stress (subtest 1)	1.00	0.18	0.15	0.62	0.73	0.87
Matching (subtest 2)	0.18	1.00	-0.24	0.51	0.07	0.37
Part of Speech (subtest 3)	0.15	-0.24	1.00	-0.11	0.37	0.41
Definitions (subtest 4)	0.62	0.51	-0.11	1.00	0.42	0.73
Sentence (subtest 5)	0.73	0.07	0.37	0.42	1.00	0.82
Total Test	0.87	0.37	0.41	0.73	0.82	1.00

3.7.4 Validity.

Let us now go back to the item types which appeared in each subtest in order to consider what was required of the learner in order to arrive at an answer. As a first step, this section will analyze each item type using the receptive/productive, recognition/recall, imprecise/precise, and embedded/discrete distinctions mentioned earlier. This analysis should also provide insights into the validity of each subtest.

Let us begin by looking back at the first subtest. Remember that the purpose of this item was to assess the students' abilities to identify the stressed syllable of target vocabulary words. Here is an example of this item type again:

Listen as your teacher reads the following words. Circle the syllable for each word that receives the *most* stress. Your teacher will read each word twice. Each correct answer is worth 1 point.

1. attitudes a—tti—ttudes

A closer inspection of this item reveals that it is probably receptive, since responding to the item involves listening for stress. However, if we consider Nation's complete definition, which states that receptive items involve going from the form of the word to the meaning, then this item cannot really be classified according to this distinction because it does not focus on meaning at all, but rather on an aspect of pronunciation. Nation may not have thought about this when writing the definitions for receptive and productive items because pronunciation is not generally thought of as an aspect of vocabulary knowledge. Additionally, this item cannot be described as either a recognition or a recall item, because it is not assessing meaning, nor is the word being elicited. Clearly, however, this is a precise item because there is only one possible answer with no possibility of demonstrating partial knowledge. Lastly, this item cannot be clearly defined as embedded or discrete. On the one hand this is similar to the multiple-choice item mentioned earlier because it focuses on a specific aspect of the target word (stress) and has only one syllable for the correct response. However, the item is also embedded because it is testing listening.

Now, let us take a look back at an item from the second subtest:

Write the letter next to each word to show its definition. Each correct answer is worth 1 point.

1. nondairy ___ A. not very interesting
B. not made with milk or cream

It is clear that this is a receptive item because answering the item involves going from the form of the word to its meaning and answering the item involves a little bit of reading. It can also be classified as a recognition item because answering the item involves identifying the meaning. Answering the item requires precise knowledge because there is only one clearly acceptable answer that can be chosen. Additionally, this is a traditional discrete item, and this item type has frequently been used to exemplify the discrete distinction in the testing literature because it is testing one specific aspect of vocabulary knowledge (meaning). However, this item also illustrates why it is better to think of the embedded/discrete distinction as a continuum rather than an all or nothing classification because while it can be argued that this item is a

measure of vocabulary knowledge, it still requires a little bit of reading and could, thus, also be said to be testing reading to some extent.

Moving on, we can consider the item below from the third subtest:

Read the sentences below and write the correct part of speech (noun, verb, etc.) for each of the underlined words in the blank. Each correct answer is worth 1 point.

1. The proposed cuts have caused considerable controversy.

_____

This item can be characterized as receptive because it requires reading, but is inconsistent with the distinction as defined by Nation because it does not involve going from the form of the word to the meaning. Engaging with this item does not require consideration of meaning at all. This item fits Read's definition of recognition in the sense that the target word is initially presented to the learner. It is, however, inconsistent with Read's definition because, as just mentioned, this item is not related to meaning. Answering this item requires precise knowledge because only one part of speech can be identified as the correct answer. Lastly, this item is embedded because a sentence context is provided to help the learner identify the part of speech. However, engaging with the context may not be necessary for answering all of the items in this subtest correctly because the learners could potentially use other strategies to arrive at the correct answer. For example, if we consider the word *universally* which also appears in this subtest, it is possible that learners could recognize this as an adverb because they had learned that adverbs commonly end with an -ly suffix. If the learners used this strategy, then it seems that this subtest is also testing knowledge of word parts.

Next,, we can analyze an item from the fourth subtest:

Write a definition for each word from Part 4 based on how it is used in the sentence above.

1. controversy:

Means_____

The categorization of this item as receptive/productive is questionable. On the one hand, this item seems to be productive because the response involves writing. However, the students must also use reading to identify the word, which is an element of the receptive distinction. An additional element of this item that suggests a receptive categorization is that it involves going from the form of the word to its meaning. This item clearly fits Read's recognition definition, because students are shown the word and asked to show that they understand it by writing the meaning. It is an imprecise item because although partial credit is

not given, students still have some flexibility in how they define the word because they are expected to create their own definitions. Thus, some variation in the words chosen to write the definitions can be expected. Lastly, this item is discrete because it does not provide any context and focuses only on the meaning aspect of the word.

Lastly, let us analyze the item type presented in subtest 5:

Write a sentence for each of the following words to show that you know what the word means and how it is used. You may choose a different form of the word than is given if you wish. For example, if the word given is *complicated,* then you may also use *complicate,* or *complication*.

1. attitude:

_____

This item seems to be productive because it involves a considerable amount of writing to give a response. However, it also seems to be receptive because it involves going from the form of the word to its meaning. It is a recognition item because the target word is given and the student is asked to show an understanding of the meaning by writing a sentence. It is also imprecise because partial credit is given for sentences with grammatical errors. Lastly, this item seems to be embedded in the sense that the item is testing the ability to use the word through sentence writing and with attention to grammar. However, this item also is discrete in the sense that there is no context given for the item – that is to say, the word is presented in isolation.

The analysis of these items according to the four distinctions presented in the literature review will have hopefully provided some insights into the validity of each subtest by considering whether or not each subtest was measuring what it was designed to measure. From the above analysis, it should be clear that careful consideration was given about what the items are actually assessing in order to make accurate judgments about the learners.

In addition to the underpinnings of the items for each subtest, hopefully this analysis has also revealed additional insights that should be valuable when making judgments about vocabulary items. Just as Read has problematized the receptive/productive distinctions with reference to vocabulary assessments, it seems that three of four classifications used to describe vocabulary items can also be problematic when the items being analyzed are less conventional than the traditional matching format. Just as Read has noted that [the testing of all language skills depends on vocabulary to some extent] (Read, 2012b, p. 307) from this item analysis, one could just as easily argue that testing vocabulary involves the testing of other language skills to some extent. Additionally, this analysis reveals a need for further clarifications in the definitions of terminology used to describe vocabulary assessment items.

4.0 Conclusion and Discussion

This paper began by discussing the concepts *vocabulary* and *vocabulary knowledge* and reviewing many recommendations that have been made for constructing and piloting vocabulary tests. The description above regarding the process of constructing and analyzing the results of a test construction project that followed many of these recommendations also showed that many of these recommendations can facilitate the process of creating a vocabulary test with desirable validity and reliability.

