The C-test: Usefulness for measuring written language ability
of non-native speakers of English in high school

by

Jongsook Park

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: English (Teaching English as a Second Language/Applied Linguistics)

Major Professor: Carol A. Chapelle

Iowa State University
Ames, Iowa
1998
Graduate College
Iowa State University

This is to certify that the Master’s thesis of

Jongsook Park

has met the thesis requirements of Iowa State University

__________________________
Major Professor

__________________________
For the Major Program

__________________________
For the Graduate College
TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1

CHAPTER 2. TEST DEVELOPMENT 10

CHAPTER 3. METHODS 51

CHAPTER 4. RESULTS AND DISCUSSION 62

CHAPTER 5. CONCLUSIONS 83

APPENDIX A. INSTRUCTIONS FOR TEST PROCTOR/RATER 96

APPENDIX B. QUESTIONS AND ANSWERS FOR LOGICAL EVALUATION OF TEST USEFULNESS 97

APPENDIX C. GUIDELINES TO CONSTRUCT A C-TEST 103

APPENDIX D. THE SAMPLE C-TEST MATERIAL 104

REFERENCES 108

ACKNOWLEDGMENTS 111
CHAPTER 1. INTRODUCTION

In this chapter, first, I will explain the reason for developing the C-test as a test of written language ability and, thus, the purpose of this study. I will also introduce the C-test, the initial test development, and the concept of test usefulness qualities and their roles in test development. Finally, I will address the research questions in this study and how those questions will be examined throughout this study.

How did this study start? In order to answer the question and, therefore, see this study in context, it is necessary to address the reason for developing the C-test as a test of written language ability. The reason for starting the development of a version of the C-test was that there is a need for a test which can assess written language ability of non-native English speaking (NNES) students for English as a Second Language (ESL) instruction in high school in rural Iowa. We have to look at the situation of NNES students in high school to understand why a test of written language ability is necessary. All NNES students are expected to study in English-medium class as soon as they enroll in schools. Some NNES students could have high-level English ability, or near native-like English ability. In this case, those students could compete on an equal basis with native speakers with little or no language barriers in the classroom. The others may have difficulties in participating in the classroom as their English abilities are not fully developed to compete with native speakers. Those students are considered limited English proficient (LEP) students. According to a handbook for administrators and teachers (Iowa Department of Education, 1996a: 2), “Limited English Proficient (LEP) students are students with a home language background other than English,
whose English language skills are not yet well enough developed for them to be able to participate successfully in classrooms where all academic instruction is provided in English.” Therefore, LEP students may need additional help with their English in accomplishing academic requirements successfully in academic domains in school such as following instructions, reading materials, taking notes, taking tests, and/or participating in discussion in and outside class.

In order to help LEP students participate successfully in the classroom, schools are expected to provide appropriate language programs such as ESL programs or bilingual aids/education in order to provide an equal educational opportunity. Many schools have been providing ESL instruction; others are expected to do it. The need to provide an equal educational opportunity for LEP students has been stated in various documents such as in the following:

Numerous acts, laws, court decisions, and guidelines have been written over the years for these [LEP] students. They combine to create and clarify the current legal responsibilities of all United States school districts for the education of LEP students (Iowa Department of Education, 1996a: 2).

In addition, schools need to provide funds for instructional programs for LEP students as stated in the same handbook (1996a: 9) as follows:

Inherent in a school district’s obligation to take “appropriate action to overcome language barriers that impede equal participation by its students” (Equal Educational Opportunity Act of 1974) is the obligation to finance these programs.

Therefore, it is the schools that need to act to help LEP students improve their English ability to participate successfully in academic domains in school. The first step for schools to do this job would be to identify/select LEP students and, then, place them in appropriate ESL or bilingual programs.
Then, how do schools select LEP students and place them in ESL programs? That is, what kinds of assessment packages do schools utilize to accomplish this mission? In high schools in rural Iowa, current assessment packages usually include informal observations by mainstream teachers and/or ESL staff, interviews, Home Language Survey questionnaires (Iowa Department of Education, 1996a) and/or oral proficiency tests such as the Bilingual Syntax Measure II developed in the 1970's (Bilingual Syntax Measure II, 1978). The Bilingual Syntax Measure II, one of the most widely used placement tests (Iowa Department of Education, 1996a: 37), is an oral proficiency test which assesses some syntactic features of spoken language ability using cartoon drawings. Therefore, in most cases, there is no formal test to assess written language ability which should be considered a crucial element of language ability required in academic domain in high school.

The fact that there is no formal assessment of written language ability might cause inappropriate selection or identification of LEP students and placement of them for ESL instruction. Only with appropriate assessments in selecting and placing LEP students in ESL programs, schools could be able to provide more appropriate educational opportunities for LEP students. Therefore, there is a need to develop a test of written language ability in order to help schools do their jobs better in providing appropriate ESL programs for the current situation; this became a background of this study.

Then, what is the purpose of this study in the above context? The purpose of this study is to examine whether a version of the C-test is useful as a test measuring "written language ability" of non-native English speaking high school students. The results of the test, then, may help teachers and decision-makers select and place LEP students for ESL
instruction in high schools in rural Iowa. Here, I call the C-test a test of “written language ability” instead of addressing a “reading” test or a “writing” test. The reason is that in the C-test, even though reading and writing skills are required, the scope of using reading and writing skills is quite limited. That is, in the C-test, the process of decoding and encoding could be carried out with a very limited range of language ability involving written texts. For example, the C-test does not require test takers to answer questions about the main idea of a text; test takers do not need to write even a sentence level answer. The only type of responses in the C-test is writing parts of words after reading a mutilated text; some answers could even easily be cued by the existing halves of the words without understanding the whole content of a text. Therefore, by defining the C-test as a test of “written language ability,” I would like to emphasize the fact that even though the C-test requires test takers to use language ability to read written texts and, then, produce written responses, the degree of using language ability involving the reading and writing processes in the C-test is quite limited. In addition, the language ability (Bachman, 1990, quoted in Bachman and Palmer, 1996) consists of language knowledge and metacognitive strategies. Among the areas of language knowledge, the C-test mainly requires a test taker’s grammatical knowledge; using metacognitive strategies are also limited. (See discussion about language ability in Chapter 2.) Therefore, in this study, the C-test is considered a test of “written language ability.”

C-test usefulness will be examined by answering research questions which are addressed later in this chapter. In order to explain why I have chosen the C-test method for the above purpose, it is necessary to introduce what the C-test method is--its history,
definition, and evaluation—and explain the initial development of the C-test as a test of written language ability for the above situation which was a basis of this current study.

The C-test method was developed by Raatz and Klein-Braley (1981) in order to improve shortcomings of the classical cloze test methods, which produce fill-in-theblank tests. Therefore, the C-test method shares its theoretical background with the cloze tests. However, its deletion procedures and other requirements are different from those of the cloze test methods. In the cloze tests, words are deleted in a fixed or random ratio, usually in a 5th to 10th word ratio. In order to get enough deletions, at least 50 deletions, one longer text is presented. Therefore, in order to successfully write answers, test takers might need topical knowledge for that particular area in addition to language ability. Also, in the close tests, as each word is entirely deleted, there are likely various acceptable answers for some deletions. This might require more time to score the cloze tests unless the rater counts only exact words as correct responses.

On the other hand, in the C-test method, parts of words are deleted in a fixed ratio in several short texts; the deletion procedure is done in a fixed ratio by deleting parts of words, usually the second half of every second word starting in the second sentence. Having 100-120 deletions is recommended to produce good reliability coefficients. This procedure requires less number of words in total in order to get a same number of deletions compared to the cloze test methods. Therefore, the C-test method requires less scoring time. As five to six short texts of independent topics are given, the C-tests may require a greater variety of topical knowledge than the cloze test. For scoring the C-tests, usually two scoring methods have been recommended. One is the exact scoring method; with this scoring method, only an exact
word in the original text is counted as a correct response. Another one is the acceptable scoring method; with this scoring method, each acceptable word is counted as a correct response if the meaning of the word does not change the meaning of the exact word in the original text. Each correct response gets 1 point; no point is given for an incorrect or no response.

The C-tests were originally developed for proficiency and placement testing (Klein-Braley, 1985; Klein-Braley and Raatz, 1984; Chapelle, 1994). According to Klein-Braley and Raatz (1984), the C-tests were explicitly developed as a test of "general language competence" which includes various aspects of language ability. Language ability consists of language knowledge and strategic competence, or metacognitive strategies (Bachman and Palmer, 1996), which will be explained more in Chapter 2. As there is no reason to believe that the C-tests measure the test taker's spoken language ability (Klein-Braley, 1996), it could be said that the C-tests measure ability in written language.

Several studies have reported the usefulness of the C-tests in their particular settings. The C-tests have been known to be reliable in almost all settings and valid in most settings. (See discussion in Chapter 2 for more details.) The fact that the C-tests have produced satisfactory reliability and validity coefficients in most settings (Klein-Braley and Raatz, 1984) is very important because the reliability and validity are crucial elements of qualities of test usefulness. Examining reliability and validity will be also the core part of the C-test usefulness evaluation in this study. The qualities of test usefulness will be briefly introduced later in this chapter and will be discussed in depth in the next chapter. For the purpose of ESL selection and placement of LEP students, the C-tests seem to work in some settings whether in ESL or
EFL situations in assessing the language ability of non-native speakers; this was the basis for choosing the C-test method to develop a test of written language ability in the first place.

The first version of the C-test used in this study was developed to fulfill a requirement in a testing course, English 519, in the fall of 1996. At that time, I was able to complete the C-test development as a group project with two other graduate students who were also in the Department of English with concentration of Teaching English as a Second Language/Applied Linguistics. One of them was a consultant for English Speakers of Other Languages (ESOL) at a local area education agency (AEA) and introduced the need for developing a test of written language ability to improve assessment procedures for ESL selection and placement in secondary schools in rural Iowa.

After I examined the C-test usefulness in other testing situations described in the literature, I decided to develop a version of the C-test for the high school level described above. The process of the test development was quite thoroughly discussed in and outside the classroom. I examined several steps of test development including both theoretical and empirical considerations: the purpose of the test, the tasks in the target language domains, the characteristics of the test takers, the definition of the construct to be measured, the test tasks, the scoring rubric, and the usefulness of the test. At that time, the usefulness of the C-test was addressed by examining the logical analysis and mini-empirical analysis involving only one NNES high school student. In order to examine the C-test usefulness, a larger scale empirical study was considered crucial to enhance the possibility of real use of the C-test, which triggered to start this research study.
In order to address why the test usefulness needs to be examined for the real use of the C-test, it is necessary to discuss the qualities of test usefulness and their roles in the process of test development and use. According to Bachman and Palmer (1996), there are six qualities of test usefulness: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Bachman and Palmer emphasize that it is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness. Their notion of usefulness is expressed such as in the following (Bachman and Palmer, 1996: 18):

\[
\text{Usefulness} = \text{Reliability} + \text{Construct validity} + \text{Authenticity} + \text{Interactiveness} + \text{Impact} + \text{Practicality}
\]

Among the six qualities, the reliability and construct validity have been known as two major qualities of test usefulness. The reliability tells us to what extent a test is consistent in measuring the test taker's language ability. The construct validity tells us whether a test is valid for measuring what would be expected to be measure. However, the other qualities also play important roles in their own way in the process test development and use. That is, authenticity tells us to what extent a test task is close to a task in the target language domains in real life. Interactiveness is also important as it tells us the extent of a test taker's involvement in accomplishing a test task. We also want to know impact of a test on individuals and educational systems involved in the test taking and use. Finally, practicality is related to the relationship between required resources and available resources in developing and using a test. Therefore, all the qualities have to be examined to determine the overall usefulness of a test in each testing situation. All six qualities of test usefulness will be defined and their roles will be discussed in Chapter 2.
In this study, the C-test usefulness will be examined in the following manner. All six qualities of test usefulness will be examined in answering the following six research questions, which will be the basis for determining the overall usefulness of the C-test in this study:

1. Is the C-test reliable to measure written language ability of ESL students?
2. Does empirical analysis provide construct validity evidence for the C-test?
3. To what extent is the C-test similar to authentic tasks in target language use domains?
4. To what extent is the C-test task interactive?
5. What would be the impact on individuals (test takers and teachers) and educational systems/society? Is the impact positive?
6. Is the C-test practical to use?

Answers for the above research questions will be examined and addressed in the following manner. Chapter 2 will examine all six qualities of test usefulness based on the logical analysis of test usefulness as part of test development, which will answer the above six questions fully or partially. Chapter 3 will report methods used for data collection and analysis for the pilot and main tests as a basis for the empirical evaluation of test usefulness. Chapter 4 will discuss the results of the empirical study examining reliability, construct validity, impact, and practicality, relevant to the research questions 1, 2, 5, and 6. Chapter 5 will summarize answers for all six research questions and draw the conclusions about the C-test usefulness; then, it will address recommendations for the future use of the C-test along with the discussion about the areas of improvement and implications for future research.
CHAPTER 2. TEST DEVELOPMENT

In this chapter, I will examine the process of test development of the C-test. In order to that, first, I will address the purpose of the test. Next, I will define the qualities of test usefulness, which will be the basis for further discussion in this chapter. I will also provide a literature review about the C-test and its usefulness, which will give us relevant information for the C-test usefulness in this study. Then, I will define the target language use (TLU) domains and describe the characteristics of two TLU tasks, which are likely related to the test task in the C-test. I will describe the characteristics of the test takers in detail as well. This will give us information about the target population of the C-test as a test should be developed for a particular target group in mind. I will also define the construct to be measured on the C-test and, then, examine the test task in the C-test. Finally, I will examine the C-test usefulness based on the logical analysis of test usefulness qualities. The logical evaluation of test usefulness will answer the research questions 1 through 6 for reliability, construct validity, authenticity, interactiveness, impact, and practicality, respectively. In addition, for reliability, construct validity, impact, and practicality, more evidence will be examined based on the results of empirical research in Chapter 4; this will answer the research questions 1, 2, 5, and 6 from the perspective of empirical evaluation. The methods for the empirical research will be addressed in Chapter 3.
The purpose of the test

The purpose of the test in this study is to make inferences about written language ability in a high school academic domain in which written language ability is necessary to understand and carry out academic requirements successfully. Written language ability will be discussed when I define the construct later in this chapter. Ideally, the inferences from the test are expected to be used as part of information in the whole assessment package that schools would take into consideration in making decisions about test takers and decisions about ESL programs.

Decisions about ESL programs (at the program level)

Based on the inferences from the test, schools may consider to start ESL program(s) and/or hire ESL staff. That is, schools that currently do not offer any ESL or bilingual help for LEP students may consider offering ESL program(s) and/or hiring ESL staff when NNES student(s) enrolled in school are considered LEP student(s) based on the inferences from this test and other assessments.

Decisions about test takers (at the individual level)

The inferences from the test may help schools and teachers select or identify LEP students and place them in ESL program(s). Based on the inferences of the test as part of the whole assessment, if an NNES student scores high on the test and other assessment, he or she may be exempt from ESL instruction: that is, written language ability of the student may be considered at least sufficiently high to participate successfully in the academic domain of
school. If LEP student(s) are identified, they may receive additional language help for their academic achievement, which will improve educational environment for all the students in school.

Now, we need to ask if the C-test can be used as a test to accomplish the above purpose. As the first step to answer this question, I will provide a literature review about the C-test in other studies. Then, the next step will be examining the remaining process of test development for the C-test for the purpose of the test addressed above.

The qualities of test usefulness

The qualities of test usefulness will be examined following Bachman and Palmer’s study (1996). According to them, there are six qualities of test usefulness: the reliability, construct validity, authenticity, interactiveness, impact, and practicality. In Bachman and Palmer’s framework, each quality should be seen in a compensatory relationship, not in conflict --e.g. weaker reliability may be compensated by stronger validity, etc. Therefore, it is the overall test usefulness that needs to be maximized, not an individual quality, in each particular testing situation. It is necessary to define the qualities of test usefulness and their roles in the process of test development and use as it will be the basis for further discussion in this chapter.

Reliability is defined as the degree of consistency of scores of test takers. Reliability is a crucial element of test usefulness as a test should consistently measure test takers’ performances or language abilities on a test. If performances of test takers on the test result
Construct validity is a core part of the process of validating a test (Chapelle, 1994). Construct validity is another crucial element of test usefulness as a test should measure what would be expected to be measured in the test. In a traditional or old view of validity, a test could be considered valid and be used without further examination after it has been validated once even in one particular testing situation. However, a new view of construct validity has emerged recently, which is the concept referred to in this study. The new view claims that the construct validity is on-going process (Messick, 1991) and cannot be said that a test is valid for all time. That is, the construct validity of test use should be examined in each particular setting.

Among the six qualities, the reliability and construct validity are critical for test development and are referred to as essential measurement qualities. This is because they are the qualities that would provide the major justification for using test scores as a basis for making inferences of decisions.

Authenticity is defined as the degree of correspondence between the characteristics of a given language test task and the features of a target language use (TLU) task (Bachman and Palmer, 1996: 23). The authenticity is considered to be an important test quality because it relates the test task to the domain of generalization to which we as test designers or developers want our score interpretations to generalize. It is an important test quality because it potentially affects test takers’ perceptions of the test and, then, their performances.
Interactiveness is defined as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task (Bachman and Palmer, 1996: 25). The individual characteristics, which can be related to language testing, are the test taker's language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata. Interactiveness in a language testing concerns the interaction between the test taker and the test task. That is, the interaction should require the use of language knowledge if we are able to make inferences about language ability on the basis of the test taker's performance. In order to make inferences about language ability, responding to the test task must involve the test taker's areas of language knowledge and his/her metacognitive strategies (Bachman and Palmer, 1996). It is also a critical quality of language test tasks as it is related to construct validity.

