8-2016

# An evaluation of a post-entry test: An item analysis using Classical Test Theory (CTT)

Suthathip Thirakunkovit
*Purdue University*

*Graduate School Form*
*30 Updated 12/26/2015*

**PURDUE UNIVERSITY**
**GRADUATE SCHOOL**
**Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Suthathip Thirakunkovit

Entitled

AN EVALUATION OF A POST-ENTRY TEST: AN ITEM ANALYSIS USING CLASSICAL TEST THEORY (CTT)

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

April Ginther
Chair

Anthony Silva

Richard Johnson-Sheehan

Shelley Staples

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): April Ginther

Approved by: Ryan Schneider 4/21/2016

Head of the Departmental Graduate Program Date

# AN EVALUATION OF A POST-ENTRY TEST:

# AN ITEM ANALYSIS USING CLASSICAL TEST THEORY (CTT)

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Suthathip Thirakunkovit

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

For my parents and friends:

I would like to dedicate my dissertation work to my loving parents and my best friends. A special feeling of gratitude is first given to my parents. Even though both of you cannot be here to celebrate my success, I always think of you, and I know that I always have your support. Moreover, I would like to thank my close friends, Ning, Jeab and Ben, who are always there throughout my entire doctorate program. All of you have been my amazing cheerleaders. Thank you for constant support and encouragement during the challenges of my graduate school and life. Lastly, I would like to give a special thank to my best friend, Tyler Carter. I am truly thankful for having an opportunity to know you. You have taught me how to love and be loved.

# ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Associate Professor April Ginther for continuous support of my PhD. study and research. Her insight in language assessment tremendously guided me throughout the process of my dissertation writing. I could not have imagined myself able to finish my dissertation without her support and encouragement.

Secondly, I would like to thank all my committee members Professor Anthony Silva, Professor Richard Johnson-Sheehan, and Assistant Professor Shelley Staples, who were more than generous with their expertise and precious time.

Last but not the least, I would like to thank my colleague, Dr. Lixia Cheng, Testing and Assessment coordinator of the PLaCE Program, who helped me collect the data and provided valuable statistical guidance.

TABLE OF CONTENTS

Page

LIST OF TABLES ..... vii

LIST OF FIGURES ..... viii

ABSTRACT ..... ix

CHAPTER 1. INTRODUCTION ..... 1

1.1 Statement of the Problem ..... 1

1.2 Objectives of the Study ..... 6

1.3 Research Questions ..... 7

1.4 Organization of the Study ..... 7

CHAPTER 2. THEORETICAL FRAMEWORKS ..... 9

2.1 Knowing a Language ..... 10

2.2 Concept of Communicative Competence ..... 12

2.3 Cognitive Academic Language Proficiency (CALP) ..... 17

2.4 Language Automaticity ..... 19

2.5 Using C-Test and Cloze Elide Tasks to Assess Language Automaticity and Academic Language Proficiency ..... 22

CHAPTER 3. LITERATURE REVIEW ..... 26

3.1 Cloze Procedure ..... 26

3.1.1 What is Cloze Testing? ..... 26

3.1.2 What are Cloze Tests Said to Measure? ..... 29

3.2 C-Tests and Cloze-Elide Tests ..... 40

3.2.1 What are C-Tests? ..... 40

3.2.2 What are Cloze-Elide Tests? ..... 41

3.2.3 Research On C-Tests and Cloze Elide Tests ..... 44

Page

CHAPTER 4. RESEARCH METHODOLOGY ..... 50
4.1 Overview of the ACE-In ..... 50
4.2 The Ace-In at the Piloting Stage ..... 52
4.3 C-Test and Cloze-Elide Test Passages ..... 53
4.4 Testing Procedure ..... 56
4.5 Participants of the Pilot Study ..... 57
4.6 Test Analysis ..... 58
4.7 Classical Test Theory ..... 58
4.7.1 Descriptive Statistics of the Test Scores ..... 60
4.7.2 Test Item Difficulty (Test Item Facility) ..... 60
4.7.3 Test Item Discrimination ..... 61
4.7.4 Test Reliability ..... 62
CHAPTER 5. RESULTS ..... 64
5.1 Demographics Data of the Participants ..... 64
5.2 Readability Indices of Text Passages ..... 64
5.3 Analysis of the Data ..... 67
5.3.1 Descriptive Statistics of Test Scores ..... 68
5.3.2 Average Item Difficulty and Item Discrimination Values for Each Test Passage ..... 71
5.3.3 Item Performance of Each Test Item and their Syntactic Property ..... 72
5.3.4 Reliability Analyses ..... 78
CHAPTER 6. DISCUSSION AND CONCLUSIONS ..... 79
6.1 Summary of the Study ..... 79
6.2 Directions for Future Test Development ..... 82
6.3 Limitations of the Study ..... 84
6.4 Implications of the Study ..... 85
LIST OF REFERENCES ..... 86
APPENDICES
Appendix A Distributions of Test Takers across Their L1 and Study Programs .... 98

Page

Appendix B Flesch Kincaid Grade Level Readability Scores of C-test and Cloze-elide Passages ..... 99
Appendix C TextEvaluator Complexity Scores of C-test and Cloze-elide Passages ..... 100
Appendix D Average Means and Standard Deviations of Item Difficulty and Item Discrimination of Each Test Passage ..... 102
Appendix E Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Item in the Pilot Data ..... 104
Appendix F The Items with the Highest Item Difficulty Values for Each Test Form in the Pilot Data ..... 119
Appendix G Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Cloze-Elide Item in Pilot Data ..... 122
VITA ..... 129