The remainder of this paper will discuss the limitations of the test construction project, pedagogical implications of the study, and suggestions for future administrations of this vocabulary assessment instrument.

4.1 Limitations

While this study revealed some evidence of the final test's validity and reliability, there are a few limitations that must be considered when interpreting these findings.

Although the scores on the final tests were used to make decisions about students' achievement of certain vocabulary objectives, evidence could have been collected prior to testing to ensure that the vocabulary objectives on the final test were appropriate for these learners. Since students were not pretested on the vocabulary for this test there is no way of knowing what aspects of vocabulary knowledge the students knew before taking the course, or what aspects of vocabulary knowledge students learned during the course. Pretesting students on the target vocabulary and vocabulary objectives at the beginning of the course would have provided data that could reveal the students' initial vocabulary knowledge in order to confirm that the vocabulary objectives selected were relevant to the needs of the learners. Since the pilot and final tests were both developed weeks after the course began, however, this was logistically impossible.

Additionally, the scoring procedures and scoring criteria for some of the subtests may have inflated their reliability coefficients. For example, it has often been noted in the testing literature that having a single rater tends to inflate the reliability of a test (which was the case in this study). Because of this it would have been desirable to establish interrater reliability to make these test scores more objective. To do this multiple raters would all need to be in agreement about the test and all of its elements in order to score it consistently. If these raters arrived at consistent scores, this would suggest greater reliability for the test. However, it seems unlikely that having additional raters would alter the scores for the first three subtests

because the format of these items was extremely objective. Different methods for developing the scoring criteria for subtests 4 and 5 could also lead to greater reliability. Since the scoring criteria for these subtests were developed by a single test constructor/scorer, the results of these subtests may have been influenced by personal biases. For example, the definition of partial-credit responses for subtest 4 and "meaningful errors" for subtest 5 were based on the opinion of just one person. By collaborating with other instructors in order to further develop these scoring criteria, these criteria could be made more objective.

Also, the validity of some of the subtests is questionable because this study was unable to confirm whether the students memorized answers to test questions from their vocabulary charts or whether they had actually attained the necessary skills to arrive at a correct answer. For example, if students had completed all of their vocabulary charts, they would have already underlined the stressed syllable for words in subtest 1. Although it would have required a lot of memorization, since the students did not know which words out of all of the words studied would appear on each subtest, students could still have indicated a correct answer for subtest 1 by memorizing the answers in their chart instead of through listening. Likewise, students could have memorized sentences that were presented in class for subtest 5 and have written them as answers on the test instead of developing them originally. Use of memorization for this task would not indicate the ability to use the target words.

Lastly, since the number of students who took the final test was only 11, claims for the test's validity and reliability must be interpreted cautiously. According to Read, (2012b) claims for validity and reliability require a sample size of 30 or more. Such a large sample size was not possible in this testing situation because I only had access to the students in my own class for the final test.

4.2 Pedagogical Implications

Several pedagogical implications were revealed through this study.

First of all, creating "tally sheets" like the ones shown in this paper might prove useful to classroom teachers who want to revise a test, but who do not have the time to revise and analyze it as thoroughly as in this Double Plan B project. As the results of the final test revealed, using the tally sheets to "eye ball" the IFs of the items on the pilot most certainly contributed to the final test's reliability. This method can be useful for quickly identifying items that are obviously too easy, such as the ones that all students answered correctly, without performing any calculations. This recommendation is not meant, however, to underestimate the value of considering ID. In order to create the best quality tests, a thorough analysis of both IFs and IDs for items will probably be necessary. However, if items are answered correctly by all learners, the tally sheet will also reveal the ID for these items because any item that is answered

correctly by all learners will always have an ID of 0.00. Use of a tally sheet could be a way to substantially improve the quality of a test with minimal effort.

Secondly, organizing the pool of words that you would potentially like to test in a list like the one in Appendix A can facilitate the construction of some common vocabulary items and make it easier to follow many of the recommendations mentioned in the literature review of this paper (see 2.0) that can improve the quality of a vocabulary test. For example, if the items are taken from a textbook, grouping them by unit will help the test constructor to ensure that a representative sample of each of those words comes from each unit if testing all of them is impractical. Including the definitions on a word list will obviously help in the construction of matching items, and including the parts of speech for each word will make it easy to create blocks that include words that are all the same part of speech. Once the test constructor has decided the format for the items on the test, some indication of what subtest each word will appear on could be noted on the word list. This will help the test constructor to keep track of how many times a word appears on the test so that priority is not given to some words over others.

Additionally, the experiences reported concerning more context-dependant items such as those in subtests 4 and 5 reveal that several challenges can result from choosing these formats. These items are clearly more difficult to score than more discrete-item formats. These subtests also show that using less discrete-item formats can make it more challenging to interpret resulting scores because the constructs assessed by these items will be more difficult to define. When selecting item formats such as these, teachers may want to consider if an item-type with greater practicality can also be used to assess the same, or similar, constructs. If a more practical item format is not possible, teachers must be prepared to work with carefully defined scoring criteria to ensure the reliability of these items.

One item format that might be especially useful for teachers of a listening/speaking course is the word-stress format that appeared in subtest 1 of the tests in this study (see 3.2.1). English is different from many other languages because the meaning of a spoken word can be dependant on which syllable is stressed. For example, *construct* can be either a noun or a verb in spoken English depending on whether the first or second syllable receives stress. Learners need to be aware of stress in order comprehend and use spoken English. By making students more aware of stress in the vocabulary that they encountered frequently in context, through lectures and other listening texts mentioned earlier, my goal was to help students interpret the listening texts presented in the course more meaningfully and also help them develop an awareness of stress that would transfer to the learning of new words outside the classroom. Learning to comprehend also naturally precedes the ability to use, and although it was not assessed in the vocabulary achievement test, another goal of listening for stress was to bridge the gap between reception and production so that learners could eventually produce the words

they studied accurately. Unfortunately, a limited amount of class time prevented this from becoming a major component of instruction in the course.

The ease of constructing and scoring the word-stress item type also shows its high level of practicality. As the data showed, this item can also be extremely reliable. When constructing these items, it is recommended that teachers try to include words with a large number of syllables whenever possible in order to increase reliability.

Lastly, use of a vocabulary chart, such as the one described in this study (see 3.1), revealed that this can be a valuable tool for teaching aspects of word knowledge that extend beyond definitions. This activity is helpful for getting students to focus on several aspects of target vocabulary and can also make vocabulary learning student-centered and communicative. During the "teaching" phase of this activity, students can also be encouraged to correct their peers on pronunciation and grammar. In this way, this activity can also be meaningful for other course objectives as well, especially in a speaking and listening course. For a more detailed description of using the vocabulary chart activity described in this paper see Appendix H.

4.3 Suggestions for Future Test Administration

If this test were given again, several other procedures could be taken to provide additional insights about the testing instrument in an effort to improve it further.