Impact is a quality of test usefulness that refers to how test taking and using a particular test impact on society, educational systems, the individuals within those systems. Impact operates at two levels: a macro level, in terms of the societal or educational system in general; and a micro level, in terms of the individuals (e.g., test takers and teachers) who are most directly affected by the particular test use (Bachman and Palmer, 1996). As using a test implies specific values and goals of an educational system or society, test developers and users always need to examine potential consequences of using a test for a particular purpose.

Practicality can be defined as the relationship between the resources required in the design, development, and use of the test, and the resources available for these activities. The practicality should be considered at every stage in the process of test development. That is, at every stage, test designers or developers need to examine what kinds of resources are required
and what kinds of resources are available to develop and administer a test. As the required and available resources will vary from one situation to another, practicality can only be determined for a specific testing situation. This test quality is different from the other five qualities. While those five qualities pertain to the uses of test scores, practicality pertains primarily to the ways in which the test will be implemented, and, to a large degree, whether it will be developed and used at all (Bachman and Palmer, 1996: 35). If the required resources do not exceed the available resources at any stage in test development, then the test is practical and the test development and use can proceed. Considerations of practicality, therefore, might affect test developers’ decisions at every stage in test development and might lead test developers to reconsider some of earlier specifications of test development.

Literature review about the C-tests

Basically the C-test method was developed by modifying the cloze test methods as a reaction to the drawbacks of the cloze test methods. Both methods claim that they could measure language ability in written texts with reduced redundancy where the test taker is expected to utilize all aspects of his/her language ability to restore deleted words or parts of words. Therefore, in the literature, the C-test method is often referred to one of the cloze test methods (Chapelle and Abraham, 1990). Actually, the theoretical basis of the C-test method shares a same hypothesis with that of the cloze test methods.

As the theoretical basis of the C-test method, Klein-Braley (1985) uses Oller’s claim about pragmatic language tests which was originally the basis of the classical cloze tests:
any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to related sequences of linguistic elements via pragmatic mappings to extra-linguistic context (Oller, 1979: 38)

According to the above claim, cloze tests are pragmatic tests because they meet the naturalness criteria for language tests: in order to give correct responses, the test taker may need to operate on the basis of both immediate and long range contextual constraints. Also, she or he must utilize information that is inferred about the facts, ideas, events, relationships, states of affairs, social settings, and the like that are pragmatically mapped by the linguistic sequences contained in the passage (Oller, 1979: 43-44). Examples of cases of pragmatic mappings where extralinguistic context and the linguistic context are interrelated could be found in using so-called deictic words (e.g., here, now, then, there, this, and that), pronouns that refer to persons or things, tense indicators, aspect markers on verbs, adverbs of time and place, determiners and demonstratives in general, etc. (Oller, 1979: 43). As claimed above, pragmatic tests, therefore, are always integrative. On the other hand, discrete point tests like multiple choice tests cannot be pragmatic. For example, there is no normal language use context where a language learner may be asked to listen to and distinguish between minimal pairs of phonological contrasts. In this sense, integrative tests are often pragmatic; cloze procedure and dictation are examples of pragmatic tests (Oller, 1979: 38). As Klein-Braley (1985) claimed, as both cloze procedure and C-test procedure meet the pragmatic naturalness criteria for language tests by using authentic materials as the basis of test construction, Oller’s comments apply equally to the C-tests.

Both C-test method and cloze test methods are considered tests of reduced redundancy. In order to discuss tests of reduced redundancy, it is necessary to understand the
concept of redundancy. Here, redundancy is related to language use in a natural context which is a basis of pragmatic tests. That is, language itself or language use is presumably redundant (Gradman and Spolsky, 1975). In other words, there are a variety of clues in normal discourse; those clues could include grammatical, lexical, syntactical, pragmatic, etc. In normal discourse, language users utilize those clues in understanding the utterances or texts depending on a particular language use situation. By adding noise to the background, or deleting some parts of utterances or texts, redundancy of language could be reduced to some degree leaving enough clues for a native speaker, but not for a non-native speaker, to interpret the message.

Reduced redundancy test methods, therefore, test the test taker’s ability to make use of the remaining redundancy of the language as a whole in order to restore the missing parts of words in the text. Klein-Braley (1985) explains the required use of language ability in a situation where the test taker faces reduced redundancy. That is, in the cloze tests and the C-tests, reduced redundancy is achieved by deleting parts of texts; then, the test taker needs to restore the deleted parts in the original text or a possible/acceptable text. The process of restoring requires the test taker’s knowledge of the language or about the language (Klein-Braley, 1985). This leads the assumption that as the test taker’s control of the target language improves, he or she may be able to make more successful use of the redundancy provided by natural language in authentic texts, and will thus achieve a higher score on the test (Klein-Braley, 1985). As Klein-Braley (1985) claims, in order to restore the text, the test taker needs to utilize all kinds of clues: grammatical, syntactical, lexical, semantic, collocational, contextual, pragmatic, logical, situational, etc. Therefore, the final score obtained by the test
taker on a test of reduced redundancy can be considered a numerical estimate of his/her ability (Klein-Braley, 1985; Raatz, 1985). In order to achieve the purpose of tests of reduced redundancy, it is necessary to obtain random samples of the test taker’s performance by using a random deletion procedure for test construction as tests of reduced redundancy do not directly related any specific areas of language abilities.

Then, we need to look at what the problems with the cloze test methods were and how the C-test method could likely improve the problems? One major problem was that the cloze test methods produced the unpredictable results because the results were obtained by various fixed ratio deletion procedures (Alderson, 1979; Chapelle and Abraham, 1990). The every second word procedure on the C-tests were claimed to improve the fixed ratio cloze procedure by producing a large number of random samples within a same length (Klein-Braley, 1985: 84). A second problem with the cloze test methods was the effect of text topics and difficulty on the test performance of the test taker (Chapelle and Abraham, 1990). This problem could be minimized by presenting several different short texts (Klein-Braley, 1985: 84). A third problem was the lack of criterion reference point which should be defined by performances of educated native speakers (Chapelle and Abraham, 1990; Klein-Braley, 1985). That is, educated adult native speakers should be able to score high on a test measuring language ability without or less topical knowledge. However, the cloze tests could be quite difficult even for adult educated native speakers as their performances on the cloze tests could be very much affected by topical knowledge. This would not be the case on the C-tests as Klein-Braley (1985: 84) claims: “Adult educated native speakers achieve virtually perfect scores.” Here, a criterion reference point for the C-test is defined by performances of
educated adult native speakers, not by performances of non-native speakers with better
language ability in a particular testing situation. The reason is that there are unlikely to be
enough number of non-native speakers with better language ability in each testing situation; it
is also difficult to define better language ability compared to language ability of the entire
target non-native speaker group in that testing situation. The C-test method, therefore, was
introduced as an improved test of reduced redundancy.

The C-test method is intended as a means of constructing norm-referenced tests to be
used for proficiency and placement testing (Klein-Braley and Raatz, 1984; Klein-Braley,
1985). A norm-referenced test (NRT) is a test to discriminate test takers along a continuum
with scores on average 50 percent, which creates a normal distribution (Brown, 1996). An
NRT is different from a criterion-referenced test (CRT) because, in a CRT, the test taker is
expected to show his/her understanding of a course as in a final examination of a course.
Therefore, whereas an NRT is usually administered at the beginning of a program for selection
and/or placement, a CRT is likely given as an achievement test at the end of a course.

Klein-Braley and Raatz (1984) and Klein-Braley (1997) make the following claims
about the C-tests:

1. they are easy to construct;
2. more items are possible with much shorter texts;
3. scoring is objective because there is almost always only one possible solution;
4. scoring is quick for the native speaker or the teacher because it takes only slightly more
time than is needed for reading the texts;
5. C-tests are very easy for native speakers;
6. the deletion procedure in every second word likely raises the probability of obtaining a representative sample of all the word classes in the text;

7. as C-tests consist of several different texts, the sampling of content categories is considered better.

Also, Raatz (1985: 74) reports that the C-tests comprise individual test parts that are homogeneous and that "...the total score can be assumed to be unidimensional and interval-scaled." All of the above claims about the C-tests are related to the C-test usefulness research, which is the focus of this study.

The C-test usefulness has been investigated since it was introduced in 1981 (Chihara, Cline, and Sakurai, T., 1996; Chapelle, 1994; Chapelle and Abraham, 1990; Coleman, 1996; Dornyei and Katona, 1992; Huhta, 1996; Jafarpur, 1995: Klein-Braley, 1984; Klein-Braley, 1996; Koller and Zahn 1996; Sigott and Koberl, 1996). Most studies have reported that the C-test method is reliable, producing high reliability coefficients (Coleman, 1996; Dornyei and Katona, 1992; Jafarpur, 1995).

For validating the C-tests, various kinds of criteria have been used. Klein-Braley and Raatz (1984) report that they used teacher ratings or school grades as criteria for validating the C-test. In the same study, they also provide evidence and theoretical justification for the C-test validity in a variety of contexts. Other researchers also have provided evidence based on other kinds of criteria (e.g., standardized proficiency tests, selection/placement tests in their institutions, etc.). In one study (Chihara, Cline, and Sakurai, 1996), TOEFL scores were correlated with the C-test results. Dornyei and Katona (1992) present evidence for the C-test against four different language tests which include the Test of English for International
Communication (TOEIC), the department proficiency test, the oral interview, and the cloze test. In another study (Chapelle and Abraham, 1990), evidence for the C-test is provided against the placement test of a university consisting of four parts: vocabulary, reading, listening, writing. As another way of investigating the construct validity, experimental studies comparing the performances of adult native speakers with those of non-native speakers have reported that native speakers greatly outperform non-native speakers, mostly ranging in between 80% and 95% depending on testing materials and situations (Huhta, 1996; Jafarpur, 1995).

However, there have been some arguments related to the validity of the C-tests because researchers have different opinions in defining the construct of the C-tests. Some report that the C-test is useful for general proficiency testing (Dornyei and Katona, 1992; Klein-Braley and Raatz, 1984). Others present a problem in using the C-test for general proficiency testing (Jafarpur, 1995; Huhta, 1996); they argue how the C-test requiring mainly grammatical knowledge on the test represents spoken or reading ability of the test taker.

Even in one study (Singleton and Little, 1991), the researchers argued for the quality of the C-test as a second language vocabulary test; however, their argument was questioned in another study (Chapelle, 1994). Accordingly, the correlational studies have reported various results even though the C-test was considered a general proficiency test in most cases. In some studies, the C-test was more correlated with integrative tests like the department proficiency test consisting of vocabulary, grammar and listening parts, TOEIC total, the cloze test, or dictation rather than with discrete tests (Dornyei and Katona, 1992; Huhta, 1996). In one study (Chapelle and Abraham, 1990), however, the C-test was more closely correlated with a
vocabulary test than with listening, reading, or writing tests. There have been also arguments concerning the face validity of the C-tests. Some report that the C-test has face validity (Klein-Braley and Raatz, 1984; Klein-Braley, 1996); others find that the C-test does not possess face validity (Huhta, 1996; Jafarpur, 1995). Therefore, the face validity may be quite differently perceived by test takers and/or teachers in different testing situations.

Regarding impact on society and educational systems, and upon individuals (e.g., test takers, teachers), several studies report positive impact in their testing situations: that is, the C-tests were used for selection and/or placement testing and research purposes especially in Europe. One study (Huhta, 1996) reports that the C-test was administered to analyze a testing situation in the Department of English at a university in Finland. Another study (Coleman, 1996) reports that the C-test was used for comparative proficiency testing in the UK, Germany, and Austria for learners of English, French, Spanish, German, and Russian. In addition, Klein-Braley (1996) reports that the C-test has been used in Germany to select secondary school students for a nation-wide English proficiency competition. Increasing uses of the C-tests in real life may be seen as an indication of positive impact.

Since Klein-Braley and Raatz (1984) reported that the C-tests are easy to construct and score and easy to administer, most of C-test studies have reported that the C-tests are practical to use in their particular testing situations (Jafarpur, 1995; Coleman, 1996). One study (Huhta, 1996), however, reports a problematic area in constructing the C-test. That is, some text types appeared to be problematic as texts of the C-test; the often used average sentence length and type-token ratio were not very useful for test design in that study.
Description of tasks in the target language use (TLU) domain

We have to define what the target language use domain is before describing tasks in the TLU domain. Bachman and Palmer (1996: 44) define “a target language use domain as a set of specific language use tasks that the test taker is likely to encounter outside of the test itself and to which we want our inferences about language ability to generalize.” According to them, there are two general types of TLU domains relevant to the process of test design and test development: real life domains and language instructional domains. While the general procedures for identifying tasks for potential development as test tasks are essentially the same for these two domains, the considerations with respect to usefulness are somewhat different (Bachman and Palmer, 1996: 103). That is, in a real life domain, a test may be designed to be used in making decisions (e.g., hiring decisions for a job) which are more directly relevant to the test taker’s performance on tasks; real life tasks may be a basis for developing test tasks if we know real life conditions that the test takers will face and if the test takers’ language ability is high enough to perform test tasks based on real life tasks (e.g., a test of English for business communication). On the other hand, in a language instructional domain, language is used for the purpose of teaching and learning of language; we may design a classroom quiz or achievement test based on language instructional tasks. The characteristics of language instructional TLU tasks may match those of real life TLU tasks in some cases; in other achievement testing situations, it may be difficult to determine what an appropriate real life domain may be. In the process of the C-test development, we may have to consider both real life tasks and language instructional tasks bases for developing the test tasks in the C-test. That is, the results of the C-test may be used to decide test takers’ future
to receive ESL instruction or to be exempt from it. At the same time, the test is related to
teaching and learning of the target language as the test takers will be placed in ESL
program(s) based on the test results and will need to participate in the classroom, or in the
academic domain of school.

In order to describe the characteristics of specific TLU tasks which could probably be
related to the C-test task, at first, we need to identify what kinds of TLU tasks are likely
relevant to the test takers in the academic domain of school. Then, we need to select TLU
tasks which might be considered potential test task(s) on the C-test and, then, describe the
characteristics of those selected TLU tasks; this will help us see the correspondence between
those TLU tasks in real life and the C-test task. Describing the characteristics of those TLU
tasks will be a basis for discussing authenticity of the C-test later in this study. After the
characteristics of the TLU tasks are identified, they need to be evaluated in terms of their
potential contribution to the usefulness of the test. That is, we need to compare the
characteristics of one TLU task with another to find out if there are any distinctive
characteristics overlapped among those TLU tasks; this will help us identify and describe the
characteristics of the C-test task.

For the test development, TLU tasks are restricted to tasks requiring written language
ability including reading and writing skills. Written language ability, which will be defined in
the section of the definition of the construct, is considered crucial to carry out academic tasks
in English-medium mainstream classes along with spoken language ability including listening
and speaking skills. The possible tasks in the TLU domain involving written language ability
are taking notes, writing a term paper, writing an essay exam, writing homework assignment,
taking tests, reading text books, reading instructions, reading schedules, etc., which are the tasks necessary to participate in the classroom. For the purpose of this study, the TLU tasks are narrowed down into the following two tasks involving written language ability: reading a text book (for answering questions provided in each chapter) as TLU task 1 and taking a test (e.g., a fill-in-the-blank type test) as TLU task 2. These two TLU tasks are likely related to the C-test task as the C-test is supposed to measure the test taker’s ability to use written language based on some degree of reading and limited production of writing.

I will describe the characteristics of the above two TLU tasks, or TLU tasks 1 and 2, in terms of the characteristics of the setting, characteristics of input, characteristics of expected response, and relationship between input and expected response. In doing that, I will use a check list (see Table 2.1) suggested by Bachman and Palmer (1996: 108).

In that frame work, the characteristics of the setting include physical characteristics, participants, and time of task. Physical characteristics can be examined in terms of location, noise level, temperature and humidity, seating conditions, lighting, and materials and equipment. The characteristics of input can be explained in terms of format: channel, form, language, and type. The characteristics of expected response will be explained in terms of format: channel, form, language, and type. Relationship between input and response can be described in terms of reactivity, scope, and directness.
Table 2.1. TLU checklist

<table>
<thead>
<tr>
<th>TLU TASK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristics of the setting:</strong></td>
</tr>
<tr>
<td>Physical setting</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>Time of task</td>
</tr>
</tbody>
</table>

| **Characteristics of input:** |
| Format |
| Channel |
| Form |
| Language |
| Type |

| **Characteristics of expected response:** |
| Format |
| Channel |
| Form |
| Language |
| Type |

| **Relationship between input and response:** |
| Reactivity |
| Scope |
| Directness |

**TLU task 1: reading a text book (for answering questions provided in each chapter)**

1. Characteristics of the setting: the characteristics of the setting include physical characteristics, participants, and time of task. Physical characteristics include location, noise level, temperature and humidity, seating conditions, lighting, and
materials/equipment, etc. Location will be in a classroom, at the library, or at home. Noise level will vary: that is, it may be quite at the library and somewhat quiet or noisy at home. Temperature and humidity will vary. Seating conditions will vary: individual desks and chairs or a big table shared with other students. Lighting will vary including well lit. Materials and equipment include text book(s), papers, and pens for answering questions. Participants will be familiar with materials and equipment. Participant(s) will be a student, himself/herself. Time of task will vary including daytime, evening, and/or weekend(s).