# LIST OF TABLES

Table Page

4.1. The Overview of the ACE-In at the Piloting Stage 60 .....53

4.2 Overview of Test Passages on Each Test Form.....54

4.3. The Content of the Test Passages .....55

5.1. Descriptive Statistics of C-Test of International Students .....69

5.2. Descriptive Statistics of Cloze-Elide Test of International Students .....69

5.3. Descriptive Statistics of C-Test Performance of Native Speakers of English .....70

5.4. Descriptive Statistics of Cloze-Elide Performance of Native Speakers of English....70

5.5. One Way ANOVA Repeated Measures for Comparing the C-test Performance of Native Speakers of English and International Students .....71

5.6. One Way ANOVA Repeated Measures for Comparing the Cloze-Elide Performance of Native Speakers of English and International Students .....71

5.7. Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Item in the Pilot Data .....72

5.8. The Items with the Highest Item Difficulty Values for Each Test Form in the Pilot Data.....74

5.9. Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Cloze-Elide Item in Pilot Data .....76

# LIST OF FIGURES

Figure Page

2.1. Models of the Developments of Communicative Competence Components .....14

4.1. The Structure of the ACE-In .....51

# ABSTRACT

Thirakunkovit, Suthathip. Ph.D., Purdue University, August 2016. An Evaluation of a Post-Entry Test: An Item Analysis Using Classical Test Theory (CTT). Major Professor: April Ginther.

This study is an analysis of test reliability of two screening tasks (C-test and cloze-elide) in the Assessment of College English-International test (ACE-In), a post-entry test developed at Purdue University. The study uses Classical Test Theory (CTT) to assess the reliability of these test items. CTT is selected because this theory is the standard comprehensive procedure for developing, evaluating, and scaling test items (DeVellis, 2006). This reliability analysis is important because it is a prerequisite to the test validation process. This study has three major research questions:

1. What is the item characteristics of C-test and cloze elide?
2. What are the average values of item difficulty and item discrimination of C-test and cloze elide items?
3. What are the internal consistency coefficients for and correlation coefficient between the C-tests and cloze elide tests?

The results of the pilot study showed that the average score of C-test is 77.8 (SD = 9.98), and that of cloze-elide test is 36.59 (SD = 14.86). Considering the average values of item difficulty and item discrimination of both tasks, C-test items are generally considered easy (item difficulty > 0.7), while cloze-elide items are of medium difficulty (item

difficulty ≈ 0.6). Even though C-test items have acceptable discrimination i.e., the average biserial correlation indices ($r_{pb}$) are 0.3, cloze-elide items are shown to have much better discrimination values on average i.e., $r_{pb}$ indices are higher than 0.5. The Cronbach's alpha coefficients, a measure of internal consistency, of C-test and cloze-elide are .88 and .96, respectively. The Pearson product-moment correlation analysis revealed that the correlation between the C-test and cloze-elide is high (r = .66), and it is significant with the p-value less than .01. These analyses of test reliability indicated that the test items were measuring the same underlying construct – generally language proficiency.

Even though the key results of the item analyses showed that C-test did not meet the standard of item difficulty and discrimination, it does not necessarily mean that C-test cannot sufficiently serve its intended purpose as a preliminary screening tool. After examining the score distributions of both C-test and cloze-elide scores, the scores of both tasks range widely. With fairly wide standard deviations, there is a potential to combine the scores of these two screening tasks to identify the students who had a uniformly low performance across both tasks.

# CHAPTER 1. INTRODUCTION

This dissertation is an analysis of test reliability of two screening tasks (C-test and cloze-elide tasks) in the Assessment of College English-International test (ACE-In), a post-entry test developed at Purdue University. The ACE-In consists of three modules. However, I decided to focus only on the first two tasks of Module 1, which are C-test (word completion) and cloze-elide (word deletion) as they both are variations of the cloze procedure. The study uses Classical Test Theory (CTT) to assess the reliability of these test items. CTT is selected because this theory is the standard comprehensive procedure for developing, evaluating, and scaling test items (DeVellis, 2006). This reliability analysis is important because it is a prerequisite to the test validation process. If the test is unreliable, there is no need to spend the time investigating whether the test is valid or not. The goal of my study is to provide information that can be used to assess validity of the test. Essentially, it may provide guidance to the test developers related to revising test items for future test administrations.

## 1.1 Statement of the Problem

The increasing number of international students has created several challenges for many universities across the United States, including Purdue University (Haan, 2009). This trend has become a concern because many admitted students encounter difficulties

when they listen to lectures, read textbooks, participate in class discussions, and give presentations (Read, 2015; Wall, Clapham & Alderson, 1994). Even though international students may meet entry-level requirements for English proficiency, they may not have developed adequate academic language proficiency to handle reading, writing, speaking, or listening requirements at the level needed to thrive in university-level content courses. However, this does not mean that they are not intelligent enough to handle university courses, but rather that they need time and specific instruction to develop their English-language skills to the levels that can meet the rigorous demands of college-level work. This is a common scenario that describes the experience of many international students who come to the United States, who may need English language support to be fully prepared to succeed in U.S. universities. Unless universities offer English language support in academic language areas, international students can easily become overwhelmed not only by the difficulties in academic language but also by the differences in the performance expectations of Western educational systems (Read, 2013a; Read, 2015).