Verbal reports could be used with students to find out more about how they arrive at answers to each of the items. This could be accomplished either by asking the students to write down why they chose certain answers for items on the test, or by conducting an oral interview with students after the test. Verbal report data might be especially helpful for increasing the reliability of subtest 2 because this information could be used to improve the distractors. Verbal report data could also help to evaluate the validity of subtest 3. If students reported that they used the sentence context to identify the parts of speech for these items, then this would indicate that this subtest is assessing the knowledge that it was designed to assess. However, if most of the verbal report data revealed that students were using other strategies, like knowledge of prefixes to correctly answer these items, this would indicate that subtest 3 is not functioning properly.

Another approach to improving subtest 2 would be to ask a pilot group to write their own definitions for the words in subtest 2. If many of these students gave a similar incorrect definition for any of the target vocabulary words, these definitions might be used as high-quality distractors since they would suggest a common misunderstanding that could potentially be the same for the final group of test-takers. Distractors created using this method might improve the reliability of this subtest since it would be more likely that only the students with the highest abilities would get them correct.

An alternative method could also be used for designing the scoring criteria for subtest 5, regarding acceptable grammatical mistakes. Since this test was intended for students in an intensive English program, grammar objectives from lower level grammar courses could be used to make decisions about acceptable mistakes. By examining the pedagogy and working with teachers from lower-level courses, acceptable grammatical mistakes could be established for subtest 5 based on what students are expected to master before moving to level 3 in the MELP program.

Since this test was designed to measure word knowledge in a speaking and listening course, alternative item formats could also be designed that would require the students to use speaking and listening to answer the items. For example, the words in subtest 2 (matching) could be presented via audio in the format shown below.

1. __ A. not very interesting
B. not made with milk or cream

Although the directions for this subtest would essentially be the same, answering questions using this format would require the students to recognize vocabulary words through listening before selecting the definitions and therefore make the task more listening-focused. Another possible item format for the subsequent test could require the students to create an audio recording in order to add a speaking element to the test. Before these additional item-types could responsibly be added to a subsequent version of the test, however, they would need to be piloted before they were used to make decisions about students.

Lastly, an approach that could provide some interesting insights would be to set up an *intervention study*. While this concept is discussed in depth in Brown (2005), the basic procedure would be to administer the test at the beginning of the course and analyze the items which the students answered incorrectly. In an ideal situation, the IFs of the items would all be 0.00, indicating that the students needed to acquire the skills necessary to answer those items correctly. The results of the pretest would then be used to determine the instruction that would be necessary for students to acquire those vocabulary skills. The instructor would provide the necessary instruction before administering the test again. A comparison of the IFs of the items on the first test with the items on the second test could be used to evaluate how effective the teacher's instruction was. If the IFs on the pretest were very low and the IFs on the final were very high, an inference could be made that students had acquired the necessary language skills and that this was due to high quality instruction, although other factors would still need to be considered.

Clearly there are many issues to consider when constructing and piloting a vocabulary assessment. While numerous issues have been discussed in this paper, it is likely that other issues will arise under different circumstances and with different types of items. However, the

construction of most vocabulary tests will likely involve consideration of many, if not all, of the questions mentioned earlier. Hopefully, the insights presented in this paper have provided a sense of how to improve the reliability and validity of vocabulary tests and will be useful for other vocabulary test construction projects.

References Cited

Beglar, D., & Murray, N. (2009). *Contemporary topics 3*. (3rd ed.). White Plains, NY: Pearson Education.

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

Cohen, A. D. (1994). *Assessing language ability in the classroom*. 2nd ed. Boston, MA: Newbury House/Heinle & Heinle.

Douglas, D. (2012). *Understanding language testing*. Abingdon, England: Hodder Education.

Lertap (2003-2011). *Laboratory of Educational Research Test Analysis Package*. 5th version. Curtin University, West Australia. Distributed by Assessment Systems Corporation, St. Paul, MN. <http://www.assess.com/xcart/product.php?productid=3>

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching, 44*(4), 529-539.

Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.

Read, J. (2012a). Assessing vocabulary. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 257-263). Cambridge, England: Cambridge University Press.

Read, J. (2012b). Piloting vocabulary tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 307-320). New York, NY: Routledge.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, England: Cambridge University Press.

Zimmerman, C. B. (2009). *Word knowledge: A vocabulary teacher's handbook*. Oxford, England: Oxford University Press.

Appendix A - Initial Pool of Vocabulary

Unit 1 Vocabulary

1. attitudes: (n.) opinions and feelings
2. constantly: (adv.) all the time
3. construct: (v.) create or build
4. evolving: (v.) developing
5. expanding: (v.) increasing
6. identity (n.) a quality that makes someone distinct
7. inevitable: (adj.) unavoidable
8. phenomenon: (n.) a remarkable thing
9. reinforces: (v.) supports
10. widespread: (adj.) common

Unit 2 Vocabulary

1. alert: (adj.) watching and listening carefully
2. aptitude: (n.) a natural ability or skill
3. devoted: (v.) given time and perhaps money to some activity
4. exhibit: (v.) show something so that it's easy to notice
5. imagination: (n.) the ability to form creative ideas in your mind
6. inconsistencies: (n.) two or more pieces of information that do not agree with one another
7. motivation: (n.) eagerness and willingness to do something
8. predominant: (adj.) more powerful than others
9. strategy: (n.) a plan used to achieve a goal
10. underlying: (adj.) hidden and not easy to discover

Unit 4 Vocabulary

1. converging: (v.) becoming similar
2. enables: (v.) makes possible
3. entire: (adj.) complete
4. guaranteed: (v.) certain
5. homogeneous: (adj.) similar
6. ideological: (adj.) based on beliefs or ideas
7. media: (n.) TV, radio, and newspapers
8. promote: (v.) draw support for
9. universal: (adj.) worldwide
10. verbal: (adj.) spoken

Unit 6 Vocabulary

1. anthropologists: (n) people who study societies and their beliefs
2. attachment: (n.) a feeling of strong loyalty, love, or friendship to a person or thing
3. characteristic: (n.) a special quality or feature that someone or something has
4. emotion: (n.) a strong human feeling such as love or hate
5. enhance: (v.) make better
6. hormone: (n.) a substance in our body that influences our growth, development, and health
7. invoke: (v.) cause something to happen
8. mutual: (adj.) feeling or action that is felt or done by two or more people toward one another
9. prospective: (adj.) likely to do a particular thing or that an event is likely to happen
10. romantic: (adj.) related to the concept of love

Unit 8 Vocabulary

1. civil liberties: (n.) rights and freedoms people have in society
2. commercial: (adj.) having to do with business or trade
3. composite: (adj.) made up of different parts
4. controversial: (adj.) causing a lot of disagreement
5. deterrent: (n.) something that prevents people from doing something
6. security: (n.) measures taken by agencies to protect us
7. sophisticated (adj.) advanced or complex
8. suspected: (adj.) believed to be responsible for doing something wrong
9. techniques: (n.) methods or ways of doing something
10. via: (prep.) through, by, or by way of

Unit 12 Vocabulary

1. adulthood: (n.) the period of life when a person is completely grown
2. confirmed: (v.) determined that something is definitely true
3. interracial: (adj.) between different races of people
4. legitimacy: (n.) acceptance; validity
5. matured: (v.) become fully developed physically or emotionally; grown up
6. nationalistic barriers: (n.) a block people face because of their national beliefs
7. norm: (n.) the usual or acceptable way of doing something
8. population: (n.) people in a particular area or members of a particular group
9. pregnant: (adj.) when a woman is carrying an unborn offspring in her body
10. social class: (n.) a group of people with a similar rank in society

Unit 9 Vocabulary

1. accent: (n.) the way a person pronounces words
2. discrete: (adj.) when ideas or things are separate from each other
3. distinct: (adj.) clearly different or separate
4. flexible: (adj.) able to change or be changed easily
5. generation: (n.) a group of people born around the same time
6. impressive: (adj.) something that makes a strong impression or causes admiration
7. nonverbal behavior: (n.) expressing meaning without words
8. precise: (adj.) exact
9. random: (adj.) existing in a way that seems to be without reason; unpredictable
10. ultimately: (adv.) in the end; finally

Appendix B - The Pilot Version of the Test

Note: Items marked with an * were omitted from the final test.