2. Characteristics of input: the characteristics of input will be explained in terms of format which includes channel, form, language, and type. Channel is visual. Form is mainly written language but may include figures like maps, charts, or graphs. Language in use is the target language, English, appropriate for academic discipline in high school. As participant(s) will be student(s) to learn/review knowledge from reading before and/or after topics are presented in school, language use is expected to be somewhat limited to topic areas.

3. Characteristics of expected response: the characteristics of expected response will be explained in terms of format examining channel, form, language, and type. Channel is visual to be read. Form is mainly language. Language in use is the target language, English. Participants are expected to use language appropriate for the academic domain of school, demonstrating grammatical knowledge including syntactic and vocabulary knowledge to organize sentences and textual knowledge to form texts if necessary. Language use is expected to be appropriate for topic(s) and the grade level. Type of expected response is answering chapter questions with selected item(s), limited
production, and/or somewhat extended production. Expected responses can be produced at the lexical level, sentence level, and/or paragraph level depending on specific questions.

4. Relationship between input and response: it includes reactivity, scope, and directness.
Reactivity is non-reciprocal as there is no reaction or interaction between them. Scope is considered broad. Relationship is considered direct to a great extent as answers are expected to be found or inferred from the input.

**TLU task 2: Taking a fill-in-the-blank test**

1. Characteristics of the setting: the characteristics of the setting include physical characteristics, participants, and time of task. Physical characteristics can be examined in terms of location, noise level, temperature and humidity, seating conditions, lighting, and materials/equipment. Location will be in classrooms. Noise level is expected to be low unless there is noise from outside classrooms. Temperature and humidity are expected to be comfortable in most cases. Seating conditions will be the same for all participants including individual desks and chairs. Lighting will be well lit. Materials and equipment include test booklet(s) and pen(s) to take the test. Participants will be familiar with them. Participant(s) will be students and a teacher; all of who are likely to be familiar to each other. Time of task will be daytime during school hours.

2. Characteristics of input: the characteristics of input will be explained in terms of format examining channel, form, language, and type. Channel is visual to be read. Form is mainly written language; it is unlikely to have figures. Language in use is the target language,
English, appropriate for academic discipline in high school. As topic(s) are expected to be presented in school, language use is expected to be general but topic related.

3. Characteristics of expected response: the characteristics of expected response will be explained in terms of format which includes channel, form, language, and type. Channel is visual to be read. Form is mainly language. Language in use is the target language, English. Participants are expected to use language appropriate for the academic domain of school, mainly demonstrating grammatical knowledge including syntactic and vocabulary knowledge to organize individual sentences. Type of expected response is limited production like a word or phrase level answer required for each specific blank.

4. Relationship between input and response: it includes reactivity, scope, and directness.

Reactivity is non-reciprocal as there is no reaction or interaction between them. Scope is considered narrow. Relationship is considered direct only to some degree; answers may be or may not be directly found from the input and they may be inferred from the input.

Table 2.2 shows the summary of the characteristics of both TLU tasks 1 and 2.

### Characteristics of the test takers

The characteristics of the test takers need to be examined in the process of test development as we want to develop a test appropriate for the target test takers who may take the C-test. The characteristics of the test takers will be described in terms of personal characteristics, topical knowledge of test takers, levels of language knowledge of test takers, and possible affective responses to taking the test.
<table>
<thead>
<tr>
<th>Description of a task in TLU</th>
<th>TLU TASK 1</th>
<th>TLU TASK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristics of the setting:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical characteristics:</td>
<td>Classroom, library, home.</td>
<td>Classroom.</td>
</tr>
<tr>
<td>Location</td>
<td>Varied: quite at library, somewhat quiet or noisy at home.</td>
<td>Quite.</td>
</tr>
<tr>
<td>Noise level</td>
<td>Varied.</td>
<td>Usually comfortable.</td>
</tr>
<tr>
<td>Temperature/humidity</td>
<td>Varied: individual desk, chair, etc.</td>
<td>Good: individual desk, chair.</td>
</tr>
<tr>
<td>Seating conditions</td>
<td>Varied: likely well lit</td>
<td>Well lit.</td>
</tr>
<tr>
<td>Lighting</td>
<td>A text book, papers, pens to answer questions. Students will be familiar with them.</td>
<td>Test booklets, pencils, pens. Students will be familiar with them.</td>
</tr>
<tr>
<td>Materials/equipment</td>
<td>Student.</td>
<td>Students, teacher; they are likely familiar with each other.</td>
</tr>
<tr>
<td>Participants</td>
<td>Varied: daytime, evenings, or weekends.</td>
<td>During school hours.</td>
</tr>
<tr>
<td>Time of task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics of input:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>Written; charts, graphs.</td>
<td>Written.</td>
</tr>
<tr>
<td>Form</td>
<td>Target, topic specific.</td>
<td>Target.</td>
</tr>
<tr>
<td>Language</td>
<td>Text.</td>
<td>Text.</td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Characteristics of expected response:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>Written.</td>
<td>Written.</td>
</tr>
<tr>
<td>Form</td>
<td>Target.</td>
<td>Target.</td>
</tr>
<tr>
<td>Language</td>
<td>Answering chapter questions: limited, or extended production.</td>
<td>Limited production: a word or phrase level production.</td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationship between input and response:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reactivity</td>
<td>Non-reciprocal.</td>
<td>Non-reciprocal.</td>
</tr>
<tr>
<td>Scope</td>
<td>Relatively broad.</td>
<td>Narrow.</td>
</tr>
<tr>
<td>Directness</td>
<td>Direct.</td>
<td>Relatively indirect.</td>
</tr>
</tbody>
</table>
Personal characteristics

Test takers' personal characteristics will be described in terms of age, gender, nationalities, native language(s), immigrant status, length of residence in the U.S., level of general educational background, and preparation or prior experience with ESL proficiency test(s). These characteristics will affect the usefulness of specific test tasks as test developers or designers would want to develop a test for a specific target group in a particular setting. That is, in order to make the most appropriate test task(s) for that testing situation, it is necessary to develop a test for a specific TLU domain described above and specific test takers in mind as personal characteristics might affect the test performance of test takers.

Test takers are expected to be non-native English speaking students in high school in ages between 14 and 19 including both males and females. Their nationalities and native languages might be widely varied. Their immigrant status may also be varied: some might be legal or illegal immigrants; others may be temporary residents. Length of their residence in the U.S. is expected to be varied ranging in between a few months and a few years. Levels of their general educational background may be varied. Some may have received general education equivalent to their grade levels in the U.S. or in their native countries; other little or no formal education. Most of test takers are unlikely familiar with ESL proficiency test(s).

Topical knowledge

Test takers’ topical knowledge might be in a relatively wide range as their general educational levels and length of their residence in the U.S. are expected to be varied. If test
takers have received general education appropriate for their ages, it is expected that some may be familiar with several topics on the test and others might be familiar with a few at least.

Levels of language knowledge

Levels of language knowledge of test takers are expected to be varied from the beginning to the advanced level based on length of their residence in the U.S. and levels of general education. Levels of written language ability may be varied also. Some may demonstrate a lower level of written language ability but may produce relatively good oral communication skills. Others may show a high level of written language ability but may demonstrate a relatively lower level of oral communication skills.

Possible affective responses to taking the test

Possible affective responses from the test takers may be varied depending on their written language ability. Lower level test takers are likely to feel frustration about taking the test as they probably think that they are not going to perform well on the test and/or if they may be under stress due to their language ability. High level test takers may be relatively positive in taking the test as they might be more confident in their language ability.

The construct to be measured on the C-test

Before we examine the test tasks in the C-test, we need to define the construct to be measured on the C-test. The construct is what the test is intended to measure. We have to define the construct of a test in its particular testing situation for the following three purposes (Bachman and Palmer, 1996: 116):
1. to provide a basis for using test scores for their intended purpose(s),
2. to guide test development efforts, and
3. to enable the test developer and user to demonstrate the construct validity of these interpretations.

In order to define the construct of the C-test, it is necessary to define language ability as a basis for defining the construct to be measured in the C-test. I will use a theoretical model of language ability proposed by Bachman (1990, quoted in Bachman and Palmer, 1996: 67) as it is a useful framework for defining the construct for this testing situation. In this framework, language ability is an ability which enables us to use language properly in a particular setting. Language ability consists of language knowledge and strategic competence, or metacognitive strategies. It is the combination of language knowledge and metacognitive strategies that may provide language users with the ability to produce and interpret the target language discourse such as in responding to tasks on language test(s) or in real life situation(s). In this framework, language knowledge includes two broad categories: organizational knowledge and pragmatic knowledge; each category can be divided into several areas. Table 2.3 summarizes the areas of language knowledge as presented in Bachman and Palmer (1996: 68). In fact, most of language tests including the C-test may focus on only one or a few of the areas of language knowledge. In the following section, the areas of language knowledge and metacognitive strategies will be explained, respectively.
Table 2.3. Areas of language knowledge

<table>
<thead>
<tr>
<th>ORGANIZATIONAL KNOWLEDGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(how utterances or sentences and texts are organized)</td>
</tr>
<tr>
<td>Grammatical knowledge</td>
</tr>
<tr>
<td>(how individual utterances or sentences are organized)</td>
</tr>
<tr>
<td>Knowledge of vocabulary</td>
</tr>
<tr>
<td>Knowledge of syntax</td>
</tr>
<tr>
<td>Knowledge of phonology/graphology</td>
</tr>
<tr>
<td>Textual knowledge</td>
</tr>
<tr>
<td>(how utterances or sentences are organized to form texts)</td>
</tr>
<tr>
<td>Knowledge of cohesion</td>
</tr>
<tr>
<td>Knowledge of rhetorical or conversational organization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRAGMATIC KNOWLEDGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)</td>
</tr>
<tr>
<td>Functional knowledge</td>
</tr>
<tr>
<td>Knowledge of ideational functions</td>
</tr>
<tr>
<td>Knowledge of manipulative functions</td>
</tr>
<tr>
<td>Knowledge of heuristic functions</td>
</tr>
<tr>
<td>Knowledge of imaginative functions</td>
</tr>
<tr>
<td>Sociolinguistic knowledge</td>
</tr>
<tr>
<td>Knowledge of dialects/varieties</td>
</tr>
<tr>
<td>Knowledge of registers</td>
</tr>
<tr>
<td>Knowledge of natural or idiomatic expressions</td>
</tr>
<tr>
<td>Knowledge of cultural references and figures of speech</td>
</tr>
</tbody>
</table>

Language knowledge: organizational knowledge and pragmatic knowledge

Organizational knowledge is involved in controlling the formal structure of language for producing or comprehending grammatically acceptable utterances or sentences, and for organizing these to form texts, both spoken and written (Bachman and Palmer, 1996: 67). According to this framework, there are two areas of organizational knowledge: grammatical
knowledge and textual knowledge.

Grammatical knowledge is involved in producing or comprehending formally accurate utterances or sentences. This includes knowledge of vocabulary, syntax, phonology, and graphology. Textual knowledge is involved in producing or comprehending texts, which are units of language that consist of two or more utterances or sentences. There are two areas of textual knowledge: knowledge of cohesion and knowledge of rhetorical or conversational organization. Knowledge of cohesion is involved in producing or comprehending the explicitly marked relationships among sentences in written texts or among utterances in conversations. Knowledge of rhetorical or conversational organization is involved in producing or comprehending organizational development in written texts or in conversation.

In the C-test, grammatical knowledge may be required more than any other knowledge areas as the test elicits lexical level responses within contextual constraints. In addition, textual knowledge is also required to fill in the blank appropriately, which will help test takers understand each text in context and use textual knowledge as a basis for correct response(s) appropriate in context. That is, test takers need to understand the text cohesively in order to use appropriate tenses of verbs, appropriate plurality of nouns, etc.

The next area of language knowledge is pragmatic knowledge. Pragmatic knowledge enables us to create or interpret discourse by relating utterances and texts to their meanings, to the intentions of language users, and to relevant characteristics of the language use setting. There are two areas of pragmatic knowledge: functional knowledge and sociolinguistic knowledge.
Functional knowledge enables us to interpret relationships between utterances or sentences and texts and the intentions of language users. There are four categories of language functions: knowledge of ideational, manipulative, instrumental, and imaginative to use language to express or interpret meaning in terms of our experience of the real world (e.g. explanations, and expressions of sorrow or anger, etc.), to affect the world around us (e.g. suggestions, warnings, compliments, apologies, etc.), to extend our knowledge of the world around us (e.g. teaching, learning, problem-solving, etc.), and to create an imaginary world for humorous or aesthetic purposes (e.g. jokes, use of figurative language, poetry, etc.), respectively. Usually language use involves the performance of multiple functions in context of connected discourse, not in isolated utterances.

Sociolinguistic knowledge enables us to create or interpret language appropriate for a particular language use setting; this includes knowledge of the conventions for appropriate use of dialects, registers, idiomatic expressions, figures of speech, etc.

Pragmatic knowledge may be not strongly required to produce appropriate response(s) on the C-test but is expected to be used whenever it is applicable—for example, when there are blank(s) of idiomatic expression(s).

Strategic competence or metacognitive strategies

Strategic competence, or metacognitive strategies can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities. Using language involves the language user's topical knowledge and affective schemata, as well as all the areas of language knowledge discussed above. There
are three areas where metacognitive strategies operate: goal-setting, assessment, and planning.

In the C-test, goal-setting, assessment, and planning strategies are necessary to complete the test successfully. These strategies, however, will not be explicitly included in the definition of the construct as we want to measure written language ability of test takers, not the metacognitive strategies. Therefore, we want to present explicit instructions and example(s) in order to eliminate potential misuses of metacognitive strategies.

The definition of the construct to be measured in the C-test

The construct to be measured in the C-test is written language ability based on theory-based construct definitions. The theory-based construct definitions are based on a theoretical model of language ability as discussed above, rather than the contents of a language teaching syllabus. As stated earlier, the purpose of the test is making decisions about selection and placement of LEP students, not assessing the achievement of specific syllabus objectives. Making decisions will be based on test takers' performances on a written test, the C-test, where the type of the input and expected responses is written language; we do not expect test takers to use their ability of spoken language on the C-test. By defining the construct as written language ability, we exclude spoken language ability which needs to be assessed using a test designed to measure spoken language ability.

Even though we are interested only in measuring language ability in the C-test, we could not simply exclude metacognitive strategies and topical knowledge in the definition of the construct of the C-test. However, the degree of involvement of metacognitive strategies and topical knowledge should be minimal in order to not obscure performance of language
ability. That is, even though metacognitive strategies are likely involved in taking a test to some degree, it is necessary to reduce inappropriate uses of test takers' metacognitive strategies on the test by presenting explicit instructions and examples. In addition, the type of metacognitive strategies required in the C-test is considered simple and straightforward. First, the test taker needs to read the instructions and examples. Then, the test taker is required to read each text and fill in the blanks. In restoring the missing parts of words, the test taker may utilize various clues including clues from existing parts of words or clues from long range contextual constraints. Therefore, in taking the C-test, the type of metacognitive strategies may not be different from one test taker to another. Regarding involvement of topical knowledge, even though we do not want to measure the test takers' language ability affected by their topical knowledge, it is expected that the test takers have general topical knowledge appropriate for the high school academic domain. By presenting various topics, however, it is necessary to try to reduce the effect of topical knowledge on the test takers' scores obtained from the test.

The test task in the C-test

In determining what characteristics the test task and the test will have, we need to consider distinctive characteristics of the TLU task types, the purpose of the test, the definition of the construct to be measures, and the resources that are available for test development and use, which will be a basis for discussing qualities of test usefulness. Figure 2.1 shows how design components are related to each of qualities of test usefulness in developing test tasks as illustrated in Bachman and Palmer (1996: 172).
In order to consider the characteristics of the test task, first, we need to identify distinctive characteristics of TLU tasks 1 and 2 as a basis of the test task. They are as follows: the location is likely a classroom; the input is written English appropriate for a high school academic domain; the expected response is limited production of written English for
word-level answers; the scope of the relationship between the input and expected response is moderate to narrow depending on the text. The C-test task, therefore, will have these general characteristics of TLU tasks. Students, however, may not face the same type of C-test task, filling in half of a word. For the complete details of C-test task, therefore, the characteristics of C-test task will be described based on specifications of the test itself. The specifications of a test task include:

1. the specific purpose of the test task,

2. the definition of the construct to be measured, and

3. the framework of test task characteristics.

The purpose of the C-test and the definition of the construct on the C-test are the same as the ones described above as there is only one type of test task on the C-test. The characteristics of C-test task will be described basically in the same framework used for describing the characteristics of TLU tasks. The framework of test task characteristics, therefore, includes the characteristics of the test setting, the test rubric, the input, and the expected response, and relationship between input and response. In the following section, the test task characteristics of the C-test will be described in the framework.

**Characteristics of the setting**

The characteristics of the setting include physical setting, participants, and time of task. Regarding physical characteristics, the locations are to be varied including a classroom, an office, or other quite areas that can be arranged in the school. The noise level is varied but it is expected to be low. Temperature and humidity are expected to be comfortable but we
expect some seasonal variances like a hot summer day at the end of a spring term or at the beginning of a fall term. Seating conditions are individual desks and chairs to take a test comfortably. The room is well lit. The materials and equipment include a large clock/a timer, test booklets, pencils, erasers, and a pencil sharpener if possible. Test takers are likely to be familiar with these materials and equipment which are used in the classroom and at home.

The participants include test taker(s) and test proctor(s). Test taker(s) are newly enrolled high school student(s) with their first language(s) other than English. Test proctor(s) are teachers and/or school administrator(s) in the school, and/or ESOL consultant(s) in a local Area Education Agency. Test proctors need to be familiar with the test and need to be informed and be trained about how to give the test; the instructions for test proctor will help. (See Appendix A for the instructions for test proctor/rater.)