The minimum entry requirement for undergraduate admission at Purdue University is 80 for the total TOEFL Internet-Based test score (TOEFL-iBT), but 88 for the College of Engineering, College of Science, and Krannert School of Management. Minimum sub-section scores are 20. These threshold entry-level scores, however, only represent minimum language proficiency and literacy. The students who enter the university at or near the minimum level are likely to encounter language challenges in their studies because of limitations in their language proficiency (Light, Xu & Mossop, 1987; Read, 2015). Many professors in different departments at Purdue University,

especially those in degree programs that have large numbers of international students, already complain that they have too many students despite having met the TOEFL entry requirements struggling with the language demands of their courses. This is the reason why many professors at Purdue have criticized that standardized tests simply "don't work" i.e., they cannot fulfill their purpose as adequately measuring language proficiency of the students (Ginther, 2013). Some professors even believe that some students are cheating on the tests (Ginther, 2013). In most cases, they mistakenly assume that meeting the cut-score requirement means that students' proficiency is relatively high and that they should be able to succeed without support. However, students at the cut-score are more likely to need extra language support to succeed in their studies. Some might argue that these international students come with strong study skills. However, many of them lack writing and speaking skills when they first arrive in the United States. Their limited language proficiency or lack of these productive language skills may hinder their integration or contribution to the class.

Even though some departments or schools may choose to raise their cut-scores, Purdue University's minimum language requirement is still set at 80 for the TOEFL-iBT. Since Purdue University offers a large number of academic programs in science and mathematics, it may lose its competitive advantage to its competing institutions if international applicants with outstanding academic backgrounds in these scientific fields defer their decision to apply if they struggle to meet a higher language requirement. Due to this marketing reason, developing a post-entry test can be a possible alternative to raise the existing admission requirements (Read, 2015). Therefore, the post-entry test called the Assessment of College English-International (ACE-In) is being developed to help

identify incoming students who may need extra language support and to help the university create support courses for these students.

Some people might ask whether the scores of standardized tests such as the TOEFL or IELTS tests are sufficient for making decisions about whether students need extra language support while studying their main subjects. Even though many studies indicate that the scores of these heavily-researched tests are reliable, the main purpose of many standardized tests is only to measure general language proficiency for admission, and these test scores are not meant to be specifically used for screening or placing students into different ESL courses – i.e., these tests are not designed to be sensitive enough to detect small proficiency differences (Kokhan, 2012; Mullen, 2009; Read, 2013b). When they are used for different purposes, incorrect placements or negative consequences may result (Mullen, 2009).

Currently, some large universities use the TOEFL or IELTS tests for English course placement. However, the results of some studies are not satisfactory. For instance, two studies conducted at the University of Illinois at Urbana-Champaign showed that using standardized test scores alone did not work well for placing the majority of the international students into its ESL courses (Kokhan, 2012; Kokhan, 2013). Mullen (2009) conducted a similar study at a large university in Canada to examine the impact of using a standardized test as a placement tool. In particular, Mullen (2009) investigated whether the Test of English for International Communication (TOEIC) could correctly place students in appropriate course levels. However, the results of a survey of the classroom instructors showed that 126 out of 551 students were misplaced and the misplacement affected the students' willingness to improve their English and the teachers' decisions to

give a pass or fail grade to students who were misplaced. Similarly, in one of the plenary sessions at the Midwest Association of Language Testers Conference (2015), Associate Professor Christine Tardy mentioned that the University of Arizona previously used the TOEFL-iBT writing score to place new students into different writing courses. However, its survey results showed that approximately 80% of the students were misplaced into the courses that were too difficult or too advanced for them. As a result, the university has decided to develop its own writing test that is specifically tailored to meet the needs of the institution and interests of a diverse student population.

Looking at the findings of these studies, it can be argued that standardized tests may not be appropriate for placement in a local university context. Therefore, Kerstjiens and Nery (2000), Kokhan (2012, 2013), and Read (2015) suggested that a university should have a locally developed and administered placement test and then carry out its own research to determine whether the test could sufficiently meet particular needs of an institution.

The ACE-In, the post-entry test being developed at Purdue University for identifying incoming international students who may benefit from English support, may become mandatory for all international undergraduate students whose TOEFL subtest scores are below cutoff points. The preliminary test tryout started in the fall semester of 2014; it is necessary for test developers to ensure that the test items function effectively. Before the test is fully operational, it is important for the test developers to examine the test reliability as a starting point for creating a valid argument about the test. Therefore, this study will adopt several statistical analyses to examine the reliability of the test as a precursor to developing arguments for reliability and validity of the ACE-In test.

## 1.2 Objectives of the Study

With the large number of incoming international students every academic year at Purdue University, which is the context of the study, there is a need to test students in a minimal amount of time and resources while yielding reliable and valid test results. Therefore, the ACE-In has been developed for identifying any international students who are likely to benefit from language support courses offered by the Purdue Language and Cultural Exchange (PLaCE) program. This group of students should be encouraged or even required by the university to improve their language skills by taking extra language courses parallel with the courses in their departments.

The ACE-In consists of three modules. The first module consists of three sub-sections: C-test (word completion), cloze-elide (word deletion), and elicited imitation (listen and repeat). The second module includes short-answer speaking tasks, and the third module is an essay. This study will particularly analyze the item performance of the C-tests and cloze-elide tests, which are the variations of the standard cloze procedure, in order to investigate test reliability of these two sub-sections. This study will examine the results of item analysis (descriptive statistics, item difficulty, and item discrimination) and examine the reliability of test scores through the calculations of Cronbach's alpha and Pearson's correlation.

The main goals of this analysis are twofold. First, we have to determine whether or not the ACE-In meets minimal expectations for test reliability. In the context of this study, the test should make it possible to discriminate, to the greatest possible extent, students who would benefit from extra language support from students who do not need

additional language support. Second, we want to identify the need for item modifications for future test administrations before the ACE-In is fully operationalized.

## 1.3 Research Questions

To be more specific, the primary purpose of this study is to explore how the overall performance of C-test and cloze elide sub-sections function and see whether the tests can effectively identify students who can potentially benefit from extra language support. To that end, the following research questions are posed:

1. What is the item characteristics of C-test and cloze elide?
2. What are the average values of item difficulty and item discrimination of C-test and cloze elide items?
3. What are the internal consistency coefficients for and correlation coefficient between the C-tests and cloze elide tests?