Name:_____ Score:_____/125

Contemporary Topics 3 Vocabulary Practice Test

Part 1: Identifying the stressed syllable (20 points)

Listen as your teacher reads the following words. Circle the syllable for each word that receives the *most* stress. Your teacher will read each word twice. Each correct answer is worth 1 point.

Example: 1. money mo—**ney**

Note: Answers here appear in bold, although students were expected to circle the answers on the actual test.

Item	IF
1	0.60
3	0.50
5	0.40
7	0.40
9	0.50
*11	0.70
13	0.40
*15	0.80
17	0.40
19	0.30

1. attitudes	**a**--tti--ttudes	2. constantly	**con**--stant--ly
3. expanding	ex--**pand**--ing	4. imagination	i--ma--gi--**na**--tion
5. inconsistencies	in—con—**sis**—ten—cies	6. strategy	**strat**--e--gy
7. guaranteed	guar--an--**teed**	8. promote	pro--**mote**
9. universal	u--ni--**ver**--sal	10. attachment	a--**ttach**--ment
11. invoke	in--**voke**	12. prospective	pro--**spec**--tive
13. controversial	con--tro--**ver**--sial	14. suspected	su--**spect**--ed
15. accent	**ac**--cent	16. discrete	dis--**crete**
17. impressive	im--**pre**--ssive	18. interracial	in--ter--**ra**--cial
19. legitimacy	le--**gi**--ti--ma--cy	20. population	po--pu--**la**--tion

Item	IF
*2	0.70
4	0.50
*6	0.80
*8	0.70
*10	0.80
*12	0.70
*14	0.90
16	0.60
18	0.60
20	0.50

Part 2: Matching the words with definitions (30 points)

Write the letter next to each word to show its definition. Each correct answer is worth 1 point.

Example: 1. nondairy B

A. not very interesting
B. not made with milk or cream

Matching Part A

1. identity	J	A. a natural ability or skill
2. phenomenon	M	B. a society in an advanced state of development
3. aptitude	A	C. the usual or acceptable way of doing something
4. motivation	N	D. a person who studies living things
5. anthropologist	O	E. the period of life when a person is completely grown
6. hormone	G	F. something that prevents people from doing something
7. deterrent	F	G. a substance that influences growth and development
8. generation	P	H. the chemical process that changes food into energy
9. adulthood	E	I. readiness to take action
10. norm	C	J. a quality that makes someone distinct
		K. a source of supply or support
		L. a form of transport
		M. a remarkable thing
		N. eagerness or willingness to do something
		O. a person who studies societies and their beliefs
		P. a group of people born around the same time

Item	IF
21	0.60
22	0.30
23	0.40
24	0.40
25	0.80
26	0.50
27	0.10
*28	0.90
*29	1.00
30	0.50

Note: **I** was another potential answer choice for number 4 (*motivation*).

Matching Part B

1. construct	I	A. become similar
2. evolve	F	B. make possible
3. reinforce	J	C. gradually increase in quantity or size
4. devote	O	D. experience or go through
5. exhibit	L	E. make better
6. converge	A	F. develop
7. enable	B	G. keep
8. enhance	E	H. determine that something is definitely true
9. confirm	H	I. create or build
10. mature	N	J. support
		K. locate
		L. show something so that it's easy to notice
		M. stop doing something
		N. develop physically or emotionally
		O. give time and perhaps money to some activity
		P. increase the speed of something

Item	IF
*31	0.70
32	0.30
33	0.10
34	0.40
35	0.50
36	0.40
*37	0.90
38	0.30
39	0.60
40	0.40

Note: **F** was another potential answer choice for number 10 (*mature*

Item	IF
41	0.50
42	0.70
43	0.60
*44	1.00
45	0.30
46	0.70
47	0.20
48	0.40
*49	0.80
*50	1.00

Matching Part C

1. inevitable	**F**	A. able to continue
2. alert	**J**	B. made up of different parts
3. underlying	**P**	C. similar
4. homogeneous	**C**	D. not objective
5. ideological	**L**	E. absolutely necessary
6. commercial	**O**	F. unavoidable
7. composite	**B**	G. clearly different or separate
8. distinct	**G**	H. separate from physical realities
9. flexible	**N**	I. undependable
10. pregnant	**M**	J. watching and listening carefully

K. official
L. based on beliefs or ideas
M. when a woman is carrying an unborn offspring in her body
N. able to change or be easily changed
O. having to do with business or trade
P. hidden and not easy to discover

Part 3: Identify parts of speech (15 points)

Read the sentences below and write the correct part of speech (noun, verb, etc.) for each of the underlined words in the blank. Each correct answer is worth 1 point.

Item	IF
*51	0.90
52	0.50
53	1.00
*54	0.80
55	1.00
56	1.00
*57	0.80
58	0.70
*59	0.80
60	0.70

Example: **1. Jim is unreliable because he never keeps a promise. adjective**

1. They have made their fortunes from industry and commerce. __**noun**_____

2. The proposed cuts have caused considerable controversy. ____**noun**_____

3. Supporters of the death penalty argue that it would deter criminals from carrying guns. ____**verb**_____

4. His government is looking distinctly shaky. _____**adverb**_____

5. He said it is still not possible to predict the ultimate outcome. ____**adjective**_____

6. The flexibility of the lens decreases with age; it is therefore common for our sight to worsen as we get older. ____**noun**_____

7. They called John an enabler because he bought beer for the young boys. ____**noun**_____

8. This administration is not entirely free from suspicion. ____**adverb**_____

9. The editors can give no guarantee that they will fulfill their obligations. _____**noun**_____

10. The disadvantage is that it is not universally available. _____**adverb**_____

Item	IF
61	0.80
62	0.40
63	0.50
*64	0.90
65	0.80

11. George was unable to verbalize how he felt about a girl in his class. ____**verb**_____

12. This program is not a legitimate use of the taxpayers' money. _____**adjective**_____

13. Humans experience a delayed maturity; we arrive at all stages of life later than other mammals.

____**noun**_____

14. I received a confirmation by email after I registered for the test. ____**noun**_____

15. All airports in the country are working normally today. ____**adverb**_____

Part 4: Write definitions (30 points)

Write a definition for each word from Part 3 based on how it is used in the sentence above.