Time of task may vary because each school administers the test depending on its own needs. It is likely to administer the test at anytime during school hours. In order to avoid pulling test taker(s) out from their mainstream classes, we recommend testing either in the morning before class or in the afternoon after class. For the convenience of test takers and school(s), however, the test is likely given during school hours. In any case, testing in the morning is recommended for better performance(s) of test taker(s) as test taker(s) are likely tired at the end of a day.

**Characteristics of the test rubric**

The characteristics of the test rubric include the structure of the test, instructions, the duration of the test, and how the language that is used will be evaluated or scored.
Regarding structure, the test task consists of six parts. Each part is indicated with an alphabetic letter (e.g., A, B, etc.) or a number (e.g., 1, 2, etc.) so that each part is clearly distinguished from one another. The order of the parts is fixed; however, test takers may move from one to another within the given time constraint. Each part carries equal weight in the final score. There are 20 blanks for each part and, therefore, 120 blanks for the whole C-test. However, the first part will be a practice passage so that it will not be scored; by presenting one easy practice passage, we expect that test takers may have a better chance to be familiar with the test task. Five parts following the practice one are to be scored to produce a final score for the whole C-test.

Instructions are given in the target language, English. As the test measures test takers' language ability, we expect test takers to understand the instructions in English. As test takers have a wide variety of native languages, we want to ensure consistency and fairness for all test takers. The test takers read the instruction as it is read aloud by the proctor. Therefore, channel of the instructions is visual and aural. The instructions explain what the test takers are expected to do on the test. Sample text(s) are presented following the instruction in the same page. The test task is asking test takers to complete missing parts of words in the test texts.

Regarding the duration of time, or time allotment, it is necessary to secure a fifty minute period for a whole test: thirty minutes for taking six parts of the test, three to five minutes for the instructions, and time for other necessary administration procedure--to hand out and collect the test booklet(s)/pencil(s), answer questions if any, etc.
The scoring method includes criteria for correctness, procedures for scoring the responses, and explicitness of criteria and procedures. The criterion for correctness is to measure written language ability based on theory-based definition of the ability to use language in the task. The test task is to be scored objectively. Correctness will be judged based on test takers’ limited production responses; test takers restore missing parts of words on the test tasks. There are three different scoring methods available: the exact, acceptable, and spelling-error tolerance scoring methods. With the exact scoring, only exact words found in the original texts are considered correct responses. With the acceptable scoring, appropriate alternatives are also considered correct responses. For example, depending on a particular context, 'cold' can be an acceptable answer for 'cool' and 'this' for 'thg.' With the spelling-error tolerance scoring, minor spelling-errors are also considered correct responses—e.g., 'planing' for 'planning.' Test takers’ responses will be scored ‘1’ (one point) for a correct answer depending on which scoring method is used or ‘0’ (zero point) for an incorrect or no written answer. The total sum of ‘1’s on the whole C-test will be the score of the test taker on the test. For now, we recommend the acceptable scoring method to be included in the instructions for test proctor/rater in order to ensure the objectivity (see Appendix A.)

Regarding procedures for scoring response(s), limited production responses are read and scored by a single rater according to a scoring key which provides exact and expected acceptable answers. If unexpected alternative(s) are found, the rater may contact either the school’s ESL coordinator or an ESL consultant at a local Area Education Agency for the decision.
Regarding explicitness of criteria and procedures, test takers are not explicitly informed about the scoring criteria; this is to make the instructions as simple as possible. Even though the scoring criterion is not mentioned in a separate part, it is fully implied in the instructions and sample text(s) because the characteristics of expected responses are explicitly presented in the instructions and sample text(s). It is, therefore, unlikely for test takers to misunderstand the scoring criteria on the test.

**Characteristics of input**

Input consists of the material contained in a given test. The material is described in terms of format, language characteristics, and topical characteristics.

Format has to do with the way in which the input is presented and includes the following characteristics: channel, form, language, length, type, degree of speededness, and vehicle. Channel is visual. Form is language. Language is the target language, English, because we want to test English. Length of the task is considered medium. Each part of the task is a simple text with 70-100 words. Type of input is a text for the task. The task is not designed to require speededness; however, test takers need to complete the task within time constraint which is necessary to ensure the consistency and fairness for all test takers. Within the time constraint, test takers need to allocate appropriate time for reading the input in order to save time for writing responses. The time limit is considered sufficient for processing or understanding the input for most test takers. Vehicle is considered not live as it is a written test.
Language of input is related to language characteristics and topical characteristics. Regarding language characteristics, there are organizational and pragmatic characteristics. Organizational characteristics include grammatical and textual characteristics. Grammatical characteristics are general vocabulary, morphology, and syntax of written academic English for secondary school students; language is typewritten. Regarding textual characteristics, input is cohesive in the test task. The six parts of the test task are organized orderly. Following the instructions in the first page, the test task is given in the second page through the fourth page: two parts per page. Pragmatic characteristics include functional and sociolinguistic characteristics. There are no specific functional characteristics in the task. As the materials are from text books or supplementary text books likely used in the classroom, sociolinguistic characteristics are written academic language for young readers, moderately formal register, natural language, and no explicit cultural language.

Topical characteristics are general and various appropriate for the academic domain of school; passages presented are from text books or supplementary text books at the secondary school level.

Characteristics of the expected response

The characteristics of the expected response include format, language characteristics, and topical characteristics.

Format includes channel, form, language, length, type, and speededness. Channel is visual. Form is language. Language is the target language, English. Length is short as the expected response is part of a word for each blank. Type is limited production, restoring part
of a word in each blank. The expected response of the test task is not designed for speededness; however, test takers have to finish the task within the given time constraint.

Language characteristics include organizational and pragmatic characteristics. Regarding organizational characteristics, test takers are expected to use general vocabulary, morphology and syntax of written academic English appropriate in the classroom. As the expected response is part of a word with one to several letters for each blank, textual cohesion is not expected. Pragmatic characteristics include functional and sociolinguistic characteristics. There are no specific functional characteristics except filling in the blank, a simple function to restore part of a word in the context of the immediate and/or extended environment. Sociolinguistic characteristics are academic written dialect appropriate for the classroom in high school, moderately formal, natural language, no specific cultural references.

No specific topical characteristics are expected in response. As topics are general and various, topical characteristics may vary in each topic to some degree.

**Relationship between input and response**

Relationship between input and response include reactivity, scope of relationship, and directness of relationship. Reactivity is non-reciprocal because there is no reaction for any response at the time of testing. Scope of relationship is considered narrow to medium. That is, some responses are affected by their immediate environment, the first half of a word; other responses require more understanding in the distant environment and/or the context of the whole passage. Test takers with higher-level language ability are more likely utilizing the more extended environment. Relationship is considered direct only to some degree: that is,
the relationship varies from one blank to the next and from the content of the text. In producing responses, clues from input can be explicit and direct for some blanks and implicit and indirect in other blanks.

**The logical evaluation of C-test usefulness**

The previous sections in this chapter were discussed as a basis for examining the logical evaluation for C-test usefulness. In this part, the usefulness of C-test will be examined based on logical evaluation which involves value judgments. All six qualities of test usefulness will be examined; this will answer the six research questions from the perspective of logical evaluation. The results from empirical study will provide more evidence of the C-test usefulness for reliability, construct validity, impact, and practicality in Chapter 4.

Bachman and Palmer (1996) suggest a framework consisting of various questions pertaining to examining systematic facets of logical analysis for the six qualities of test usefulness. The questions are designed to examine the logical evaluation of the test usefulness involving value judgments; we need to judge the extent to which facets of a quality are satisfied and how. (See Appendix B for complete details for questions and answers for logical evaluation of test usefulness.)

Based on the answers for to what extent qualities are satisfied, we could conclude the following:

1. Reliability of the test is considered considerably high based on examining systematic facets of reliability--the test setting, test rubric, test input, expected response, and relationship between input and response. The questions ask to what extent the systematic facets vary
in an unmotivated way from one part of the test to another, or on different forms of the
test. Based on the logical evaluation, the characteristics of most systematic facets do not
vary in an unmotivated way from one part of the test to another. However, we may
expect that the characteristics of some facets might vary in an unmotivated way to some
extent (e.g., the time of testing, physical setting, topical knowledge, etc.) as they likely
vary in each testing situation. Nevertheless, in general, qualities of most systematic facets
of reliability are satisfied to a great extent, which provides evidence for the research
question 1 from the perspective of logical analysis.

2. Construct validity is considered fairly high in terms of clarity and appropriateness of the
construct definition, and the appropriateness of the task characteristics with respect to the
construct definition, which provides evidence for the research question 2 from the
perspective of logical analysis. However, there may be possible sources of bias in the task
characteristics to some extent. It has to do with the familiarity of the task and the topical
knowledge. Explicit instructions and example passages are considered helpful to raise the
familiarity of the task. In order to get rid of possible bias from topical knowledge, widely
varied topical texts are really recommended. If texts can be screened by several ways,
sources of bias may be eliminated. The level of texts needs to be examined prior to actual
test uses by making number of native speakers in high school take the test and collecting
the results. This process is considered important to avoid problems caused by bias related
to the texts.

3. Authenticity is considered relatively medium. Authenticity of the test task itself may be
relatively low as the test takers may not face the same fill-in-the-blank task in the academic
domain of school. However, the test takers may likely encounter similar types of topics and written language presented on the test in the classroom as the passages on the test are real materials from text books and supplementary text books at the secondary school level. Also, even though the test task requires test takers to produce limited production of written language at the lexical level, the test takers have to read and understand passages at the lexical level and the textual level in order to write correct responses. This process, which requires the test takers to utilize ability of written language for reading, understanding, and writing, could be considered very similar to the process of TLU tasks in the academic domain of school. Therefore, in general, authenticity concerning the correspondence between the TLU tasks and the test task is considered medium, which answers the research question 3.

4. Interactiveness based on the logical evaluation is considered medium. Even though the test task requires limited use of language function, the test takers need to utilize the ability of written language to write responses; the test takers have to read and understand the task in written English in order to produce correct responses. Also, in the process of test taking, metacognitive strategies are required as the C-test task requires interaction among components of language knowledge to some degree. Therefore, as the C-test task is interactive to some degree, the interactiveness is considered medium; this answers the research question 4.

5. Impact is considered fairly positive in general, which answers the research question 5 from the perspective of logical evaluation. The consequences of taking the C-test may be relevant to test takers, teachers, and schools to a great. The results of the test may affect
the future of test takers for ESL selection and placement. Decision-makers (e.g., teachers and schools) may need to adjust ESL program(s) based on the selection and placement. In addition, no specific negative consequence is concerned for now.

6. Based on the logical evaluation upon systematic facets of practicality, practicality is considered very high as many of existing resources, either human or material resources or both, are most likely available at all three stages: the design stage, operationalization stage, and administration stage. This provides evidence for the research question 6.

As mentioned before, the qualities of test usefulness should be examined to provide evidence to maximize the overall usefulness of the test at hand. Based on the logical evaluation of test usefulness of the C-test, the C-test can be considered useful for the purpose of the test to be developed. Even though authenticity and interactiveness of the C-test are considered relatively medium, we found more positive evidence for the other four qualities, or reliability, construct validity, impact, and practicality. That is, the C-test should be considered reliable and practical to use to a great extent; the C-test is also considered providing evidence for the construct validity and generating positive impact. Based on logical evaluation of test usefulness, the C-test should be considered useful for the situation in this study. In the following two chapters, the C-test usefulness will be continuously examined by providing more evidence for the reliability, construct validity, impact, and practicality based on empirical evaluation. The methods of empirical research will be addressed in Chapter 3; then, the results will be discussed in Chapter 4.
CHAPTER 3. METHODS

In Chapter 2, I discussed the process of test development. In the process, test usefulness was examined by doing logical analysis, involving value judgments, of individual qualities of test usefulness; that provided some evidence for test usefulness of the C-test by answering all six research questions from the perspective of logical evaluation. This chapter will discuss methods of empirical research. The empirical research will provide more evidence for the four qualities of test usefulness: reliability, construct validity, impact, and practicality. This chapter will discuss the subjects used for the pilot and main test in the study and the procedure used to develop the C-test materials. This chapter will also discuss the procedure used for giving the C-test and analyzing the results. The results of this empirical research will be presented and, then, discussed in the next chapter.

Subjects

Both non-native and native speakers participated in the pilot and main tests. It was desired to have both groups in a same school; the comparison of the test results from both groups could make more sense as both groups may receive a similar type/level of educational input appropriate for the grade level in the school. The pilot and main tests were administered in two different schools.

The subjects for the pilot test were 5 non-native speakers and 22 native speakers including both male and female enrolled in a same high school in Iowa. Non-native speakers were a mixed grade group ranging from the 9th grade to in the 11th grade at the time of testing.
All native speakers were in the 9th grade. According to the teacher, the proficiency levels of non-native speakers were considered varied and so was the English language ability of native speakers. The C-test was given at the end of the spring 1997 semester.

The subjects for the main test were 30 non-native speakers and 19 native speakers enrolled in another high school in Iowa including both male and female. Proficiency levels of non-native speakers varied from the beginning level to the advanced level. Non-native speakers' first languages also varied as they were from various countries: South American countries, Asian countries, Eastern Europe countries, etc. Their general educational background is expected to vary as they were put in the grade based on their ages. Non-native speakers were a mixed grade group of students in the 9th grade through the 12th grade: 7, 9, 7, and 7 students in the 9th, 10th, 11th, and 12th grade, respectively. The school offers an ESL program. Most of non-native speakers received ESL instruction; a few of them previously received ESL instruction but were no longer in the ESL program. Native speakers were in the 11th grade. The C-test was given at the middle of the fall 1998 semester.

The research in this study was approved by the Iowa State University Human Subjects Committee and by the two participating high schools. The subjects volunteered to participate in the research.

Materials

The pilot test

In order to create a C-test for the pilot test, first, passages were collected from various topics which are likely used in the classroom. We did not want passages which may require
very specific topical knowledge; or we did not want passages which may produce rather tricky deletion(s). The passages were not related each other; they had independent topics (see Appendix C for guidelines to construct a C-test).

All passages were from actual text books and supplementary books likely used in the classroom; some of them were from middle school level books and the others from high school level books. Passages were basically gathered at a high school reading level with general/basic topical knowledge relevant to the academic domain of school, preferably at a 9th grade level. Some passages were from a middle school level materials, which was considered reasonable because high school students should have little or no difficulty in processing those passages. Theoretically, using their written language ability, a majority of 9th graders in high school are expected to understand the content with no or little difficulty when they take the test. However, due to the characteristics of the C-test, they may miss some of answers or write incorrect responses at their language developmental stage.

The purpose of the pilot test was to select passages, or superitems appropriate for the main test. It is necessary to define the concept of superitems for further discussion. The concept of superitems provides a theoretical background for constructing the C-test and, therefore, is a basis for estimating the reliability of the C-test. The superitem method is based on the following assumption (Raatz, 1985: 64):

Authentic test A consists of c authentic parts A₁, A₂, ... to Aₖ. Assuming that all the parts are independent of each other, but are equivalent and measure the same thing, then the total test score is the sum of the part scores. These parts can be viewed as superitems. In this case one can calculate intercorrelations and discrimination indices for the superitems without going inside the test parts....This is the way in which we estimate the reliability of the C-Test.
For the pilot test, it was necessary to have at least 10 passages so that best suitable passages could be selected for the main test. First, approximately 40 passages were screened by the length and the content; among them, 23 passages were selected and screened again before deletion by two native speakers and me. Then, 10 passages were finally selected to construct a C-test for the pilot test. In the C-test, each passage, or superitem, had 20 blanks; so, 10 superitems had 200 blanks in total. Deletion procedures were completed following the rule of 2: starting the second word in the second sentence, the second half of every other word was deleted. If a word has an odd number of letters, one more letter was deleted. After 20 deletions were made, the remaining text was left intact. (See Appendix C for the guidelines to construct a C-test for more details.) Finally, 10 superitems were constructed and, then, screened by filling in missing letters in each blank to find out if deletions have problems. In order to produce a test booklet for the pilot test, 10 superitems were presented in the order of expected difficulty, from the easiest one to the most difficult, based on the judgmental decision. Instructions with an example text was presented before the actual test passages. A title was given to each passage to show explicitly test takers that each passage had a different topic.

The pilot test was administered in a high school in Iowa at the end of the spring 1997 semester, May 1997. Test takers took the test for this research study. All native speakers took the test at the same time in a classroom during a regular class period. After receiving instructions, native speakers were allowed 50 minutes to complete the test with 10 passages. Most of them finished in 30-35 minutes. Non-native speakers took the test individually during a self-study period within a few days after native speakers took the C-test. It was impossible
to give non-native speakers the test at the same time as they had different class schedules. As all test materials were collected after native speakers took the test, it was unlikely that non-native speakers may have got information about the content of the test.

The results from the pilot test were scored by hand using three scoring methods: the exact scoring, acceptable scoring, and spelling-error tolerance scoring; these scoring methods were explained in the section of the test rubric in describing the characteristics of the C-test task in Chapter 2. The Microsoft Excel for Windows, Version 7.0 was used to compute descriptive statistics for both non-native and native speakers. SPSS procedures were used to estimate the reliability. KR-20 was calculated to estimate a reliability coefficient for the internal consistency based on the native speakers’ scores with the exact scoring method because we had a few non-native speakers. As we wanted to select passages easy enough for native speakers with their written language ability and with little or no specific topical knowledge, the exact scoring method was considered reasonable for screening passages with tricky or confusing deletions. The overall results of the pilot test are presented below; the discussion will follow. Table 3.1 shows the descriptive statistics for the C-test scores from non-native speakers (N=5); Table 3.2 shows the descriptive statistics for native speakers’ C-test scores (N=22).