## 1.4 Organization of the Study

This dissertation has six chapters. The current chapter presents the introduction of the study, statement of the problem, purpose of the study, and research questions. Chapter 2 describes the theoretical frameworks that inform the study. This includes the concept of communicative competence, the distinctions between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP), and the concept of language automaticity. In this chapter, I will particularly discuss the key concepts and applications of these frameworks to help inform the use of C-tests and cloze-elide tests in the academic domain. The third chapter of this dissertation will review the relevant

literature on the cloze procedure and its variations and discuss their advantages and disadvantages with regard to their use. The fourth chapter will consist of the descriptions of the ACE-In, the participants of the pilot test, data collection, and data analyses. Following the methodology, there will be a discussion of the basic concept of the classical test theory (CTT) and its application to the validation of the ACE-In. Chapter 5 will contain the results of item analyses, and the last chapter will be the discussion of the results and directions for future test development.

# CHAPTER 2. THEORETICAL FRAMEWORKS

According to Spolsky (1975, 1977), the history of modern language assessment began during the 1960's, which is described as the beginning of the "pre-scientific period." During this period, most teachers had little or no concern for test reliability. Their tests mostly focused on journals, essays, translations-based skills and oral performance. According to Malone (2008), the instructors and test developers during this period lacked formal training in language assessment. Since there were no systematic guidelines for rating and scoring given to the raters, this resulted in debates about test reliability, especially when the written and oral tests were administered to large groups of examinees. Such exams were, for example, those for candidates to the civil service and for university admission.

Lado (1961) is recognized as having introduced the concepts of psychometric assessment into the field of applied linguistics. The test developers during this first period were primarily concerned with the construction of a test that could demonstrate statistically high reliability; as a result, test items such as multiple-choice and true/false were popular. Based on his argument, Lado believed, "Language is built on sounds, intonation, stress, morphemes, words, and arrangements of words" (1961, p. 25); therefore, the learners' mastery of these elements should be tested separately. The test materials used during this period mostly decontextualized, and the common testing

practices were, for example, words in isolation, phonemic discriminations, spelling, and grammatical analysis.

Around the 1970s, with the emergence of the communicative competence approach, a new shift toward the sociolinguistic period occurred. During this period, language assessments focused on the learners' abilities to use different language skills integratively to create meaningful communication. As Spolsky (1968) mentioned a language test should essentially assess the learners' ability to utilize a wide array of linguistic knowledge for communicative purposes, rather than discrete or isolated skills. He further argued that even though test takers could score very high on discrete-point tests, they might not be able to use the language for any functional purpose. His argument later led to the demands for the development and refinement of integrative assessments, which is the starting point of my discussion in this chapter.

## 2.1 Knowing a Language

The shifts in the concepts of language testing can be clearly marked by two major periods, depending on our understanding of "What does it mean to know and use language?" Indeed, different perspectives on the notion of language proficiency can lead to different approaches in assessing the learners' language proficiency.

From what Spolksy has argued, there is a distinction between knowing about a language and knowing a language. The interpretation of each can be very straightforward. Knowing about a language refers to the students' abilities to remember and recite vocabulary and grammatical rules. In contrast, knowing a language implies the learners' abilities to use the language in communication i.e., it involves the abilities to understand

different linguistic elements and then apply them simultaneously to create an infinite number of meaningful sentences in a communication context. For this reason, Lado's argument about language learning was not well received because learning a language is a complex process that actually subsumes several linguistic components, rather than isolated discrete component alone. Essentially, the learners are not only required to have grammatical competence or the abilities to use the rules, but they have to go beyond those rules by being able to use the language in communication.

According to Spolsky (1968), there are two fundamental elements of language proficiency. The first element is knowledge of grammar that determines the connection between words and meanings. The learners cannot master a foreign language without learning grammar rules because grammar allows the learners to make sense of words of their choices. For this kind of knowledge of the language system, it is what Chomsky refers to as *language competence*. However, it is not sufficient for a learner to only know grammar that makes up the language. The learners need to know how to apply that knowledge in order to get their messages across, in other words, to communicate. The ability of using language to communicate is called *language performance* or *language proficiency*, as discussed promoted by Read (2015). This kind of knowledge actually involves several kinds of underlying language components.

If we accept that knowing a language is more than a matter of knowing discrete elements of what language contains, we are now concerned with developing a language test that can assess the learners' abilities to use different elements integratively in communication. In searching for a way to test the learners' abilities in using a second language effectively, the notion of communicative competence needs to be reviewed. In

discussing the concept of communicative competence, I will review the major components of communicative competence models.

## 2.2 Concept of Communicative Competence

One of the frameworks that has been widely accepted as a basis for testing language proficiency in a second or foreign language is Dell Hymes' communicative competence model. Dell Hymes' communicative competence model is grounded in a multi-dimensional viewpoint of what is believed to be competent language learners. Based on Hymes' arguments, communicative competence not only refers to the implicit knowledge of language structures in the Chomskian sense but also includes competence of language use appropriate to a given social situation. Hymes' model of communicative competence was later refined by many applied linguists e.g., Canale and Swain (1980), Savignon (1983), Bachman (1990), and Bachman and Palmer (1996) (See Figure 2-1).

The seminal work of Canale and Swain (1980) classified communicative competence into three major components, which are linguistic competence (knowledge of language structures such as grammatical rules, spelling, and vocabulary), sociolinguistic competence (the use of appropriate language), and strategic competence (the ability to use verbal and nonverbal language that enhances the communication or compensates for communication deficiencies). In fact, it was Canale and Swain who first decided to extend the notion of grammatical competence to linguistic competence in order to avoid ambiguity because they believe that this language component should also include the knowledge of phonology in addition to morphology and syntax.