Example: 1. unreliable : Means not dependable.

You will not receive points for answers that use another form of the word.

Example: 2. unreliable: Means not reliable.

Each correct answer is worth 2 points.

Note: Answers for this section will vary, but possible answers are shown below.

Item	IF
*66	0.30
67	0.40
68	0.10
*69	0.20
70	0.20
71	0.60
*72	0.10
73	0.40
*74	0.40
75	0.70
76	0.50

1. commerce: Means____**business; trade**_____

2. controversy: Means____**debate; disagreement; argument**_____

3. deter: Means____**prevent; stop; hinder**_____

4 distinctly: Means____**particularly; especially; clearly; unmistakenly**_____

5. ultimate: Means____**final**_____

6. flexibility: Means____**ability to change easily; pliability**_____

7. enabler: Means____**someone who makes something possible**_____

8. entirely: Means____**completely; totally**_____

9. guarantee: Means____**assurance**_____

10. universally: Means____**globally**_____

11. verbalize: Means____**say; put into words**_____

Item	IF
77	0.20
78	0.50
*79	0.60
80	0.60

12. legitimate: Means____**appropriate; acceptable; valid**_____

13. maturity: Means____**development; growth**_____

14. confirmation: Means____**certification; affirmation**_____

15. normally: Means____**as usual**_____

Part 5: Sentence writing (30 points)

Write a sentence for each of the following words to show that you know what the word means and how it is used. You may choose a different form of the word than is given if you wish. For example, if the word given is *complicated,* then you may also use *complicate,* or *complication*.

A sentence that shows that you know the meaning of the word and can use correct grammar is worth full credit (2 points)

Example: 1. complicated: The math problem was really *complicated* so I didn't get the right answer.

A sentence that shows that you know the meaning of the word, but uses incorrect grammar is worth half credit (1 point).

Example: 2. complicated: The math problem was really *complication* so I didn't get the right answer.

A sentence that does not show that you know the meaning of the word receives no credit (0 points).

Example: 3. complicated: It is *complicated*.

Item	IF
81	0.80
82	0.30
*83	0.80
84	0.80

1. attitudes: **People have many *attitudes* about abortion because it is such a controversial topic.**_____

2.constantly: **When I was young, I hated my sister because she *constantly* bothered me.**_____

3. imagination **A person with a good imagination can think very creatively.**_____

4. strategy: **Tell me a *strategy* that you use for learning English so we can help others improve too.**_____

5.verbal: ***Verbal*** **communication is just a fancy name for speaking.**

6.characteristic: **One *characteristic* of my friend Sean is that he drinks a lot.**

7. romantic **My wife thought I was so *romantic* when I proposed to her on our anniversary.**

8. security: **I don't like the *security* cameras on campus because I feel like I'm being watched.**

9. sophisticated: **Computers are so *sophisticated*. I don't think I'll ever fully understand how they work.**

10 .via: **I got a message *via* email that said a famous rock band is going to play on my birthday.**

11. precise: **I can't remember the *precise* location of a sushi restaurant I visited in San Diego because its been so long since I went there.**

12. random **The teacher assigned groups at *random* by having us draw cards.**

13. ultimately: **The world will *ultimately* end, but no one knows when.**

14.interracial: **Many marriages in the US today are *interracial* which is why seeing black and white couples is not uncommon.**

15. population: **The *population* of New York City is the largest of any city in the US.**

Item	IF
85	0.60
86	0.70
*87	1.00
88	0.90
89	0.40
90	0.60
*91	0.00
*92	0.40
93	0.20
94	0.50
*95	1.00

Appendix C - Tally Sheets

Students' Answers for Subtest 1

Item	Correct	Incorrect
1	6	4
2	7	3
3	5	5
4	5	5
5	4	6
6	8	2
7	4	6
8	7	3
9	5	5
10	8	2
11	7	3
12	7	3
13	4	6
14	9	1
15	8	2
16	6	4
17	4	6
18	6	4
19	3	7
20	5	5

Students' Answers for Subtest 2

Item	Correct	Incorrect
21	6	4
22	3	7
23	4	6
24	4	6
25	8	2
26	5	5
27	1	9
28	9	1
29	10	0
30	5	5
31	7	3
32	3	7

33	1	9
34	4	6
35	5	5
36	4	6
37	9	1
38	3	7
39	6	4
40	4	6
41	5	5
42	7	3
43	6	4
44	10	0
45	3	7
46	7	3
47	2	8
48	4	6
49	8	2
50	10	0

Students' Answers for Subtest 3

Item	Correct	Incorrect
51	9	1
52	5	5
53	10	0
54	8	2
55	10	0
56	10	10
57	8	2
58	7	3
59	8	2
60	7	3
61	8	2
62	4	6
63	5	5
64	9	1
65	8	2

Students' Answers for Subtest 4

Item	Full Credit	Partial Credit	No Credit	No Attempt
66	1	2	3	4
67	2	2	2	4
68	1	0	3	6
69	0	2	3	5
70	2	0	1	7
71	2	4	2	2
72	1	0	4	5
73	1	3	3	3
74	1	3	5	1
75	4	3	3	0
76	5	0	3	2
77	1	1	3	4
78	0	5	2	3
79	1	5	2	2
80	5	1	3	1

Students' Answers for Subtest 5

Item	Full Credit	Partial Credit	No Credit	No Attempt
81	4	4	1	1
82	0	3	3	4
83	4	4	2	0
84	1	7	1	1
85	1	5	2	2
86	4	3	2	1
87	2	8	0	0
88	1	8	1	0
89	1	3	0	6
90	2	4	2	2
91	0	0	2	8
92	0	4	3	3
93	0	2	1	7
94	0	5	0	5
95	6	4	0	0

Appendix D - Peer Critique

Review of Stephen Kis' Pilot Test by Dongming Yang

There are a few things that might be worth reconsideration in the test. For the first part, I'm not sure whether it's necessary to break down the syllables for students. Since these are low advanced students, they may already be well familiar with counting syllables by themselves. A good way to increase the difficulty level of a test is not simply including more items but adding the workload of an individual item. Unless students are not yet trained to analyze syllables, they might be challenged more appropriately if the words are just laid out as a whole.

Part 3 asks students to identify parts of speech in sentences, which gives them some contexts to draw on and a good amount of reading to apply relevant rules to practice. Among 15 target words that are tested, there are 7 nouns, 4 adverbs, 2 adjectives and 2 verbs. The composition of different parts of speech is out of balance, so I don't know whether this shows the test is designed to give more emphases on nouns. This causes my concern about the test reliability. Does it give rise to the possibility that students well acquainted with the rules to distinguish nouns have a good chance of passing without knowing much about the other three parts of speech? I suggest there be equal balance among all four of them.