For the reliability analysis, the internal consistency was estimated at the superitem level as explained above. The reliability coefficient was .939 for the C-test scores from native speakers with the exact scoring method. Therefore, based on this internal consistency analysis, all superitems in the C-test were considered very consistent in measuring the construct which would be expected to be measured.
For the main test, five passages were selected. The five passages were five easiest ones for the native speakers across all three scoring methods: texts 1,4,5,7, and 10 (see Table 3.2). I chose passages which were easy enough for the native speakers because it may be assumed that those easy ones may require little or no specific topical knowledge in the C-test. As mentioned in Chapter 2, we are interested in measuring test takers’ language ability, not their topical knowledge. For the non-native speakers, five easiest ones were texts 2, 4, 7, 8, and 10 (see Table 3.1). Therefore, across two groups, texts 4, 7, and 10 were commonly easiest ones. By including passages which were easiest ones for native speakers,

Table 3.1 The C-test scores from the pilot test (non-native speakers: N=5)

<table>
<thead>
<tr>
<th></th>
<th>Exact scoring</th>
<th>Acceptable scoring</th>
<th>Spelling-error t. scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(%)</td>
<td>SD</td>
<td>Min</td>
</tr>
<tr>
<td>C-test</td>
<td>70.40(35.2)</td>
<td>21.62</td>
<td>38</td>
</tr>
<tr>
<td>Text 1</td>
<td>6.40(32.0)</td>
<td>2.33</td>
<td>4</td>
</tr>
<tr>
<td>Text 2</td>
<td>7.60(38.0)</td>
<td>3.38</td>
<td>1</td>
</tr>
<tr>
<td>Text 3</td>
<td>4.80(24.0)</td>
<td>1.47</td>
<td>3</td>
</tr>
<tr>
<td>Text 4</td>
<td>8.00(40.0)</td>
<td>3.58</td>
<td>2</td>
</tr>
<tr>
<td>Text 5</td>
<td>5.80(29.0)</td>
<td>2.23</td>
<td>4</td>
</tr>
<tr>
<td>Text 6</td>
<td>6.60(33.0)</td>
<td>3.61</td>
<td>2</td>
</tr>
<tr>
<td>Text 7</td>
<td>9.00(45.0)</td>
<td>3.52</td>
<td>5</td>
</tr>
<tr>
<td>Text 8</td>
<td>7.80(39.0)</td>
<td>3.31</td>
<td>4</td>
</tr>
<tr>
<td>Text 9</td>
<td>4.80(24.0)</td>
<td>1.33</td>
<td>3</td>
</tr>
<tr>
<td>Text 10</td>
<td>9.60(48.0)</td>
<td>3.93</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 3.2 The C-test scores from the pilot test (native speakers: N=22)

<table>
<thead>
<tr>
<th>Text</th>
<th>Exact scoring</th>
<th></th>
<th>Acceptable scoring</th>
<th></th>
<th>Spelling-error t. scoring</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(%)</td>
<td>SD</td>
<td>Min</td>
<td>Max</td>
<td>Mean(%)</td>
<td>SD</td>
</tr>
<tr>
<td>C-test</td>
<td>141.95(71.0)</td>
<td>31.25</td>
<td>82</td>
<td>190</td>
<td>149.05(74.5)</td>
<td>30.73</td>
</tr>
<tr>
<td>Text 1</td>
<td>14.82(74.1)</td>
<td>3.30</td>
<td>6</td>
<td>19</td>
<td>16.14(80.7)</td>
<td>2.53</td>
</tr>
<tr>
<td>Text 2</td>
<td>14.00(70.0)</td>
<td>3.44</td>
<td>9</td>
<td>18</td>
<td>15.09(75.5)</td>
<td>3.54</td>
</tr>
<tr>
<td>Text 3</td>
<td>10.86(54.3)</td>
<td>3.31</td>
<td>4</td>
<td>17</td>
<td>13.14(65.7)</td>
<td>3.52</td>
</tr>
<tr>
<td>Text 4</td>
<td>14.91(74.6)</td>
<td>3.18</td>
<td>8</td>
<td>20</td>
<td>15.14(75.7)</td>
<td>3.28</td>
</tr>
<tr>
<td>Text 5</td>
<td>14.95(74.8)</td>
<td>2.92</td>
<td>10</td>
<td>20</td>
<td>15.27(76.4)</td>
<td>2.93</td>
</tr>
<tr>
<td>Text 6</td>
<td>14.36(71.8)</td>
<td>4.76</td>
<td>3</td>
<td>20</td>
<td>14.41(72.1)</td>
<td>4.78</td>
</tr>
<tr>
<td>Text 7</td>
<td>16.32(81.6)</td>
<td>2.70</td>
<td>10</td>
<td>20</td>
<td>16.45(82.3)</td>
<td>2.57</td>
</tr>
<tr>
<td>Text 8</td>
<td>12.82(64.1)</td>
<td>5.26</td>
<td>1</td>
<td>19</td>
<td>13.55(67.8)</td>
<td>5.46</td>
</tr>
<tr>
<td>Text 9</td>
<td>12.95(65.0)</td>
<td>4.47</td>
<td>3</td>
<td>19</td>
<td>13.91(69.6)</td>
<td>4.55</td>
</tr>
<tr>
<td>Text 10</td>
<td>15.91(79.8)</td>
<td>4.56</td>
<td>0</td>
<td>20</td>
<td>16.00(80.0)</td>
<td>4.58</td>
</tr>
</tbody>
</table>

and easy and rather difficult ones for non-native speakers, we expect the following from the results of the main test: the main test could likely be very easy for native speakers and, at the same time, discriminate among non-native speakers’ language ability.

The order of texts for the main test was decided based on the difficulty of texts observed from the results from the non-native speakers with the spelling-error tolerance scoring method. The spelling-error scoring method was considered reasonable for test takers, or high school students, as language ability of non-native and native speakers may still be at
the developmental stage. A minor spelling error of a word could be considered acceptable even in the classroom unless students are taking a spelling test. It was expected to be helpful for non-native speakers to begin with an easy passage and, then, increase the difficulty in order.

The five passages were considered a little bit difficult even for the native speakers. That is, the mean scores of the pilot test were ranging from 15.38 (76.9%) to 16.32 (81.6%) with three scoring methods; this means that a test taker may miss 5 to 4 answers out of 20 per passage. Therefore, one adjustment was made to make passages a little easier. That is, the most difficult blank in each passage was filled in with an original word in a text; instead, another deletion was made after the final deletion to keep 20 blanks in each passage. It was also examined that the new deletion may unlikely be the most difficult one in each passage. This was done based on judgmental decision by comparing the new deletion with existing deletions. Even though it was uncertain if this adjustment could possibly raise scores, it was an effort to make passages a little easier. In addition, in order to help test takers be familiar with the C-test, a practice passage at an elementary level was included to be given as the first passage. However, I decided not to notify test takers that the first one is a practice passage because, if test takers know that the first one is a practice passage, they may not try the passage at all.

Therefore, in the main test, total six passages were presented in the order of expected ascending difficulty: a practice passage based on theoretical judgment; 5 texts based on the results of the pilot test with a little adjustment. Also, the instructions and sample texts were adjusted to make more explicit than those presented in the pilot test. Each passage had 20
blanks; therefore, the entire C-test for the main test had 100 blanks in total. (See Appendix D for the sample C-test material; this was the C-test material used for the main test.) The titles of passages were as follows:

1. Text 1 (Practice passage): How do plants form seeds?
2. Text 2: First Law of Motion
3. Text 3: Friendship
4. Text 4: Money Management
5. Text 5: Food
6. Text 6: Industrial Revolution

**Procedure**

The main test was given in a high school in Iowa in the middle of the fall 1997 semester, October 1997. Test takers took the test for this research study. It was given simultaneously to both native and non-native speakers in two classrooms. Native speakers and two non-native speakers took the test during a regular class in a classroom where a teacher and an ESOL consultant administered the test. Non-native speakers except the two took the test in another classroom where I administered the test. For most of non-native speakers, it was their regular ESL class; however, some of them came to take the test during their mainstream classes. After receiving instructions together in each classroom, test takers read the passages and wrote their answers in blanks. The test takers were allowed 30 minutes in completing 6 passages. Most of the native speakers finished in 15-20 minutes; most of the non-native speakers finished in 20-25 minutes.
Analysis

The results from the main test were scored using all three scoring methods and, then, transferred into a data set for SPSS procedures. Possible maximum points were 20 for each superitem and 100 for the whole C-test. SPSS procedures were used to compute descriptive statistics, reliability coefficients, correlation coefficients, and the t-test. KR-20 was used for estimating reliability coefficients at the superitem level as explained before. The reliability coefficients were estimated for the scores of the non-native speaker group that was the target group of this testing setting.

For the correlational study, within the non-native speaker group, the Pearson-product moment correlation coefficients were estimated between the C-test scores and the teacher-rated proficiency levels. For the empirical item analysis, a qualitative research method was used. That is, responses errors of non-native speakers were examined in order to find patterns of errors and, then, grouped into the two major categories: semantically motivated response errors and syntactically motivated response errors. Semantically motivated response errors are errors that are likely motivated by the meaning of particular vocabulary but incorrectly produced due to incorrect or unanalyzed syntactical and vocabulary knowledge. The types of observable errors are mis-spellings (e.g., 'gravaty' for 'gravity') and inflectional errors (e.g., 'knew' for 'know'). Syntactically motivated response errors are errors that might be associated with the clues of the remaining letters of words in blanks. This type of errors may be response errors that show the correct use of inflections and spellings but the incorrect use of vocabulary; therefore, observable errors are not relevant to the meaning of correct
responses (e.g., 'final' for 'financial,' 'prepare' for 'prevent,' etc.) The group difference between native and non-native speakers was also calculated by the t-test.
CHAPTER 4. RESULTS AND DISCUSSION

In this chapter, test results from the empirical study will be presented in providing
evidence for reliability, construct validity, impact, and practicality. For the reliability analysis,
the internal structure of the C-test will be examined by observing the internal consistency
reliability of the entire C-test. Construct validity will be examined by providing content
evidence, correlational evidence, empirical item analysis, and experimental research. Impact
and practicality will be discussed by examining available information collected at the
operationalization and administration stages. The overall results of the C-test (non-native
speakers) will be discussed first.

The Overall Results (non-native speakers)

The descriptive statistics are provided based on the C-test scores with three different
scoring methods: the exact scoring, acceptable scoring, and spelling-error tolerance scoring
(see Table 4.1). Figure 4.1 shows the frequencies for the C-test scores based on the spelling-
error tolerance scoring method. Table 4.2 compares the C-test distribution to the normal
distribution based on the frequencies shown in Figure 4.1.

In Table 4.1, the C-test mean score is the mean of test takers’ total scores on the C-
test excluding the scores on the practice passage. The scores from the practice passage are
included in order to be compared with the scores of other superitems, or passages. As shown
in Table 4.1, for the whole C-test, the spelling-error tolerance scoring produced the highest
mean scores across the three scoring methods; the exact scoring method produced the lowest.
Table 4.1. Descriptive statistics for the C-test scores (non-native speakers: N=30)

<table>
<thead>
<tr>
<th></th>
<th>Exact scoring</th>
<th>Acceptable scoring</th>
<th>Spelling-error t. scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(%)</td>
<td>SD</td>
<td>Min</td>
</tr>
<tr>
<td>C-test</td>
<td>47.70(47.7)</td>
<td>20.23</td>
<td>8</td>
</tr>
<tr>
<td>Practice</td>
<td>12.27(61.4)</td>
<td>4.70</td>
<td>1</td>
</tr>
<tr>
<td>Text 2</td>
<td>11.87(59.4)</td>
<td>4.43</td>
<td>0</td>
</tr>
<tr>
<td>Text 3</td>
<td>11.00(55.0)</td>
<td>4.56</td>
<td>3</td>
</tr>
<tr>
<td>Text 4</td>
<td>8.50(42.5)</td>
<td>4.17</td>
<td>2</td>
</tr>
<tr>
<td>Text 5</td>
<td>7.60(38.0)</td>
<td>5.08</td>
<td>1</td>
</tr>
<tr>
<td>Text 6</td>
<td>8.73(43.7)</td>
<td>4.88</td>
<td>0</td>
</tr>
</tbody>
</table>

The C-test scores (non-native speakers)

Figure 4.1. Bar graph for the frequencies for the C-test scores based on the spelling-error tolerance scoring method (non-native speakers: N=30)
Table 4.2. Comparison between the C-test distribution and the normal distribution (ND).

<table>
<thead>
<tr>
<th>SD</th>
<th>-3S</th>
<th>-2S</th>
<th>-1S</th>
<th>M</th>
<th>+1S</th>
<th>+2S</th>
<th>+3S</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test raw scores</td>
<td>11.5</td>
<td>32.1</td>
<td>52.7</td>
<td>73.4</td>
<td>94.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal distribution¹</td>
<td>2%</td>
<td>14%</td>
<td>34%</td>
<td>34%</td>
<td>14%</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>C-test distribution²</td>
<td>3%</td>
<td>17%</td>
<td>27%</td>
<td>33%</td>
<td>20%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Difference in percentages(%)³</td>
<td>+1</td>
<td>+3</td>
<td>-7</td>
<td>-7</td>
<td>+6</td>
<td>-2</td>
<td></td>
</tr>
</tbody>
</table>

¹ The percentages are between standard deviations under the normal distribution.
² The percentages are between standard deviations under the C-test distribution.
³ The differences are the percentage differences between the normal distribution and the C-test distribution.

At the superitem level, the spelling-error tolerance scoring produced the highest mean scores and the exact spelling produced the lowest.

Table 4.1 also shows that the scores of the superitems, or texts, conformed to our expectation about the order of presenting the superitems, from the easiest to the most difficult, to a great extent. The practice passage, or text 1, was the easiest one across the three scoring methods, followed by text 2. With the spelling-error tolerance scoring, the order of texts was the order of difficulty from the easiest to the most difficult. With the other two scoring methods, the mean scores did not follow the expected order of the difficulty; however, texts 2 and 3 were definitely easier ones and texts 4, 5, and 6 were rather difficult ones, which conformed to our expectation to a great extent. As shown in Figure 4.1, the C-test scores from the non-native
speakers produced a distribution similar to the normal distribution in an NRT. If we compare the C-test distribution to the normal distribution in Table 4.2, we find that differences of percentages between standard deviations in both distributions are not much. That is, between -1S and +1S, we expect to have 68% of students in the normal distribution; actually we have 60% of students in the same range under the C-test distribution. If we look at the percentages of students below and above the mean, we have 47% of students and 53%, below and above the mean, respectively. Also in Table 4.1, the mean scores of the entire C-test were approximately in the 50% range of the maximum score on the C-test which is supposed to be ideal in an NRT: 47.7%, 50.3%, and 52.7%, with the exact scoring, the acceptable scoring, and the spelling-error tolerance scoring, respectively. These observations suggest that the C-test may work as an NRT to a great extent for the target group.

Reliability

As reliability concerns the extent to which patterns of empirical internal consistency realize theoretical expectations, the behaviors on the test must reflect performance, or response consistency. In the C-test, each text is considered a superitem, which is a basis in estimating reliability coefficients of the C-test (Raatz, 1985; Huhta, 1996) as explained before. Therefore, the superitem method contributed to produce these high reliability coefficients of the C-test at the textual level, not at the lexical level.

In this study, the internal consistency was estimated to show to what extent the C-test is reliable. The reliability coefficients among five super items were estimated between .926 and .930 with all three scoring methods (see Table 4.3); this means that the C-test produced
very high reliability coefficients. Therefore, with any of three scoring methods, the C-test can be considered highly reliable. These high coefficients confirm that test takers’ performances on different superitems of the C-test were internally consistent. In other words, in the C-test, only approximately 7% are interpreted as random variance; approximately 93% of variance among the superitems are considered consistent variance. Therefore, it may be said that the C-test measures a construct consistently.

### Construct validity

**Content evidence**

The first step in investigating construct validity will be in examining content evidence. Content evidence refers to the judgments of experts concerning the ability that test items measure (Brown, 1996; Chapelle, 1994). The C-test was designed to measure written language ability which consists of language knowledge and metacognitive strategies based on a theoretical definition of the construct that is supposed to be measured. This theoretical analysis was examined by experts in the process of test development and administration.

### Table 4.3. Reliability coefficients of the C-test (KR-20)

<table>
<thead>
<tr>
<th>Scoring method</th>
<th>Reliability coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact scoring</td>
<td>.926</td>
</tr>
<tr>
<td>Acceptable scoring</td>
<td>.930</td>
</tr>
<tr>
<td>Spelling-error tolerance scoring</td>
<td>.927</td>
</tr>
</tbody>
</table>
At the initial development stage as a course project for the testing course in the fall of 1996, three graduate students in Teaching English as a Second Language (TESL) in the department of English were test designers. One of the group members was a former Spanish teacher in a high school; another one was an ESOL consultant at a local AEA; I as a non-native speaker of English have been in EFL/ESL situations for a long time. These three members could be considered experts.