Another model of communicative competence was proposed by Savignon in 1983 as an elaboration of Canale and Swain's model. Savignon expanded the communicative competence model by adding the discourse competence into the model. Similar to sociolinguistic competence, discourse competence is concerned with the interpretation of utterances in relation to a specific context. However, discourse competence focuses more on the abilities to use and interpret a series of sentences to form a meaning. In her explanations of the model, Savignon equates *language competence* with *language proficiency*; however, she believes that the term *proficiency* is more appropriate because the term *language proficiency* has a connotation that can essentially reflect the dynamics of human communication.

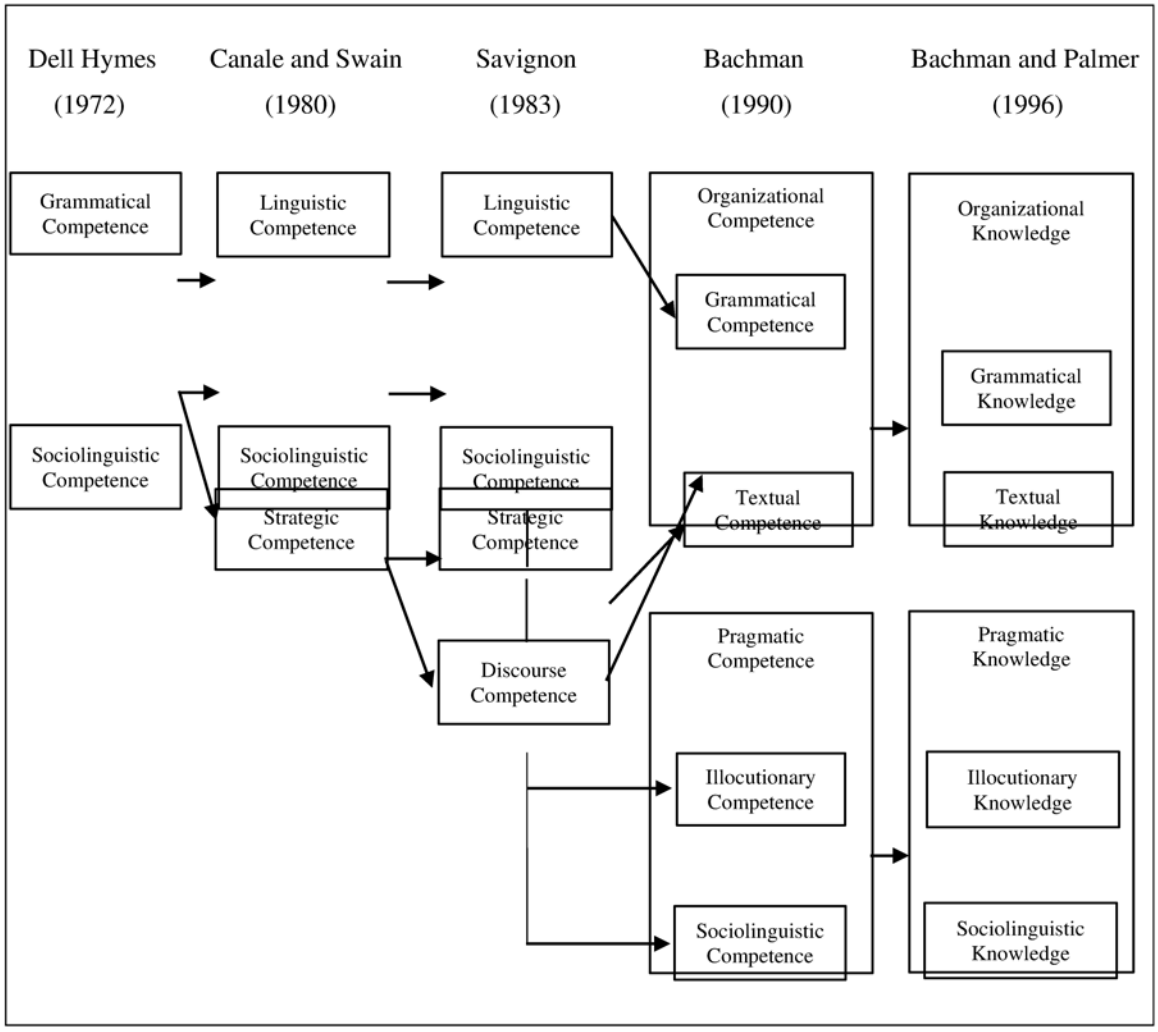


Figure 2.1. Models of the Developments of Communicative Competence Components

In 1996, Bachman and Palmer revisited and slightly altered the model. According to Bachman and Palmer (1996), the crucial characteristics of communicative competence still consist of two broad areas: organizational and pragmatic competence. However, what is new in this model is the use the term *knowledge* instead of *competence*. Both Bachman and Palmer (1996) believe that knowledge about language "can be thought of as a domain of information in memory that is available to the language user for creating and interpreting discourse in language use" (p. 44). Moreover, they added the

descriptions of *knowledge of genres* or the knowledge of the language conventions that shape communication for particular purposes into sociolinguistic knowledge.

In examining different models of communicative competence, more similarities can be found than differences. Corresponding to what Spolsky (1968) has discussed earlier, communicative competence broadly refers to two major components: 1) linguistic or grammatical competence, which is the knowledge of phonology, morphology, syntax, lexicon, and sentence-level meanings and 2) pragmatic or sociolinguistic competence, which is the knowledge of sociocultural rules of how to use the language appropriately. Canale and Swain (1979) stated that even though it was not clear how grammatical competence and sociolinguistic interact, grammatical accuracy must not be ignored in communicative assessments. Clearly, if a learner has a limited command of grammar, it is likely that he/she would not be able to communicate effectively. In a similar way, communication requires the knowledge of social contexts.