The test flows consistently from part 3 to part 4 which asks students to give definitions for the words that appear in the preceding section. Several concerns rose as I try out the test myself. At the first sight, I think both examples given in the direction are correct. To make it more straightforward, "example 1" and "example 2" could be phrased as "acceptable answer" and "unacceptable answer" respectively. My second question is related to the direction, "Write a definition…based on how it is used in the sentence above." What does it mean by "how it is used"? Are students supposed to take into account the part of speech, which is the tested skill area in part 3, when they come up with the definitions? For example, "guarantee" is used as a noun in the sample sentence before. In this case, does one still get a full credit if he defines "guarantee" as "assure"? Similarly, is "complete" an acceptable answer for "entirely"? What about defining "controversy" as "some say yes while others say no"? Another worry comes from the illustration for erroneous example, where it states that "You will not receive points for answers that use another form of the word". According to the Merriam-Webster dictionary, an "enabler" is "one that enables another to achieve an end". A "confirmation" is "an act of process of confirming". Both standard definitions actually include another form of the word. So, it might be helpful to provide a third example which tells the difference from the second one that purely uses a different form of the word without supplying extra information. In this way, it's more effective to avoid the use of test-wise strategies and still assure students to construct the definitions with another form of the word as a part of it.

Likewise, there's some ambiguity in the directions for part 5 as well. The examples are well-chosen and self-explanatory, but how to count a sentence as "incorrect grammar" is still confusing to me. What if students write something that includes the correct use of the target word but other kinds of grammar mistakes? For example, "Since the math problems is really complicated, so I didn't get the right answers." Also, why is the first item written in the plural form? Does it indicate that one should use no single form, or is it just a typing mistake?

Overall, this is a level-appropriate test on vocabulary that includes integrative and discrete-point items. It follows a pronunciation (stress) - part of speech - definition - usage format, which is reasonable and reliable as it meets students' learning pattern and expectation of the test. In addition, all the vocabulary knowledge tested seems to be covered in the textbook and the words are taken from the textbook as well. Thus, the test serves a good assessment purpose for the semester-long course. It might be helpful spell out some directions and examples more clearly to prevent any possible misunderstanding from students before administering the test. Or at least, the test designer can think them through and have a standard grading rubric in mind for certain parts which are not merely a matter of right of wrong. I think this is a good test that ties well in with the course curriculum, and thus test-taker friendly.

Appendix E – The Final Version of the Test

Name:_____ Score:_____/85

Contemporary Topics 3 Vocabulary Test

Note: Items marked with an * were omitted from the test for future administration.

Part 1: Identifying the stressed syllable (12 points)

Listen as your teacher reads the following words. Circle the syllable for each word that receives the *most* stress. Your teacher will read each word twice. Each correct answer is worth 1 point.

Example: 1. money (mo)—ney

Item	IF	ID
*1	1.00	0.00
*2	1.00	0.00
3	0.73	0.68
4	0.82	0.46
5	0.82	-0.01
6	0.73	0.24
7	0.55	0.68
8	0.82	0.80
9	0.73	0.73
10	0.73	0.74
11	0.64	0.60
12	0.82	0.22

Note: Answers here appear in bold, although students were expected to circle the answers on the actual test.

1. attitudes **a**—tti—ttudes
2. expanding ex—**pand**—ing
3. imagination i—ma—gi—**na**—tion
4. inconsistencies in—con—**sis**—ten—cies
5. guaranteed guar—an—**teed**
6. universal u—ni—**ver**—sal
7. controversial con—tro—**ver**—sial
8. discrete dis—**crete**
9. impressive im—**pre**—ssive
10. interracial in—ter—**ra**—cial
11. legitimacy le—**gi**—ti—ma—cy
12. population po—pu—**la**—tion

Part 2: Matching the words with definitions (23 points)

Write the letter next to each word to show its definition. Each correct answer is worth 1 point.

Example: **1. nondairy** **B**

A. not very interesting
B. not made with milk or cream

Item	IF	ID
13	0.91	-0.35
14	0.82	-0.23
15	1.00	0.00
16	1.00	0.00
17	1.00	0.00
18	0.73	0.33
19	0.82	0.39
20	1.00	0.00

Matching Part A

1. anthropologist **E**
2. phenomenon **M**
3. aptitude **A**
4. motivation **I**
5. identity **J**
6. hormone **G**
7. deterrent **F**
8. norm **C**

A. a natural ability or skill

B. a society in an advanced state of development

C. the usual or acceptable way of doing something

D. a person who studies living things

E. a person who studies societies and their beliefs

F. something that prevents people from doing something

G. a substance that influences growth and development

H. the chemical process that changes food into energy

I. eagerness or willingness to do something

J. a quality that makes someone distinct

K. a source of supply or support

L. a form of transport

M. a remarkable thing

Item	IF	ID
21	1.00	0.00
22	0.82	-0.23
*23	0.09	-0.02
24	0.73	0.40
25	0.73	-0.15
26	1.00	0.00
27	0.91	0.56
28	0.64	0.11

Matching Part B

1. construct **I**
2. evolve **F**
3. mature *****
4. devote **J**
5. exhibit **L**
6. converge **A**
7. confirm **H**
8. enhance **E**

A. become similar

B. increase the speed of something

C. gradually increase in quantity or size

D. experience or go through

E. make better

F. develop

G. keep

H. determine that something is definitely true

I. create or build

J. give time and perhaps money to some activity

K. locate

L. show something so that it's easy to notice

M. stop doing something

Note: The correct answer for number 3 (*mature)* was mistakenly omitted from the list of options.

Item	IF	ID
29	0.82	0.31
30	0.91	-0.35
31	1.00	0.00
32	0.91	-0.04
33	1.00	0.00
34	1.00	0.00
35	1.00	0.00

Matching Part C

1. inevitable **F**
2. alert **J**
3. underlying **M**
4. distinct **G**
5. ideological **L**
6. commercial **C**
7. composite **B**

A. able to continue

B. made up of different parts

C. having to do with business or trade

D. not objective

E. absolutely necessary

F. unavoidable

G. clearly different or separate

H. separate from physical realities

I. undependable

J. watching and listening carefully

K. official

L. based on beliefs or ideas

M. hidden and not easy to discover

Part 3: Identify parts of speech (10 points)

Read the sentences below and write the correct part of speech (noun, verb, etc.) for each of the underlined words in the blank. Each correct answer is worth 1 point.

Item	IF	ID
36	0.73	0.07
37	1.00	0.00
38	0.91	-0.14
39	0.91	0.48
40	0.55	0.13
41	0.64	0.27
42	0.91	0.48
43	0.73	0.18
44	0.91	0.48
45	0.82	0.30

Example: 1. Jim is unreliable because he never keeps a promise. adjective

1. The proposed cuts have caused considerable controversy. **noun**_____

2. Supporters of the death penalty argue that it would deter criminals from carrying guns.
verb_____

3. He said it is still not possible to predict the ultimate outcome. **adjective**_____

4. The flexibility of the lens decreases with age; it is therefore common for our sight to worsen as we get older.

noun_____

5. This administration is not entirely free from suspicion. **adverb**_____

6. The disadvantage is that it is not universally available. **adverb**_____

7. George was unable to verbalize how he felt about a girl in his class.
verb_____

8. This program is not a legitimate use of the taxpayers' money. **adjective**_____

9. Humans experience a delayed maturity; we arrive at all stages of life later than other mammals.

noun_____

10. All airports in the country are working normally today. **adverb**_____

Part 4: Write definitions (20 points)

Write a definition for each word from Part 3 based on how it is used in the sentence above.