They evaluated the C-test for the purpose of test development and agreed that the C-test may be useful as a test of written language ability. Their judgment was based on systematic content analysis by hypothesizing the abilities required by the facets of the C-test method: the test setting, the test rubric, the input to the test taker, the expected response, and the relationship between input and response (Bachman and Palmer, 1996). These facets were considered useful to examine content evidence by hypothesizing the possible effect of each facet as the test taker interacts with the test in a real situation. At first, the test taker encounters the test setting like a classroom. Then, the test taker receives the explicit test instructions which help him or her set a goal and plan for the test. After the test taker understands the instructions, he or she looks at the C-test passages as input and writes answers by restoring missing parts of words (i.e., expected output). The relationship between the input and expected response can give us clues for the specific language abilities required on the C-test. The input on the C-test will require grammatical knowledge (e.g., knowledge of vocabulary, syntax, morphology, etc.) and textual knowledge for comprehending the text in addition to basic reading ability. This input either as the first half of the word or as the context will narrow down the scope of possible responses for each blank. The test taker’s
reading ability may help to find a correct response for each blank to some extent. The
expected response is filling in missing part of a word in each blank; the test takers need to
restore the missing part by utilizing the first half of the word and/or finding implicit or explicit
clues in the given text. This hypothesis helped the group members' decision to develop the C-
test as a test of written language ability.

At the operationalization stage, a version of the C-test and the instructions for proctor
were sent to outside experts who were several consultants and the director in the Educational
Services Division in a local AEA after the initial development. Even though they have little or
no experience with ESL instruction, they are the ones who help schools improve education
environment in the areas of their expertise. All of them had no problem in considering the C-
test a test of written language ability as suggested; the director was positive in presenting the
C-test to school administrators as a new test which may be useful for ESL programs if logical
and empirical analyses can provide positive evidence for the C-test usefulness for this
particular setting.

At the administration stage for the pilot and main tests, two teachers in high schools
were asked their opinions about the C-test. After the pilot test was administered, the English
teacher in that school was asked about her opinion about the content of the C-test. She did
not provide any negative response using the C-test as a "High School English Proficiency
Test" which was the title given on the C-test. When the main test was administered in another
school, the Language Arts/ESL teacher seemed to recognize the C-test as a variation of the
cloze test method, which is actually the basis of the C-test method development, and showed a
sample test constructed with the cloze method. Even though she did not mention whether the
cloze test was currently used or not, the similarity between the cloze test and the C-test seemed to help her accept the C-test as a “High School English Proficiency Test.” The above observations suggest that two teachers judged the C-test as a proficiency test of high school English.

**Correlational evidence**

A correlational study is another way of examining evidence for construct validity. In order to do a correlational study, I obtained a list of the proficiency levels of NNSs from the teacher in the school. Because the data of two NNSs were missing in the list, this correlational study was conducted excluding those two students (N=28). The proficiency levels of NNSs are considered the teacher rated proficiency levels, one of criteria recommended for the correlational study (Klein-Braley and Raatz, 1984).

The proficiency levels of NNSs consist of four levels: the beginning, intermediate level, advanced level, and mainstream levels. If an NNS is in the mainstream level, he/she does not need to take any ESL class or only needs to take one ESL class for a subject area such as Math. The NNSs in each level were 2, 9, 3, and 14, for the beginning, intermediate, advanced, and mainstream levels, respectively. Table 4.4 shows the Pearson product-moment correlation between the C-test and the proficiency levels.

As shown in Table 4.4, the C-test scores with all three scoring methods are quite highly correlated with the teacher rated proficiency levels, particularly given the imprecise information available about proficiency levels. There is basically no difference among correlation coefficients across all three scoring methods.
Table 4.4. Correlation with the proficiency levels between the C-test and the proficiency levels (non-native speakers: N=28)

<table>
<thead>
<tr>
<th>C-test</th>
<th>Proficiency levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact scoring</td>
<td>r=.760</td>
</tr>
<tr>
<td>Acceptable scoring</td>
<td>r=.762</td>
</tr>
<tr>
<td>Spelling-error tolerance scoring</td>
<td>r=.789</td>
</tr>
<tr>
<td>(P=.000)</td>
<td></td>
</tr>
</tbody>
</table>

It may be possible to do the correlational study further if it is possible to get more information about the construct of the proficiency levels as we do not know exactly what elements constitute the construct in assessing the proficiency levels.

**Empirical item analysis**

Empirical item analysis is one way in examining evidence for construct validity. For the empirical item analysis, test takers’ response errors were analyzed. The response errors were analyzed at the lexical level by observing errors. Therefore, the results of this analysis may reveal some characteristics of response errors at the lexical level, rather than at the superitem level. Mainly two types of response errors were observed: semantically motivated errors and syntactically motivated errors. Table 4.5. shows the examples of semantically motivated response errors. Table 4.6 shows the examples of syntactically motivated response errors. The semantically motivated errors could be divided into two types of response errors which may show semantically motivated, but show aspect of incorrect, incomplete, or
Table 4.5. Semantically motivated responses indicating aspects of vocabulary knowledge

<table>
<thead>
<tr>
<th>Incorrect, incomplete or unanalyzed knowledge</th>
<th>Observable error</th>
<th>Examples of errors</th>
<th>Correct responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthographic</td>
<td>Mis-spelling</td>
<td>gravaty</td>
<td>graVITY(^1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>physicaly</td>
<td>physiCALLY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>creaks</td>
<td>creEKS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>getting</td>
<td>getTING</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pressure</td>
<td>presSURE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>watter</td>
<td>waTER</td>
</tr>
<tr>
<td>Morphemic</td>
<td>Inflectional error</td>
<td>push/pushed/pushing</td>
<td>pushES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>knows/knew</td>
<td>knOW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>machine</td>
<td>machINES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>river</td>
<td>rivERS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bought</td>
<td>bUY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>have</td>
<td>hAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>secret</td>
<td>secRETS</td>
</tr>
</tbody>
</table>

\(^1\)Capitalized letters are the missing part of word in the C-test, which is applied to the other examples.

Table 4.6. Syntactically motivated responses cued by the remaining letters of words in blanks

<table>
<thead>
<tr>
<th>Examples of errors</th>
<th>Correct responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>final</td>
<td>finaNCIAL(^1)</td>
</tr>
<tr>
<td>prepare</td>
<td>preVENT</td>
</tr>
<tr>
<td>plan</td>
<td>plAY</td>
</tr>
<tr>
<td>from</td>
<td>fOR</td>
</tr>
<tr>
<td>form</td>
<td>foRCE</td>
</tr>
<tr>
<td>gradually</td>
<td>graVITY</td>
</tr>
</tbody>
</table>

\(^1\)Capitalized letters are the missing part of word in the C-test, which is applied to the other examples.
unanalyzed knowledge (Chapelle, 1994): spelling errors and inflectional errors. Most of response errors were semantically motivated. Another type of response errors were syntactically motivated errors; test takers seemed to guess answers based on the remaining parts of words. These errors were not semantically correct but syntactically appropriate with correct spellings and appropriate inflections.

The above examples suggest that the written language ability of non-native speakers may be at the developmental level. Therefore, some of non-native speakers whose responses were semantically motivated may need more time to develop syntactic and vocabulary knowledge. Others whose responses were syntactically motivated may need more time to acquire more vocabulary knowledge. This analysis provides evidence for the construct validity.

**Evidence from experimental research**

Another way arguing for construct validity will be in examining evidence from experimental research which would provide evidence to verify that observed test performance behaves in concert with theory-based predictions. In this study, it was expected when test performances on a same C-test from both non-native speakers and native speakers were compared, native speakers would outperform non-native speakers. At the same time, it was predicted that as native speakers’ language ability could be considered still at the developmental level, their performances on the C-test may not be perfect. That is, the language ability of those students may need more time to be developed near the language ability of adult educated native speakers, a criterion reference of the C-test, as mentioned
earlier. For example, those students may still need more time to develop their language knowledge, part of language ability. Some may need to develop their lexical knowledge; others their syntactic knowledge. In order to verify these two theory-based predictions, an experimental study was conducted. The same C-test was administered to native speakers (N=17) at the same school at the same time when the C-test was administered to non-native speakers. The descriptive statistics for the C-test scores of native speakers are presented below (see Table 4.7). Figure 4.2 shows the frequencies for the C-test scores from the native speakers based on the spelling-error tolerance scoring method.

One thing has to be mentioned about the test takers. At the time of testing, there were 19 native speakers in the classroom. However, the results from the two of them were excluded from the statistical analysis because it was obvious that the two test takers did not consider the C-test seriously at all. That is, one did not try the C-test at all so that it was impossible to score it. The other one tried most of the C-test; however, his/her performance shows that he/she did not consider the C-test seriously. He/she added word(s), deleted letter(s) or part of word(s), and put silly or bad word(s) in blanks.

If we look at the overall results in Table 4.7, the native speakers’ scores also show the same pattern both at the entire C-test level and at the superitem level as observed in the results of the non-native speakers. At the entire C-test level excluding the practice, the spelling-error tolerance scoring produced the highest mean score; the exact spelling scoring produced the lowest. At the superitem level, with the acceptable word scoring and spelling-error tolerance scoring, the mean scores did perfectly conform to the order of expected difficulty: practice (text 1), texts 2, 3, 4, 5, and 6, from the easiest one to the most difficult, which satisfies our
Table 4.7. Descriptive statistics for the C-test scores (native speakers: N=17)

<table>
<thead>
<tr>
<th></th>
<th>Exact scoring</th>
<th>Acceptable scoring</th>
<th>Spelling-error t. scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(%)</td>
<td>SD</td>
<td>Min</td>
</tr>
<tr>
<td>C-test</td>
<td>79.53(79.5)</td>
<td>10.65</td>
<td>55</td>
</tr>
<tr>
<td>Practice</td>
<td>18.00(90.0)</td>
<td>1.82</td>
<td>13</td>
</tr>
<tr>
<td>Text 2</td>
<td>17.24(86.2)</td>
<td>2.18</td>
<td>12</td>
</tr>
<tr>
<td>Text 3</td>
<td>17.00(85.0)</td>
<td>2.17</td>
<td>13</td>
</tr>
<tr>
<td>Text 4</td>
<td>15.71(78.6)</td>
<td>2.24</td>
<td>10</td>
</tr>
<tr>
<td>Text 5</td>
<td>14.35(71.8)</td>
<td>3.68</td>
<td>7</td>
</tr>
<tr>
<td>Text 6</td>
<td>15.18(75.9)</td>
<td>3.20</td>
<td>10</td>
</tr>
</tbody>
</table>

The C-test scores (native speakers)

Figure 4.2. Bar graph for the frequencies for the C-test scores based on the spelling-error tolerance scoring method (native speakers: N=17)
expectation again that the easiest text should come first on the C-test to help test takers' performances. Even though the mean scores with the exact word scoring did not conform to the exact order of the expected difficulty as in the other two scoring methods, the order of the expected difficulty of texts was very similar to that with the other scoring methods. Across all three scoring methods, the practice passage and texts 2 and 3 were definitely easier ones and texts 4, 5, and 6 were rather difficult ones.

Compared to the C-test scores of non-native speakers (see Table 4.1), the native speakers definitely outperformed the non-native speakers both at the entire C-test level and at the superitem level across all three scoring methods (see Table 4.7); this confirms our first prediction. Figure 4.2 shows that the frequency of the C-test scores, which is negatively skewed as in a CRT; most of scores are near the right end of the score range, or higher scores. The C-test scores of the non-native speakers, on the other hand, produced a distribution similar to a normal distribution as discussed above. As Table 4.7 shows, the native speakers did not score (near) perfect, which confirms the second prediction. That is, with the spelling-error tolerance scoring which produced the higher mean scores at the entire C-test level, the mean is 86.2% with the minimum of 60% and the maximum of 95%. These findings confirm our two theory-based predictions.

However, we have to be cautious in comparing the C-test scores of native speakers to those of non-native speakers because both groups were not in the same grade. Whereas the native speakers are in the 11th grade, the non-native speakers were a group of mixed grades from the 9th through the 12th, and more than half of them, 16 out of 30, were in the lower grades than the native speakers. Therefore, the native speakers may have had an advantage in
taking the C-test in the area of topical knowledge, language ability (language knowledge and metacognitive strategies), etc. On the other hand, some of native speakers were not very serious in taking the C-test so that they seemed to rush to finish the test; they knew that the C-test results would not affect their grades. Therefore, it is difficult to say that the above situation may have affected the results of the C-test from the native speakers to what extent by just observing the raw scores.

In order to see the group differences between native and non-native speakers, the t-test was computed. Table 4.8 shows the mean scores and standard deviations of the whole C-tests for both native and non-native groups and the result of the t-test to see the significance of the differences between the means. The scoring method used in the t-test analysis was the spelling-error tolerance scoring.

<table>
<thead>
<tr>
<th></th>
<th>Native speakers</th>
<th>Non-native speakers</th>
<th>T-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>C-test</td>
<td>86.24(86.2%)</td>
<td>9.14</td>
<td>52.73(52.7%)</td>
<td>20.65</td>
</tr>
</tbody>
</table>

The result of the t-test shows that the C-test can differentiate between the two groups: that is, the C-test was quite easy for the native group but was rather difficult for the non-native group. Therefore, even though the above situation may have caused the performances
of the native group in a way, their performance was significantly different from the performances of the non-native group.

This experimental study satisfies theory-based predictions to great extent even though we may consider the above caution in interpreting the results from both groups. Therefore, this empirical study seems to provide evidence for the construct validity.

**Impact**

Evidence for impact will be provided by examining information collected at the operationalization and administration stages. Information was collected by asking involved individuals their opinions about the C-test. Those individuals include an ESOL consultant, several other consultants, and a director in the Educational Services Division in a local AEA in Iowa. In addition to them, there were two teachers and students who were involved in the pilot and main tests.

Impact on the educational system will be addressed by examining opinions of the consultants and the director in the Educational Services Division in a local AEA. I would like to mention who the consults are and what their roles are, which will explain why their opinions could be considered important. They all are team managers for school improvement teams in area schools; each has a specific area of expertise and research. They all work together on school improvement issues and know the numbers of LEP students in that area are increasing. They are staff members in the Educational Services division where the mission is to develop and sustain a caring for all learners (Iowa Department of Education, 1997). As the AEAs in Iowa “function as the intermediate unities among the Iowa Department of
Education, school districts, and local schools” (Iowa Department of Education, 1996b), the AEAs help the Department of Education work with districts and schools, and vice versa. The ESOL consultant was one of the group members involved in the initial test development; she was very willing to show the C-test to the other consultants and the director in the Educational Services division in a local AEA. Even though the other consultants, whose areas of expertise are not in ESL but in Math, staff development, and Language Arts, have little or no experience with ESL instruction, their opinions could be considered relevant to the C-test research which may play a role in improving educational environment for all the learners including both LEP and mainstream students. Therefore, it may be said that their opinions may reflect a positive impact on the educational system.

Basically, there were few responses/comments from the consultants in the Educational Services Division in a local AEA when the C-test was introduced to them as a test of written language ability. There were no negative responses indicating problematic areas in the C-test itself and the instructions for test proctor/rater. There was a response saying that it looks okay, though. The director liked the idea of developing a new test which may be useful for ESL instruction and wanted to the results of the C-test research; she became very supportive in administering the C-test. The ESOL consultant was very devoted to contacting schools and teachers to get their permission to administer the pilot and main test. After schools had been chosen for the pilot and main tests, the AEA was willing to state their support of the C-test research in letters to parents of students in those schools. This evidence may suggest a positive impact on the educational system.
Impact on individuals will be addressed by examining opinions and responses of teachers and students who were involved in the pilot and main tests. There were two teachers, one for each test, who helped administer the C-test and liked to see the result of the C-test, which could be considered a positive attitude toward the C-test to some extent.

There were students’ responses from both native and non-native speakers in the pilot and main tests. Most of the native speakers who participated in the pilot test wanted to get their scores so that the score reports were sent to them. Some were eager to ask correct answers for some blanks in the C-test soon after the test was finished. This may suggest that those students saw the C-test as a test which would measure their language ability. There were a few non-native speakers who took the C-test in the pilot test. Their scores and the copies of the tests were sent to them upon the teacher’s request; this may suggest that the teacher and the non-native speakers probably accepted the C-test as a test measuring students’ language ability as intended. Some of students including both NSs and NNSs, who participated in the main test, wanted to get the copies of their tests so that the copies of scored C-test were sent to them. The fact that a number of the test takers wanted to know their scores on the C-test supports that the C-test was considered a legitimate test for those high school students even though a few students did not take it seriously as mentioned earlier. The collected responses from those individuals could be considered positive in accepting the C-test as a test which would measure their language ability.

In addition, impact on individuals may need to be considered more closely related to teaching and learning situations. In doing that, we could consider possible impacts, both positive and negative, on teaching and learning in using the C-test. Possible positive impact
on teaching and learning could be using the test results to help the test takers be aware their weakness especially related to their lexical or grammatical knowledge. By analyzing response errors, teachers may find out patterns of errors for each test taker such as errors of tenses (e.g., 'know' for 'knew' for a past tense), spelling-errors, errors of subject-verb agreement (e.g., 'think' for 'thinks' in 'he thi_____'), problems of contractions (e.g., 'its' for 'it’s'), etc. Analyzing responses errors could be used as quick and initial diagnostic tools and, therefore, could be used to raise the student’s awareness in using English. This could be possible as the C-test may be given to a few students at a time in a normal situation so that a teacher may not need to spend much time in analyzing response errors of a student. On the other hand, there may be also possible negative impact on teaching and learning in using the C-test. Even though the test taker needs to utilize his/her language ability including reading writing in the C-test, the scope of using reading and writing skills is quite limited. Therefore, in order to overcome the test taker’s perception of reading and writing from taking the C-test, teachers may need to emphasize or provide more elaborated reading and writing practices afterwards as reading and writing skills are crucial tools to succeed in high school academic domain.