An understanding of the content of the definition of communicative competence is necessary because it allows test developers to hypothesize about the underlying construct of language proficiency, and then make claims about what it means to know and use the language. More importantly, these models present conceptualizations of what to test and what is testable. Through the influence of communicative competence models, the ability to use the language in communication have become the major goal of language teaching and language testing.

These models have led to a demand for integrated language assessments that can assess learners' general language proficiency by bringing together all the components of the language competence, including linguistic and pragmatic competence. This

phenomenon has later resulted in the development of direct integrative tests such as oral interviews and integrated writing. Some of these tests could be seen in TOEFL-iBT and IELTS exams. The major aim of this approach is to assess the learners' ability to use the language across interrelated skills, which is what is required in real world contexts (Plakans, 2013). However, one major issue with integrative tests or tasks is test practicality. Those tests can be expensive to administer and require rigorous rater training. Therefore, in responding to this issue, John Oller (1979) proposed the concept of psychologically integrative tests to account for communicative competence. Built upon the concept of communicative competence model, Oller introduced the cloze procedure into the fields of language teaching and testing as a form of indirect measure of integrated abilities (Bloomer, 1962; Carroll, 1961; Schneyer, 1965).

In discussing the construct of the cloze procedure, several scholars such as Cziko (1978), Klein-Braley (1997) and Oller (1979) argue that a cloze test can be an appropriate substitute for a test of productive skills because it can also measure the same underlying skills as those tested in direct communicative tests. Even though cloze testing does not focus on the actual performance of language use in communicative situations, it puts emphasis on the psycholinguistic processing of the language, and successful performance on the task requires the examinees to integrate grammatical, lexical, contextual, strategic, and pragmatic knowledge to supply the blanks with appropriate words. Given the rationale for the cloze procedure, many researchers have argued that cloze testing measures not only grammatical competence because it requires the examinees to understand the context and the language beyond the sentence level in order to choose the correct words (Brown, 1983; Chapelle & Abraham, 1990; Cziko, 1978; Oller, 1973;

Oller, 1979; Oller & Conrad, 1971), but also broader skills such as world knowledge in order to successfully complete the task. In other words, the test takers are required to demonstrate not only vocabulary and grammatical knowledge in choosing the right words and the right forms, but also knowledge beyond sentence level that is necessary for global comprehension e.g., knowledge of collocation i.e., knowing what words tend to occur together and knowledge in pragmatics in choosing words that are appropriate to the content and context of the passage (Pickering, 1976). Based on this view, it can be argued that the performance on the cloze tests depends on the abilities of the learners to integrate grammatical, lexical, contextual, and discourse-level knowledge.

## 2.3 Cognitive Academic Language Proficiency (CALP)

Although the concept of communicative competence is considered a viable model for learning and teaching, in an academic context, this concept is relatively too broad and might not sufficiently address specific needs of students coming to study at universities where English is the medium of instruction. Some international students may find their academic experience to be extremely frustrating when first coming to the United States. They might find the daily workload too heavy and overwhelming. The language used in the academic domain can be complex. If the students cannot adjust or acclimate themselves to the demands of the context, they will find their study stressful. This oftentimes results in frustration or even self-doubt in their abilities to succeed in college. This can later have a negative impact on their success in the U.S. educational environment. Wan, Chapman, and Biggs (1992) state that the success of students largely depends on their levels of specific academic language skills such as note taking, reading

articles, completing assignments in a timely manner, and participating in class discussion. If the students do not have the language skills that are necessary for handling large academic load and complex language use, they will have a tremendous amount of difficulty in their academic lives.

Even though one might argue that most of these international students already received English training in their home countries, the purpose of such language instruction may be to merely enable the students to pass standardized language tests such as TOEFL or IELTS, which are required for admission to universities. This type of intensive language training may not sufficiently prepare students to meet the college-level demands of their study programs even though the students have satisfactory test scores to be admitted to a university. In short, this group of students needs to have more than just general communicative competence.

In the discussion of the particular language demands in academic contexts, Cummins (1999, 2008) defined this particular kind of language competence as Cognitive Academic Language Proficiency (CALP). Even though the term CALP was originally used in discussions of bilingual education in elementary schools, the concept can be applied to the higher education context as well. Strictly speaking, English language learners enrolled in U.S universities must also learn how to process and use the newly acquired language appropriately in order to be considered competent users in academic contexts. Based on the early work of Cummins (1979), in which he demonstrated his ideas about the process of second language learning, CALP is not just about the acquisition of basic grammatical or vocabulary knowledge. Rather, it refers to the cognitive language abilities to comprehend reading materials and lectures in various

fields of study. However, what should be the minimum proficiency required for international students to cope with the academic demands placed upon them?

In responding to that question, Cummins (1979) believes that the students should at least know the meanings of content area vocabulary and be able to able comprehend, analyze, synthesize, and predict the information in a sophisticated manner. Many tasks in the academic domain are cognitively demanding because new information, concepts and language are usually presented to the students simultaneously (e.g., listening to classroom instructions, taking notes, and discussing an issue within a group). Therefore, in order to be successful ESL college students, they have to be able to perform the complex language tasks with accuracy and with little or no effort. According to Gatbonton and Segalowitz (2005), students with high cognitive academic language proficiency should be able to process the language and then perform complex tasks quickly enough without spending too much time thinking about sub-components of the language. If a student needs to have a great deal of conscious attention to get things right, he/she will tremendously struggle in his/her academic work. As argued by both Cummins (1999) and Gatbonton and Segalowitz (2005), these students would need not only time but also specific language instruction to help them achieve academic language proficiency in a level that is relatively comparable to that of their native-speaker peers.