Acceptable Answer: unreliable : Means not dependable.

You will not receive points for answers that use another form of the word.

Unacceptable Answer: unreliable: Means not reliable.

Each correct answer is worth 2 points.

Note: Answers for Part 4 will vary, but possible answers are shown below.

Item	IF	ID
46	0.77	0.67
47	0.36	0.35
48	0.77	0.66
49	0.55	0.34
50	0.77	0.37
51	0.73	0.15
52	0.64	0.35
53	0.45	0.05
54	0.64	0.48
55	0.64	0.10

1. controversy: Means____**debate; disagreement; argument**

2. deter: Means____**prevent; stop; hinder**

3. ultimate: Means____**final**

4. flexibility: Means____**ability to change easily; pliability**

5. entirely: Means____**completely; totally**

6. universally: Means____**globally**

7. verbalize: Means_____**say; put into words**

8. legitimate: Means_____**appropriate; acceptable; valid**

9. maturity: Means_____**development; growth**

10. normally: Means_____**as usual**

Part 5: Sentence writing (20 points)

Write a sentence for each of the following words to show that you know what the word means and how it is used. You may choose a different form of the word than is given if you wish. For example, if the word given is *complicated,* then you may also use *complicate,* or *complication*.

A sentence that shows that you know the meaning of the word and can use correct grammar is worth full credit (2 points)

Full Credit Answer: complicated**:** The math problem was really *complicated* so I didn't get the right answer.

A sentence that shows that you know the meaning of the word, but uses incorrect grammar is worth half credit (1 point).

Half Credit Answer: complicated: The math problem was really *complication* so I didn't get the right answer.

A sentence that does not show that you know the meaning of the word receives no credit (0 points).

No Credit Answer: complicated: It is *complicated*.

Item	IF	ID
56	0.55	0.05
*57	0.95	-0.12
58	0.91	0.38
59	0.77	0.65
60	0.77	0.18
*61	0.86	0.00

Note: Answers for Part 5 will vary, but possible answers are shown below.

1. attitude: **People have many *attitudes* about abortion because it is such a controversial topic.**

2.constantly: **When I was young, I hated my sister because she *constantly* bothered me.**

3.strategy: **Tell me a *strategy* that you use for learning English so we can help others improve too.**

4.verbal: ***Verbal* communication is just a fancy name for speaking.**

5.characteristic:**One *characteristic* of my friend Sean is that he drinks a lot.**

6.security: **I don't like the *security* cameras on campus because I feel like I'm being watched.**

Item	Diff.	Disc.
62	0.59	0.67
63	0.91	0.05
64	0.82	0.71
65	0.73	0.74

7.sophisticated: **Computers are so *sophisticated*. I don't think I'll ever fully understand how they work.**

8.via: **I got a message *via* email that said a famous rock band is going to play on my birthday.**

9.ultimately: **The world will *ultimately* end, but no one knows when.**

10.interracial: **Many marriages in the US today are *interracial* which is why seeing black and white couples is not uncommon.**

Appendix F - Vocabulary Test Study Guide

Vocabulary Test Study Guide

The words below are the ones that you will need to know for our final vocabulary test.

To succeed on the test you will also need to know the following information about the words:

1. You will need to know the definitions that our textbook gives for each of these words and be able to match the word to its definition. You can find this information by looking at the answers for the Build Your Vocabulary sections in each chapter we covered.
2. You will need to recognize the stressed syllable in a word after you hear it.
3. You will need to be able to recognize the part of speech for other forms of the words we've studied when you read a sentence that includes that word.
4. You will need to be able to write a definition for other forms of the words we've studied.
5. You will need to be able to write sentences for the words we've studied that show that you know how to use them meaningfully and grammatically.

Unit 1 Vocabulary

1. attitudes
2. constantly
3. construct
4. evolving
5. expanding
6. identity
7. inevitable
8. phenomenon
9. reinforces
10. widespread

Unit 2 Vocabulary

1. alert
2. aptitude
3. devoted
4. exhibit
5. imagination
6. inconsistencies
7. motivation
8. predominant
9. strategy
10. underlying

Unit 4 Vocabulary

1. converging
2. enables
3. entire
4. guaranteed
5. homogeneous
6. ideological
7. media
8. promote
9. universal
10. verbal

Unit 6 Vocabulary

1. anthropologists
2. attachment
3. characteristic
4. emotion
5. enhance
6. hormone
7. invoke
8. mutual
9. prospective
10. romantic

Unit 8 Vocabulary

1. civil liberties
2. commercial
3. composite
4. controversial
5. deterrent
6. security
7. sophisticated
8. suspected
9. techniques
10. via

Unit 9 Vocabulary

1. accent
2. discrete
3. distinct
4. flexible
5. generation
6. impressive
7. nonverbal behavior
8. precise
9. random
10. ultimately

Unit 12 Vocabulary

1. adulthood
2. confirmed
3. interracial
4. legitimacy
5. matured
6. nationalistic barriers
7. norm
8. population
9. pregnant
10. social class

Vocabulary Test Example Questions

The following examples will show you the questions you will need to answer on the vocabulary test.

Part 1: Identifying the stressed syllable

Listen as your teacher reads the following words. Circle the syllable for each word that receives the *most* stress. Your teacher will read each word twice.

Example: 1. money (mo)–ney

Part 2: Matching the words with definitions

Write the letter next to each word to show its definition.

Example: 1. nondairy B

A. not very interesting
B. not made with milk or cream

Part 3: Identify parts of speech

Read the sentences below and write the correct part of speech (noun, verb, etc.) for each of the underlined words in the blank.

Example: 1. Jim is unreliable because he never keeps a promise. adjective

Part 4: Write definitions

Write a definition for each word from Part 3 based on how it is used in the sentence above.

Example: 1. unreliable : Means <u>not dependable.</u>

You will not receive points for answers that use another form of the word.

Example: 2. unreliable: Means <u>not reliable.</u>

Part 5: Sentence writing

Write a sentence for each of the following words to show that you know what the word means and how it is used. You may choose a different form of the word than is given if you wish. For example, if the word given is *complicated,* then you may also use *complicate,* or *complication*.

A sentence that shows that you know the meaning of the word and can use correct grammar is worth full credit.

Example: 1. complicated: The math problem was really *complicated* so I didn't get the right answer.

A sentence that shows that you know the meaning of the word, but uses incorrect grammar is worth half credit.

Example: 2. complicated: The math problem was really *complication* so I didn't get the right answer.

A sentence that does not show that you know the meaning of the word receives no credit.

Example: 3. complicated: It is *complicated*.