These qualitative observations, therefore, suggest that impact on both the educational system and individuals could be considered positive to a great extent for now. When there may be a possibility of negative impact, it should be important to not ignore the possibility but improve the situation.
Practicality

Practicality will be addressed by examining information collected from the operationalization and administration stages. At both stages, there were no problems in using existing human resources and material resources. First, we have to examine the use of existing resources in an AEA at the operationalization stage. Regarding human resources, there was an ESOL consultant who contacted school and teachers for the pilot and main tests; in addition, there was clerical support in making copies of the C-test and letters to parents. Regarding material resources, it was possible to use copy machine(s) and copy papers to make the test booklets, and a phone to contact schools/teachers to seek permission in participating in the C-test research and to set up testing dates and other details. As a result, it may be said that it was possible to use existing resources at this stage.

Then, we need to examine the use of existing resources in schools and/or AEAs when the pilot and main test were administered. Regarding human resources, we need administrator(s) and rater(s). In the pilot test, both the ESOL consultant and I were administrators. In the main test, it was necessary to have more than two administrators as the C-test was simultaneously administered to two groups of students in two different classrooms. So, both the ESOL consultant and the teacher were administrators in a classroom; I was an administrator in another classroom. Regarding material resources, existing resources were used: that is, there was no problem to use testing supplies and equipment like desks, pencils, and classrooms in the schools. Regarding scoring, we need rater(s) after administering the C-test. For this study, all the C-test were scored by me as researcher. I had no problem scoring them following the instruction and the answering key; teachers or ESOL consults may have no
difficulty scoring the C-test with the answering key in the future as all the answers are at the lexical level. For scoring the C-test of 100 blanks (except the practice) given in the main test, it took less than 5-7 minutes for each student. It may require a little more time for scoring if an rater needs to consult an ESOL consultant in a local AEA when a response is acceptable but is not included in the answering key. As a result, it may be said that it was possible to use existing resources at this stage. As examined above, the C-test was practical to carry out to a great extent both at the operationalization and administration stages by utilizing existing resources.

So far in this chapter, the results of empirical research were examined to provide evidence for reliability, construct validity, impact, and practicality. We found positive evidence for those qualities of test usefulness. In the next chapter, the C-test usefulness will be summarized by answering all six research questions; this will be based on the empirical evaluation of the C-test usefulness discussed in this chapter in addition to the logical evaluation discussed in Chapter 2.
CHAPTER 5. CONCLUSIONS

In this chapter, I will summarize the answers to the research questions presented in introduction. Then, I will address recommendations/implications for the future use of the C-test for ESL instruction in high school and for further research.

The answers to the research questions will be summarized based on the logical evaluation of test usefulness discussed in Chapter 2 and based on the empirical study of test usefulness discussed in Chapter 4. In order to examine the C-test usefulness as a test of written language ability, six research questions were asked at first. I will restate the six research questions for the convenience of discussion as follows:

1. Is the C-test reliable to measure written language ability of ESL students?
2. Does empirical analysis provide construct validity evidence for the C-test?
3. To what extent is the C-test similar to authentic tasks in target language use domain?
4. To what extent is the C-test task interactive?
5. What would be the impact on individuals (test takers and teachers) and education systems/society? Is the impact positive?
6. Is the C-test practical to use?

Summary of answers to research questions

The first question asked whether the C-test is reliable to measure written language ability of ESL students in high school. For the logical evaluation, we examined to what extent the systematic facets of reliability (e.g., characteristics of the test setting, the test setting, the
test input, the expected response, and the relationship between input and response) vary in an unmotivated way from one part of the test to another, or on different forms of the test. Based on the logical evaluation, reliability of the C-test was considered considerably high as most of the facts were satisfied to a great extent. The empirical study of reliability was examined also. The reliability coefficient was very high, or approximately .93, which suggests that all superitems in the C-test were consistent in measuring the construct, or written language ability: that is, we expect only approximately 7% of random variance. This finding was also consistent with reports of other C-test studies.

The second question asked if the empirical analysis provided construct validity evidence for the C-test. Before we look at the results of empirical analysis, I would like to mention the results of logical evaluation first as the logical evaluation provided useful information about the construct definition and possible sources of bias in the task characteristics related to the construct. In terms of the clarity and appropriateness of the construct definition and appropriateness of the task characteristics with respect to the construct definition, the construct was clearly defined based on theoretical judgments and the task characteristics were considered appropriate in measuring the construct, written language ability. Some possible sources of bias, however, in the task characteristics were mentioned concerning the familiarity of the task and the topical knowledge. Therefore, in order to eliminate those sources of bias, explicit instructions along with sample passages and widely varied topics were included when the C-test was constructed.

Then, the empirical study provided more evidence for the construct validity by examining content evidence, correlational evidence, empirical item analysis, and evidence from
experimental research. As mentioned earlier, content evidence refers to the judgments of experts concerning the construct, written language ability in this study. Content evidence was examined by hypothesizing the abilities required by the systematic facets of the C-test. The systematic facets include the test setting, the test rubric, the put, the expected response, and the relationships between the input and expected response. Hypothesizing the abilities required by those facets was done at the primary test development stage by me and two other graduate students in the Department of English at Iowa State University and, then, reexamined by me for this research study. In addition, two teachers involved in administrating the pilot and main tests had no problem in approving the C-test as a high school English proficiency test measuring language ability for working with academic learning situations in high school, which was stated in the instructions included in the C-test booklet. Therefore, the above analysis provided evidence for the construct validity of the C-test.

For the correlational study, the C-test scores from non-native speakers and the teacher rated proficiency levels were compared. The correlation coefficients were estimated: .760 to .789 with all three scoring methods, which was considered high enough to say that the C-test was correlated with the teacher rated proficiency levels. However, if we know more information about the construct of the proficiency levels, we could better present this correlational evidence for the construct validity. For now, as the elements of the construct of the teacher rated proficiency levels were not clearly known to us, I would like to be cautious in using this correlational study for evidence for the construct validity of the C-test.

Another way of examining the construct validity of the C-test was the empirical item analysis. Non-native speakers' response errors were observed for this empirical item analysis.
at the lexical level. There were two types of response errors: semantically motivated errors and syntactically motivated errors. This error analysis suggests that test takers who made mistakes in writing their answers may need more time to develop either their syntactic knowledge or vocabulary knowledge or both. This evidence, therefore, revealed the characteristics of response errors which might be caused by lack of written language ability. Thus, this empirical item analysis provided evidence for the construct validity.

Evidence from experimental research was examined to provide a final piece of evidence for the construct validity. In order to verify that observed test performance behaves in concert with theory-base predictions, native English speaking students' performances on the C-test were compared with non-native English speaking students' performances. There were two predictions concerning native English speaking students' performances on the C-test. First, it was predicted that native speakers would outperform non-native speakers. Second, native speakers may not be perfect on the test as their language ability may be considered still at the developmental level. The experimental study was conducted by giving the same C-test to native speakers at the same time when the test was administered to non-native speakers in a same high school. The test results confirmed the above two predictions. That is, native speakers clearly outperformed non-native speakers: the C-test mean scores were 86.24 (86.2%) and 52.73 (52.7%) with the spelling error tolerance scoring, for native and non-native speakers, respectively. Also, native speakers were not perfect on the test: that is, the mean score was 86.24 (86.2%) with the spelling-error tolerance scoring method; this suggested that their written language ability may be still at the developmental stage.
Therefore, the above four types of studies provided evidence positively for the construct validity.

The third question was asking to what extent the C-test is similar to authentic tasks in the target language use domains. The logical evaluation provided evidence for this question by examining whether the characteristics of TLU tasks were fully described and, then, examining whether the characteristics test tasks on the C-test correspond to those of TLU tasks. The analysis results suggested that, in general, the authenticity of the C-test was considered relatively medium. In the C-test, we do not see lots of correspondences between the TLU tasks and the C-test task itself, or task of limited production. However, test takers likely encounter situations in the classroom where ability of written language and metacognitive strategies are necessary as they are required in the C-test task to some degree. Also, as the texts were real materials from text books and supplementary text books which may likely be used in the classroom, test takers may encounter or may have learned similar topics and academic written language as presented in the test; this may raise the degree of authenticity to some extent.

The fourth question asked to what extent the C-test task is interactive. The logical evaluation of interactiveness concerned the degree of involvement of the test takers' topical knowledge and language knowledge, and involvement of language functions in the test task. The C-test task may require limited use of language knowledge—mainly grammatical knowledge; however, the test task may require varied topical knowledge, some language functions including sociolinguistic and instructional, and the strategies to interact with
components of language knowledge to some degree. Therefore, interactivity was considered medium.

The fifth question was asking whether impact on individuals and educational systems is positive. First, impact on test takers and teachers was examined for the logical evaluation. For test takers, the test taking may urge them positively to study more by letting them know their weakness of the target language use on the test. After LEP students were selected based on the test results as part of the whole assessment package, they may receive ELS instruction, which is considered beneficial to them. A possible negative impact might be a case that the test taker at the beginning level may feel that the C-test is too difficult. That is, the test taking could be too frustrating for the test taker as he or she may have difficulties in using his/her written language ability along with topical knowledge, and metacognitive strategies. Teachers may value using the C-test as a test of written language ability as part of the whole package for selection and placement decisions for ESL programs as inclusion of a test of written language ability is an effort to balance the whole assessment. Inclusion of a test of written language ability, the C-test, may also help educational systems (e.g., schools, local AEAs, etc.) provide ESL programs or place LEP students in appropriate ESL programs based on the balanced assessment package. The logical analysis suggested that impact could be positive.

The empirical study also provided evidence for positive impact on test takers, teachers, and educational systems. I examined information gathered in the process of operationalization and administration of the C-test. Test takers, both native and non-native speakers, accepted the C-test as a legitimate test measuring their language ability; some of them asked me to send them their scores on the C-test. Teachers showed positive attitude toward the real use of the
C-test for ESL instruction; no negative attitude was observed at the administration stage.

Two schools volunteered to participate in the C-test study. A local AEA was very supportive at the operationalization and administration stages. The director of Educational Services in the local AEA was supportive in doing the C-test research. The ESOL consultant volunteered in contacting schools and teachers in the process of administration and in preparing test booklets by utilizing resources in the AEA. Therefore, based on the empirical evaluation, the impact on individuals and educational systems were considered positive for now.

The final question was whether the C-test is practical to use. First, based on the logical evaluation, the C-test was considered practical to use by utilizing existing resources, either human resources or material resources. Then, the practicality was examined in the process of operationalization and administration of the C-test. It was possible to use existing human and material resources. At the stage of operationalization, human and material resources in a local AEA were used to prepare letters to parents and students and the C-test booklets by providing clerical help and materials for typing and copying. Therefore, local AAEs and schools could utilize existing human and material resources at the operationalization for the future use of the C-test. At the administration stage, in addition to me, two teachers and the ESOL consultant participated in administering the pilot and main tests. Teachers and ESOL consultants could be utilized as human resources at the stage of administration for the future use of the C-test.

In this study, the usefulness of the C-test as a test of written language ability was examined by answering the six research questions concerning six qualities of test usefulness: the reliability, construct validity, authenticity, interactiveness, impact, and practicality. In
summary, the C-test was considered highly reliable in this testing situation. Empirical study provided evidence for the construct validity of the C-test based on the definition of the construct to be measured. Authenticity of the C-test was considered medium and so was interactiveness. Impact on test takers, teachers, and educational systems was considered fairly positive. The C-test was practical to use in the process of collecting data involving the operationalization and administration stages in this study.

Then, is the C-test useful in measuring written language ability for selection and placement of LEP students for ESL instruction in high school in rural Iowa? We have to answer this question by judging the overall usefulness of the C-test in this particular setting based on the analyses concerning individual qualities of test usefulness (Bachman and Palmer, 1996). We have to consider keeping a balance among individual qualities of test usefulness. That is, even though authenticity and interactiveness of the C-test were considered relatively medium, the C-test was highly reliable in measuring the construct defined in this study and the empirical study provided evidence for the construct validity of the C-test. Reliability and construct validity are the two major crucial qualities of test usefulness as a basis in making inferences based on performances on the test. These two qualities basically ensure the quality of the test itself. In addition, the C-test was also considered to create positive impact and practical to use; impact and practicality are the two important qualities closely related to the future use of the test. If a test generates negative impact on the individuals and the educational system and the test is not practical to use, the test users may not want to use the test. Therefore, based on the results of this research study, I would say that the C-test is useful in this particular testing situation for now.
Recommendations for the real use of the C-test

There are several points that I would like to mention before the C-test version of this study is used in real world. First, it is not recommended to give this version of the C-test to the very beginning level ESL students; this may only make those students frustrated in learning a new language and may give teachers negative impressions at those very beginning level ESL students if their scores on the C-test are extremely low. Therefore, it may be necessary for the test administrator to know if the test taker is able to read and write in English at least to some degree before the test administration.

Second, the C-test in this study was developed for selection and placement of LEP students; that was the original intention of the C-test development (Klein-Braley and Raatz, 1984; Klein-Braley, 1997). Therefore, the C-test usefulness was examined for that purpose. Actually, the empirical analysis in this study supported that assumption. That is, as reported in Chapter 4, the distribution of the C-test was quite similar to the normal distribution; the mean scores of non-native speakers were in 50% range across all three scoring methods. Therefore, it is not recommended to use the C-test as an exit test for now even though the possibility of using the C-test as an exit test was considered in the process of the test development. We need to include different aspects in examining the C-test usefulness focused on the possibility for an exit test; this requires further research.

Finally, if the C-test may be used to select and place LEP students in high school, we need to make sure of at least two things: setting up guidelines for selection and placement, and developing a blue print for creating parallel forms. First, we need to suggest guidelines or scales for selection and placement; this needs to be considered seriously as the decision based
on the suggested guidelines or scales may affect the test taker's future academic success in school possibly to a great extent. In this study, with the spelling-error scoring method, all native speakers scored 60 and above on the C-test whereas 43.3% of non-native speakers scored 60 and above. (See Figure 5.1 for the comparison of the frequencies of the C-test scores from both native and non-native speakers with the spelling-error tolerance scoring.) It seems surprising that more than 40% of non-native speakers scored 60 and above, and 26.5% of them 70 and above. However, if we consider the teacher rated proficiency levels of non-native speakers, the score range seems reasonable as 56.7% of them were considered in the advanced level and in the mainstream level. Therefore, the scores between 60 and 80 seem to be a possible borderline for selection. In order to narrow down the score range, we need to do further research with a larger population. For guidelines or scales for placement, we also

Figure 5.1. The comparison of the frequencies of the C-test scores from both native and non-native speakers with the spelling-error tolerance scoring method (N=17 for the native speaker group; N=30 for the non-native speaker group).
need to have more data from a larger population of non-native speakers and, then, investigate evidence from correlational studies. For now, if teachers or decision-makers may want to use the C-test in its current stage, teachers or decision-makers need to be cautious in using the scores of test takers as a basis for their decisions. Second, we need to consider creating a blue print to make parallel forms of the C-test in order for security. If only a version of the C-test is used, the test items or the topics may be spread among test takers later on; then, this might affect the scores of some test takers to some extent and, therefore, fail to create meaningful scores.

**Implications for further research**

This research study about the C-test usefulness was a limited study in the following two ways. First, very limited number of the target population participated in this study: 5 non-native speakers for the pilot test and 30 for the main test. Therefore, it was difficult to draw definite conclusions about setting guidelines for selection and placement. Secondly, there was lack of active input from ESL or mainstream teachers especially at the development stage. Therefore, if it were possible to do a similar study again, I would like to improve these two areas. For further research for this particular setting, it may be possible to study the C-test usefulness for an exit test; for that, more data from a larger population should be obtained.

Finally, I would like to briefly comment on the use of Bachman and Palmer’s framework (Bachman and Palmer, 1996) which was the basis of this study. First, it was useful to use the framework, in which a test developer or designer needs to consider all six
categories, or six individual qualities of test usefulness, because the framework helped me consider different aspects affecting the test quality and the test development itself. Without a guiding framework, a test developer such as myself could choose a couple of categories or qualities of test usefulness that are possibly easy to justify in his/her test development. Instead, considering all six categories or qualities was useful in keeping a balance among those qualities and, therefore, in possibly preventing a test developer's decision from his/her subjective perspective. Regarding the evaluation of reliability and construct validity, requiring the logical evaluation was useful at the test developmental stage. Even though the empirical evaluation of reliability and construct validity is considered important, providing the logical evaluation seemed to enhance the quality of test development from the beginning. Even though all areas of questions preparing the logical evaluation (see Appendix B) could not be relevant to all testing situations, most questions were quite relevant to this particular testing situation. Even though it was time-consuming to consider all those questions at first, the process helped me build a basis for evaluating the overall usefulness of the test.

Secondly, providing evidence for each category or quality was not very difficult once I understood the concept of the category or quality. However, studying the framework for a half of a semester seemed just enough to see the glimpse of those concepts when the framework was introduced in a testing course. It was necessary to spend much more time in understanding the framework to apply to this particular testing situation. By using the framework, considering various aspects of evidence for each category was possible. For example, at first, I thought that authenticity of the C-test should be quite low by simply looking at the type of the test task as it was difficult to match the C-test task to a type of the
target language use tasks. However, using the framework helped me see the fact that the
topics and language in the C-test were from real materials relevant to the target language use
domain, and therefore authenticity of the C-test was re-evaluated based on more detailed
facts, or evidence. As a conclusion, I could say that my overall reaction in using the
framework is quite positive. Therefore, the framework could be useful to other researchers
and test developers in their particular testing situations as well.
APPENDIX A. INSTRUCTIONS FOR TEST PROCTOR/RATER

HIGH SCHOOL ENGLISH WRITTEN PROFICIENCY TEST

The C-test is a holistic language test similar to a cloze test. It is useful as a measure of a non-native English speaking student’s level of written English proficiency. This version may be used with students in grades 9-12.