## 2.4 Language Automaticity

Lado (1961) argues that automaticity is very important in language learning because it allows the language users to select the form to create meanings effectively. Considering the acquisition of automaticity, Anderson (1992) has proposed a model that

that can be applied to many aspects of second language learning. Anderson sees the acquisition of automaticity as the development of problem-solving process -- he divides this process into three stages: the cognitive stage, the associative stage, and the autonomous stage.

During the first stage, the learners may only learn a set of skills through memorizing a set of linguistic rules relevant to the skills. The development of the skills during this stage can be slow because the learners typically have to memorize and rehearse those rules as they try to perform the skills for the first time. In the associative stage, the learners start to make connections among various elements required for successful performance which can be strengthened through trial and error. In other words, the learners will convert the knowledge that they already know and use it in a new context. Errors can be detected at this stage and can be eliminated as well. When it comes to the autonomous stage, the whole procedure of language learning becomes more rapid i.e., more automated. One indication of automaticity is that the learners are less aware of the linguistic rules when performing activities that they have already automatized, i.e., they can produce the language subconsciously, without thinking.

The role of automaticity in the process of language learning in general can be very similar to that in academic contexts. In acquiring academic language, the students have to initially learn grammar structures, key vocabulary, typical academic expressions, etc. However, simply knowing underlying linguistic rules of the academic language is not enough for the students to succeed in U.S. colleges because, in real mainstream classroom instruction, the students have to be able to understand academic materials quickly, with automaticity, and absorb teachers' instruction effortlessly. In other words,

the students are required to be able to process the language quickly and automatically, so that they can handle the demands of daily schoolwork. Once the whole process becomes automatic, the students will have a better control over language to simultaneously understand and process the content of their study.

However, some students who are already admitted to U.S. universities might not yet have acquired language automaticity; some students might find reading academic texts and listening to lectures exhausting and completing all assignments on time overwhelming. In the worst-case scenario, some of them may never truly understand the content presented to them. For these students, academic language requires too much of their attention and concentration, thus making listening, reading, and thinking overwhelming and inefficient. Their difficulties might not only show up in their reading, writing, and listening but also in their speech, where the demands of language processing are much higher in association with real time performance.

Lado (1961) and Ellis (2005) strongly argue that language automaticity can be a result of effective, repeated practice. Therefore, in order to help students develop more complex language skills in the way of becoming more automated, requiring less processing effort, additional language instruction is beneficial. Giving them extra language support, the students are given opportunities to practice the usage of the language. With sufficient language practice, they can produce the language without thinking about underlying rules while doing other tasks. Ultimately, language automaticity will be acquired.

## 2.5 Using C-Test and Cloze Elide Tasks to Assess Language Automaticity and Academic Language Proficiency

Based on what has been argued earlier, there is a strong sense that language assessment should primarily focus on integrative skills. Simply put, it may be seen inappropriate to assess knowledge of particular words and grammatical structures in isolation. Indeed, learners are required to have a certain level of language competence in order to be able to perform any integrative tasks. However, in the screening phase of a post-entry assessment, in which the test has to be administered to a large number of students and the test results are required to be processed and returned to students quickly, the use of communicative tests that ask the examinees to truly engage in interactive conversations might not be a viable option. Considering the major goal of the test development, the screening phrase of the ACE-In should make it possible to differentiate the test takers who would are adequately prepared for college study and those who are not.

Considering the context of the post-entry assessment, there could be several major implications following the use of cloze test variations (C-test and cloze-elide tasks) to measure academic language proficiency and language automaticity. First, cloze testing, in its essence, is an integrative assessment. It is important to note that the term *integrative assessment* mentioned here is not identical with integrated tasks on TOEFL or IELTS exams because cloze testing does not require the test takers to use two dominant skills e.g., listening and reading to draw together information from different sources and then synthesize it. Instead, integrative assessment that is being discussed here is built around John Oller's argument about the cognitive demands of integrated language abilities. As

argued by many scholars, cloze testing requires the learners to go beyond merely grammatical competence. The task demands the learners to comprehend and interpret the text by utilizing different kinds of language knowledge e.g., linguistic, textual, and sometimes world knowledge. The learners are required to use all these skills together in order to be able to subconsciously fill in the blanks. These language competences are believed to constitute communicative behavior (Canale & Swain, 1979).

Second, as argued by Oller (1979) and Spolsky (1968), cloze testing is a test of redundancy i.e., measuring the learners' abilities to understand a distorted message. The abilities to understand the language, make valid guesses and inferences that are appropriate to a specific language context is required for academic success in college because once the students enter their own majors, most academic classroom lectures and textbook reading assignments tend to be decontextualized. In order words, the explanations, assumptions, and meaning relations might not be overtly presented in the texts (Biber & Gray, 2010; Briere, 1972). Since they may contain few language cues, the ability to understand overall content and fill in the missing information is important. Oller (1979) has further elaborated this concept by using the term 'expectancy grammar.' For competent leaners, they should be able to make use of these redundancies to activate their internalized knowledge analogous to, or even identical with, what they already know in order to produce the language that is appropriate to a specific context. This process involves both linguistic-based knowledge and world knowledge. As Oller argued, when a competent learner reads, he/she should plan ahead by constantly guessing what should come next based on grammar-based expectancies.

Next, considering the cognitive demands of the test which require the test takers to build overall text comprehension with a reasonable speed in order to successfully complete the tasks under time pressure, cloze testing can be a reflection of how fast the test takers can process the text and how much control they have over the use of language. As noted by Oller (1979) and Grabe (2010), the full knowledge of sentence-level grammar, automaticity of word recognition, and a quick recall of English word collocations are prerequisites for successful performance on this kind of task. The demands of the task, indeed, can be much greater if the test measures the language used in the academic domain. As can be seen, the language use in this context can be very complex because the language used in instructions, textbooks, and exams can contain multiple levels of structural embedding such as finite and non-finite dependent clauses, or phrasal and clausal modifiers (Biber & Gray, 2010). Even though the knowledge of grammar is important, it might not be sufficient. The test takers are required to attain high levels of language proficiency before they can process complex texts more automatically.