Appendix G – Test for Future Administration

Name:_____ Score:_____/68

Contemporary Topics 3 Vocabulary Test

Part 1: Identifying the stressed syllable (10 points)

Listen as your teacher reads the following words. Circle the syllable for each word that receives the *most* stress. Your teacher will read each word twice. Each correct answer is worth 1 point.

Example: 1. money (mo)—ney

Note: Answers here appear in bold, although students were expected to circle the answers on the actual test.

1. imagination i—ma—gi—**na**—tion
2. inconsistencies in—con—**sis**—ten—cies
3. guaranteed guar—an—**teed**
4. universal u—ni—**ver**—sal
5. controversial con—tro—**ver**—sial
6. discrete dis—**crete**
7. impressive im—**pre**—ssive
8. interracial in—ter—**ra**—cial
9. legitimacy le—**gi**—ti—ma—cy
10. population po—pu—**la**—tion

Part 2: Matching the words with definitions (22 points)

Write the letter next to each word to show its definition. Each correct answer is worth 1 point.

Example: 1. nondairy B

A. not very interesting
B. not made with milk or cream

Matching Part A

1. anthropologist	**B**	A. a society in an advanced state of development
2. phenomenon	**G**	B. a person who studies societies and their beliefs
3. aptitude	**F**	C. a person who studies living things
4. motivation	**E**	D. the chemical process that changes food into energy
		E. eagerness or willingness to do something
		F. a natural ability or skill
		G. a remarkable thing

Matching Part B

1. identity	**G**	A. something that prevents people from doing something
2. hormone	**D**	B. a source of supply or support
3. deterrent	**A**	C. a form of transport
4. norm	**E**	D. a substance that influences growth and development
		E. the usual or acceptable way of doing something
		F. opinions and feelings
		G. a quality that makes someone distinct

Matching Part C

1. construct	**E**	A. give time and perhaps money to some activity
2. evolve	**B**	B. develop
3. devote	**A**	C. increase the speed of something
4. exhibit	**G**	D. gradually increase in quantity or size
		E. create or build
		F. experience or go through
		G. show something so that it's easy to notice

Matching Part D

1. converge	**E**	A. determine that something is definitely true
2. confirm	**A**	B. make better
3. enhance	**B**	C. keep
		D. locate
		E. become similar
		F. stop doing something

Matching Part E

1. inevitable	**F**	A. hidden and not easy to discover
2. alert	**D**	B. clearly different or separate
3. underlying	**A**	C. not objective
4. distinct	**B**	D. watching and listening carefully
		E. absolutely necessary
		F. unavoidable
		G. able to continue

Matching Part F

1. ideological	**_D_**	A. having to do with business or trade
2. commercial	**_A_**	B. separate from physical realities
3. composite	**_F_**	C. undependable
		D. based on beliefs or ideas
		E. official
		F. made up of different parts

Part 3: Identify parts of speech (10 points)

Read the sentences below and write the correct part of speech (noun, verb, etc.) for each of the underlined words in the blank. Each correct answer is worth 1 point.

Example: 1. Jim is unreliable because he never keeps a promise. adjective

1. The proposed cuts have caused considerable controversy. **noun**_____

2. Supporters of the death penalty argue that it would deter criminals from carrying guns.
verb_____

3. He said it is still not possible to predict the ultimate outcome. **adjective**_____

4. The flexibility of the lens decreases with age; it is therefore common for our sight to worsen as we get older.

noun_____

5. This administration is not entirely free from suspicion. **adverb**_____

6. The disadvantage is that it is not universally available. **adverb**_____

7. George was unable to verbalize how he felt about a girl in his class. **verb**_____

8. This program is not a legitimate use of the taxpayers' money. **adjective**_____

9. Humans experience a delayed maturity; we arrive at all stages of life later than other mammals.

noun_____

10. All airports in the country are working normally today. **adverb**_____

Part 4: Write definitions (20 points)

Write a definition for each word from Part 3 based on how it is used in the sentence above.

Acceptable Answer: unreliable : Means not dependable.

You will not receive points for answers that use another form of the word.

Unacceptable Answer: unreliable: Means not reliable.

Each correct answer is worth 2 points.

Note: Answers for Part 4 will vary, but possible answers are shown below.

1. controversy: Means **debate; disagreement; argument**

2. deter: Means **prevent; stop; hinder**

3. ultimate: Means **final**

4. flexibility: Means **ability to change easily; pliability**

5. entirely: Means **completely; totally**

6. universally: Means **globally**

7. verbalize: Means **say; put into words**

8. legitimate: Means **appropriate; acceptable; valid**

9. maturity: Means **development; growth**

10. normally: Means **as usual**

Part 5: Sentence writing (16 points)

Write a sentence for each of the following words to show that you know what the word means and how it is used. You may choose a different form of the word than is given if you wish. For example, if the word given is *complicated,* then you may also use *complicate,* or *complication*.

A sentence that shows that you know the meaning of the word and can use correct grammar is worth full credit (2 points)

Full Credit Answer: complicated**:** The math problem was really *complicated* so I didn't get the right answer.

A sentence that shows that you know the meaning of the word, but uses incorrect grammar is worth half credit (1 point).

Half Credit Answer: complicated: The math problem was really *complication* so I didn't get the right answer.

A sentence that does not show that you know the meaning of the word receives no credit (0 points).

No Credit Answer: complicated: It is *complicated*.

Note: Answers for Part 5 will vary, but possible answers are shown below.

1. attitude: **People have many *attitudes* about abortion because it is such a controversial topic.**

2.strategy: **Tell me a *strategy* that you use for learning English so we can help others improve too.**

3.verbal: ***Verbal* communication is just a fancy name for speaking.**

4.characteristic: **One *characteristic* of my friend Sean is that he drinks a lot.**

5.sophisticated: **Computers are so *sophisticated*. I don't think I'll ever fully understand how they work.**

6.via: **I got a message *via* email that said a famous rock band is going to play on my birthday.**

7.ultimately: **The world will *ultimately* end, but no one knows when.**

8.interracial: **Many marriages in the US today are *interracial* which is why seeing black and white couples is not uncommon.**

Appendix H-Instructions for Completing Vocabulary Chart Activity

Handout the vocabulary chart and assign students to groups. Write group numbers on the board and assign different words for each group. Set a time limit for students to complete the information for their assigned words without using a dictionary, so they can test themselves on what they already know.

After the time limit has been reached hand out dictionaries or allow students to use their own resources to correct and complete their charts. Tell the students they should reach a consensus as a group about the information that they want to maintain for their assigned words so that each student in the group has the same information when they are finished.

Assign students to new groups so that every member in the group has information for a different set of words from the chart.

Tell the students to take turns teaching the words they researched to their new group members. Tell them that they should not show their charts to their group members in order to use speaking and listening to complete the task.

The student teaching the new words is told to check that the other group members record the information correctly by pointing out any mistakes that the group members make while writing. Additionally, the students writing can be encouraged to correct any pronunciation mistakes that they hear from the student teaching while they are writing.

The teacher collects the charts when the activity is over and can clarify any common misunderstandings in the following class.