Give the test to a native English speaking volunteer at least once before testing a non-native English speaking student. Increasing your familiarity in this way will produce a more consistent, reliable response.

Administer the test in a small classroom, office, or other quiet area that will be free from interruptions for a 50 minute period. The room should be well lit, with appropriate temperature control. If the test is given to a group of students, be sure to schedule the test in a room with adequate space to ensure meaningful and useful scores. Have a supply of sharpened pencils available in case there is a need. You will need a stopwatch or timing device.

Several oral instructions are important to tell the student so that he or she understands the purpose of the test. To ensure consistency and fairness for all test-takers, instructions must be given in English. Tell the student this is a test measuring his/her written language ability. Also, tell the student all parts of the test carry equal weight in the final score.

Read the instructions and sample to the student (also in English). Begin timing when the student turns to Section A, or the test passages. Tell the student to stop writing when thirty minutes have passed.

When you have collected the test(s), thank the student(s) for their efforts. Tell the student(s) when they will receive their score and explain that this is only one part of the assessment package.

Score the test in a private area using the answer key provided. The total weight for each blank is one point. No point is given for an incorrect response or no response. The total sum of the whole test is the score of the student. When scoring, exact words from the original text(s) are considered correct answers. Grammatically and semantically appropriate alternatives are also considered acceptable as correct answers; if you find an answer that appears correct but is not on the key, have your school’s ESL coordinator read the test or call the ESL consultant at your local AEA. Spelling errors are counted as incorrect. There is a sample test and key:

It w______ a very co________, windy afte________. The stud______ didn’t fe______
much li______ going anyw______.

<table>
<thead>
<tr>
<th>Text 1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1</td>
<td>was\textsuperscript{1}</td>
<td>cool/ld\textsuperscript{2}</td>
<td>afternoon</td>
<td>students</td>
<td>feel</td>
</tr>
<tr>
<td>Line 2</td>
<td>like</td>
<td>anywhere</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{1}Letters underlined are answers.
\textsuperscript{2}The first word ‘cool’ is an answer; ‘cold’ is also considered an acceptable answer.
# APPENDIX B: QUESTIONS AND ANSWERS
FOR LOGICAL EVALUATION OF TEST USEFULNESS

<table>
<thead>
<tr>
<th>Questions for logical evaluation of usefulness</th>
<th>Extent to which quality is satisfied</th>
<th>Explanation of how quality is satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RELIABILITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. To what extent do characteristics of the test setting vary in an unmotivated way from one administration of the test to another?</td>
<td>Quality satisfied to some extent.</td>
<td>Characteristics of the test setting vary to some extent with regards to the time of the day or physical setting in each school. It is recommended that test takers be tested early in the day or in a place without interruptions for optimal performance.</td>
</tr>
<tr>
<td>2. To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality completely satisfied.</td>
<td>Characteristics of the test rubric do not vary in an unmotivated way from one part of the test to another.</td>
</tr>
<tr>
<td>3. To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, from one task to another, and on different forms of the test?</td>
<td>Quality completely satisfied.</td>
<td>Characteristics of the test input do not vary in an unmotivated way from one part of the test to another.</td>
</tr>
<tr>
<td>4. To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality satisfied to a great extent.</td>
<td>Characteristics of the expected response do not vary in an unmotivated way from one part of the test to another by including various topics; however, the test taker may have an advantage if a topic is relatively familiar to him/her or he/she may feel confusing in filling in parts of words as this procedure rarely happens in real life.</td>
</tr>
<tr>
<td>5. To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality completely satisfied.</td>
<td>Characteristics of the relationship between input and response do not vary in an unmotivated way from one part of the test to another.</td>
</tr>
<tr>
<td><strong>CONSTRUCT VALIDITY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarity and appropriateness of the construct definition, and the appropriateness of the task characteristics with respect to the construct definition.</td>
<td>Quality satisfied to a great extent.</td>
<td>The language ability construct is clearly and unambiguously defined based on</td>
</tr>
</tbody>
</table>
unambiguously defined?
7. Is the language ability construct for the test relevant to the purpose of the test?
8. To what extent does the test task reflect the construct definition?
9. To what extent do the scoring procedures reflect the construct definition?
10. Will the scores obtained from the test help us to make the desired interpretations about test takers’ language ability?

Possible sources of bias in the task characteristics.
11. What characteristics of the test setting are likely to cause different test takers to perform differently?
12. What characteristics of the test rubric are likely to cause different test takers to perform differently?
13. What characteristics of the input response are likely to cause different test takers to perform differently?
14. What characteristics of the expected response are likely to cause different test takers to perform differently?
15. What characteristics of the relationship between input and response are likely to cause different test takers to perform differently?

AUTHENTICITY
16. To what extent does the description of tasks in the TLU domain include information about the setting, input, expected response, and relationship between input and response?
17. To what extent do the characteristics of the test task...
correspond to those of TLU tasks?

Test task. Utilizing ability of written language for the test task itself may be similar to many TLU tasks in the classroom even though the degree of using the ability for the test task is quite limited compared the ability required for many TLU tasks. Passage on the test may contain topics and written language relevant to TLU tasks as the texts are real materials from text books likely used in the classroom. Therefore, in general, the correspondence could be considered medium.

INTERACTIVENESS

Involvement of the test takers' topical knowledge.

18. To what extent does the task presuppose the appropriate area or level of topical knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?

Moderate to low. Test task presupposes test takers' language ability and topical knowledge appropriate for their general educational level. The range of language ability and topical knowledge are likely widely varied as test takers may have various educational background.

19. To what extent are the personal characteristics of the test takers included in the design statement?

Included in considerable detail. Personal characteristics are included in the design statement in detail. Test takers will have a wide variety of personal background.

20. To what extent are the characteristics of the test tasks suitable for test takers with the specified personal characteristics?

Generally suitable except very beginning level test takers. Test tasks are considered suitable for test takers in the high beginning through the advanced level but are unlikely suitable for the very beginners.

Involvement of the test takers' language knowledge.

21. Does the processing required in the test task involve a very narrow range or a wide range of areas of language knowledge?

A medium range of areas of language knowledge. Test task is asking specific questions: to figure out the second half of a word based on some of grammatical and topical knowledge; on the other hand, test task also requires a bit of comprehension of the text to get the second half of the word in many cases. Therefore, in general, the processing involves a medium range of areas of language knowledge.

Involvement of language functions in the test tasks.

22. What language functions, other than the four major language functions, are involved?

Sociolinguistic Filling in missing part of a word is more
than the simple demonstration of language ability, are involved in processing the input and formulating a response?

function; instructional function.

23. To what extent are the test tasks are interdependent? Highly interdependent.

24. How much opportunity for strategy involvement is provided? To some extent.

25. Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best? Considerable.

26. To what extent might the experience of taking the test of the feedback received affect characteristics of test takers that pertain to language use? Considerable. Low face validity may affect motivation and test takers’ self image; however, taking the test may urge test takers positively to study more by letting them know their weakness of the target language use on the test.

27. What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers, in the design and development of the test? Considerable. Test takers should be interviewed during try-out and allowed to provided feedback, which would be used to refine the test in a real test development.

28. How relevant, complete, and meaningful is the feedback that is provided to test takers? (Are the scores that are reported meaningful? Is qualitative feedback provided?) Considerable. The feedback may help test takers know their language ability and help them be placed in ESL program if necessary.

29. Are decision procedures and Yes. All test takers follow the same procedures

IMPACT

I. Impact on individuals.

Impact on test takers.

26. To what extent might the experience of taking the test of the feedback received affect characteristics of test takers that pertain to language use? Considerable. Low face validity may affect motivation and test takers’ self image; however, taking the test may urge test takers positively to study more by letting them know their weakness of the target language use on the test.

27. What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers, in the design and development of the test? Considerable. Test takers should be interviewed during try-out and allowed to provided feedback, which would be used to refine the test in a real test development.

28. How relevant, complete, and meaningful is the feedback that is provided to test takers? (Are the scores that are reported meaningful? Is qualitative feedback provided?) Considerable. The feedback may help test takers know their language ability and help them be placed in ESL program if necessary.

29. Are decision procedures and Yes. All test takers follow the same procedures
criteria applied uniformly to all groups of test takers?  
30. How relevant and appropriate are the test scores to the decisions to be made?  
31. Are test takers fully informed about the procedures and criteria that will be used in making decisions?  
32. Are these procedures and criteria actually followed in making the decisions?  

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?</td>
<td>Yes</td>
<td>The educational system values competence in written language proficiency which is one of necessary requirements in mainstream classes.</td>
</tr>
<tr>
<td>To what extent do the values and goals of the test developer coincide or conflict with those of society and the education system?</td>
<td>Essentially complete agreement.</td>
<td>The values of the test develop coincide with those of society and the educational system to provide equal educational opportunity.</td>
</tr>
<tr>
<td>What are the potential consequences, both positive and negative, for society and the education system, of using the test in</td>
<td></td>
<td>As potential positive consequences, the educational system can help LEP students and teachers at the same time by providing appropriate ESL program or starting ESL</td>
</tr>
</tbody>
</table>

Impact on teachers.

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>How consistent are the areas of language ability to be measured with those that are included in teaching materials?</td>
<td>Somewhat consistent.</td>
<td>The language ability to be measured is the sum of several areas of language ability included in teaching materials.</td>
</tr>
<tr>
<td>How consistent are the characteristics of the test and test tasks with the characteristics of teaching and learning activities?</td>
<td>Less consistent.</td>
<td>There are no obvious consistencies as test tasks are not selected from TLU tasks.</td>
</tr>
<tr>
<td>How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?</td>
<td>Highly consistent.</td>
<td>The test is developed to select LEP students for ESL help to function well in mainstream classes and to start ESL programs, which will be highly valued of the teacher and of the instructional program.</td>
</tr>
</tbody>
</table>

Impact on society and education system.

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are the scores designed to be directly relevant and appropriate to the decisions to be made, combined with other assessment of the whole package.</td>
<td>Highly appropriate.</td>
<td>Test taker are informed about test taking procedures and about general aspect of language ability. They are not explicitly informed about the procedures of scoring. The above is consistent with the intent of the test developer to provide enough information to enable the test takers to do their best but not excessive information for which that test takers have no direct use. Generally no options are expected for making decisions. But, some options are expected based on each school’s situation.</td>
</tr>
<tr>
<td>Are test takers informed about test taking procedures and about general aspect of language ability. They are not explicitly informed about the procedures of scoring.</td>
<td>Reasonable informed.</td>
<td></td>
</tr>
<tr>
<td>The scores are designed to be directly relevant and appropriate to the decisions to be made, combined with other assessment of the whole package.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test taker are informed about test taking procedures and about general aspect of language ability. They are not explicitly informed about the procedures of scoring.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The above is consistent with the intent of the test developer to provide enough information to enable the test takers to do their best but not excessive information for which that test takers have no direct use. Generally no options are expected for making decisions. But, some options are expected based on each school’s situation.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
39. What is the most desirable positive consequence, or the best thing that could happen, as a result of using the test in this particular way. And how likely is this to happen?

40. What is the least desirable negative consequence, or the worst thing that could happen, as a result of using the test in this particular way, and how likely is this to happen?

PRACTICALITY

41. What type and relative amounts of resources are required for: (a) the design stage, (b) the operationalization stage, and (c) the administration stage?

42. What resources will be available for carrying out (a), (b), and (c) above?
APPENDIX C. GUIDELINES TO CONSTRUCT A C-TEST

The following guidelines explain how to construct a C-test measuring written language ability of high school ESL students.

1. Define the target population, or high school ESL students, and the test format.

2. Select appropriate texts:
   1) choose a text from written materials which are designed for high school academic domain;
   2) choose a text that may not contain specialized vocabulary and content;
   3) choose a text that may be complete in itself within the length;
   4) choose a text that contains 70-100 words appropriate to make 20 deletions;
   5) collect suitable texts, more than necessary.

3. Make texts into the C-test format:
   1) the first line of each text remains intact;
   2) starting the second word in the second sentence in each text, delete the second half of every second word until the number of blanks, 20 blanks, are obtained;
   3) if the number of letters of a word are odd, delete one more letter than half: that is, if a word is “there,” make it “th______”;
   4) the underline for each deletion needs to be uniformed so that any test taker may not have an advantage from the length of the underline;
   5) exceptions: single-letter words like “a” (the indefinite article), proper nouns, numbers, etc., will not be included in word counting;
   6) after the last deletion, the 20th one, each text continues until it reaches its natural conclusion.

4. Screen the formatted texts:
   1) the formatted texts need to be screened by giving the texts a few native speakers equivalent or similar to the target group in order to find out if there are any problematic deletions;
   2) if some deletions elicit unexpected responses, they are considered tricky or confusing; change and/or add deletion(s) for those responses;
   3) if a text is too difficult for native speakers, it may not be a good text for the C-test; do not include the text in the C-test.

5. Select the best suitable formatted texts, 5 to 6 texts, to be included in the C-test booklet from prepared formatted texts. Make sure the C-test booklet present various topics.

6. Finalize the C-test booklet by including instructions and examples. A practice text can be included; in that case, the practice one needs to be the first part and to be very easy compared to the other parts to enhance the test taker’s familiarity with the C-test task.
APPENDIX D. THE SAMPLE C-TEST MATERIAL

HIGH SCHOOL ENGLISH PROFICIENCY TEST

This test is designed to measure language ability for working with academic learning situations in high school.

Directions: In the test, there are six (6) paragraphs. Each one has twenty (20) blanks. Fill in each blank with the half of the word that is missing. Do not add any extra words. It may help to read the whole paragraph before trying to fill in the blanks. Write your answers clearly in the blanks. You have thirty (30) minutes to complete the entire test. You may return to previous paragraphs of the test if you finish early.

Example:

In the test, you will see blanks in each paragraph such as the following:

The dinosaurs died because the earth began to change. When t_______
dinosaurs li_______ on t_______ earth, t_______ climate w_______
暖 a_______ mild every_______. Then t_______ earth slo_______
started t_______ change t_______ a coo_______ climate. T_______
swamps be_______ to d_______ up.

You have to write your answer in each blank shown here:

The dinosaurs died because the earth began to change. When t_______
dinosaurs li_______ on t_______ earth, t_______ climate w_______
warm and mild everywhere. Then t_______ earth slowly
started to_______ change to_______ a cooler climate. T_______
swamps began_______ to dry_______ up.

Now begin work on the test on the next page.
1. **How do plants form seeds?**

Many plants form seeds. Some plants form seeds as ball-like bodies. Other plants form very tiny seeds. How can you group seeds? You may have wondered where seeds come from. Some plants have cones. Seeds form inside the cones. Other plants have flowers. Seeds form inside the flowers.

2. **First Law of Motion**

When you sit in a chair, many forces act on your body. The pressure of the atmosphere pushes on your body and gravity pulls your body downward. You sit comfortably because your body pushes upward with equal force against the atmospheric pressure. Your chair pushes up against the force of gravity to keep you from falling to the ground. All the forces acting on you are balanced. As you sit, your body obeys Newton’s first law of motion.
3. **Friendship**

Friendship means understanding. You do_____ have t______ pretend wi______ a friend. I______ okay t______ be who you a_______. They’ll under______ when yo_____ sad a______ be ha______ with y______ when y______ feel go_______. True fri______ tell ea______ other sec______ and kn______ that th______ won’t b______ spread a______ over. True friends won’t make fun of you if you get a bad grade on your test.

4. **Money Management**

Whatever money you have, managing it is important. Money manag______ means get______ the most satisfaction fr______ your fina______ resources. I______ means us______ knowledge wis______ to pl______ for yo______ future. Plan______ is t______ key. Sho______ you b______ something o______ not? Sho______ you b______ now o______ wait f______ a sale? Wo______ one expe______ pair of slacks or two pairs at a lower price better meet your needs? Such decisions are part of money management.

[Go to the next page.]
5. *Food*

Food is an essential resource. You can survive without it. Wise use of the resource has a major effect on your life. It can make a healthy body—both physically and mentally. Food can help prevent disease. On the other hand, unwise use of food can even play a part in world food shortages. Learn all about this vital resource.

6. *Industrial Revolution*

In the early stages of the Industrial Revolution, the new machines used water as their source of power. Water wheels, which we turned by running water from creeks and rivers, supplied the power for the machines. By the first part of the nineteenth century, however, steam engines were used to supply energy for machines. Coal, which was used to produce steam, was so plentiful in England and so important to England’s wealth that it was called black gold.
REFERENCES


ACKNOWLEDGMENTS

I would like to express my sincere thanks to the following:

My POS committee:
Carol Chapelle, Dan Douglas, and Fred Duffelmeyer who have been so patient with me and who have guided this project toward a right direction.

The Educational Service Division in the Northern Trails Area Education Agency:
   Troyce Fisher, Director, Colleen Hovinga, ESOL consultant, and others who helped this project become real in between schools/teachers and me.

The schools:
who participated in the pilot and main tests.

The teachers:
who helped to administer the pilot and main tests.

The students:
who volunteered to take the pilot and main tests.

My family:
my husband, Byungjun Ahn, and son, Saejin. Without your patience, support, understanding, and encouragement, I could not complete my study successfully.