Last but not least, considering the practicality of cloze testing, it has been found to be an efficient tool of assessing language proficiency in general. According to Dörnyei and Katona (1993) and Klein-Braley (1997), for example, the tests can be used to test a large group of students in one test administration, the test scoring process is relatively easy and quick, and the test development is less time consuming when compared with multiple-choice format.

Noting these major implications, the test developers hope that C-test and cloze-elide tasks would effectively identify the students who might benefit from support in order to meet the demands of academic contexts. Proficient language learners should not

take a long time to think consciously about what they should fill out in each blank. They should be able to recall basic word meanings and then rely on acquired syntax to accurately assemble inter and intra sentence-level units.

# CHAPTER 3. LITERATURE REVIEW

Determining an appropriate, but simple method in screening students in terms of language proficiency is not an easy task. Many kinds of tests have been proposed as useful and reliable measures for screening or placing students into appropriate courses. Among those that have been successfully employed are the cloze test and its variations (Bachman, 1985; Elder & Von Randow, 2008; Oller & Conrad, 1971; Oller, 1973; Read & Von Randow, 2013).

## 3.1 Cloze Procedure

### 3.1.1 What is Cloze Testing?

Of the numerous integrative assessments, the cloze test is the best-known form of testing. The use of cloze procedure was developed by Wilson Taylor who initially developed the cloze procedure as a measure of "readability" of written materials for native speakers. The term 'cloze,' which is pronounced like the verb 'close' is derived from the word 'closure' in Gestalt Psychology, which emphasizes learners' ability to maintain meaningful perceptions of unrelated or missing elements in the world (Taylor, 1953). The central principle of this concept is that the world is viewed as the whole rather than the sum of its small elements. The concept of the cloze procedure is related to Gestalt Psychology because the cloze task asks the learners to understand the context of a

text as a whole by using both linguistic and semantic knowledge, as well as world knowledge, to accurately fill out parts of the text that are missing. In order to complete the task successfully, the learners must have sufficient language proficiency because they will be required to synthesize the information that is given and then make an inference about different textual elements that are mapped by the linguistic sequences in the passage when filling in the missing segments of the texts (Klein-Braley, 1985, 1998, 1997; Oller, 1979; Spolsky, 1985).

In fact, the ability to understand an incomplete message and make an educated guess is part of the concept of reduced redundancy of language, discussed by Spolsky (1969). According to Spolsky (1969, 1985), the phenomenon of reduced redundancy occurs in everyday language use. For example, in a noisy environment, a person might be required to guess the missing information by relying on available contexts. Or when we hear someone talking on the phone, we might need to guess what the other side of the conversation is about (Spolsky, 1985). The tests of reduced redundancy such as the cloze test can be thought as the simulations of reality because the testing procedure also presents examinees with mutilated texts, and then requires them to restore the incomplete text, based on partial information. The excerpt below is an example of a cloze test passage. This cloze passage follows the standard deletion, which is usually every 7$^{th}$ word. The first sentence of the passage is left intact.

> Living abroad is a unique experience that comes with a lot of responsibility and independence. Along with the stress of traveling _____ starting college in the US, many _____ with culture shock. Culture shock is _____ that is tough to see coming, _____ the good news is there are _____ to help you prepare.

> Living abroad is a unique experience that comes with a lot of responsibility and independence. Along with the stress of traveling and starting college in the US, many struggle with culture shock. Culture shock is something that is tough to see coming, but the good news is there are ways to help you prepare.

According to Klein-Braley (1983), the cloze tests are the tests that have been most researched during the past 50 years, first as a measure of readability, then as a measure of general language proficiency (Alderson, 1978; Ajideh, 2009; Oller, 1972; Oller & Conrad, 1971; Grotjahn & Schiller, 2014), and more recently as a measure of reading comprehension (Williams, Ari & Santamaria, 2011; Trembley, 2011; Yamashita, 2003) because the cloze tests are shown to have moderate to strong correlations with reading comprehension tests. However, the use of cloze test measures seemed to go out of favor when the communicative approach gained favor in the 1970's and early 1980's (Wood, 1993), and a number of researchers questioned the face validity of the cloze test because it, at first glance, appears inauthentic i.e., "incommunicative" (Bradshaw, 1990; Klein-Braley, 1985).

However, face validity is arguably the weakest form of test validity. Indeed, it is very important for a test to appear to test what it reports to measure because stakeholders tend to be more supportive of a test that has face validity. What is more important is construct validity. The cloze test might not match what students do in the real world, but that does not mean that the test is not valid. As many researchers have proposed (Bachman, 1985; Chavez-Oller, Weaver, & Oller, 1977; Gamarra & Jonz, 1987, Ginther, 1986; Yamashita, 2003), the construct underlying the cloze tests is general language

proficiency. Under time constraints, the cloze task requires a great deal of language processing.

Even though the cloze procedure has been mainly criticized though the arguments about face validity, it is still widely used and researched in the field of language testing (Bailey, 1998). The survival of cloze tests in language testing may be explained by two major reasons. First, the cloze procedure is argued to be a reliable tool to assess general language proficiency and reading skills of the learners (Bachman, 1982, 1985; Bormuth, 1969; Brown, 1988; Oller, 1972, 1973; Trembley, 2011), and second, cloze tests are relatively easy to construct and administer (Bormuth, 1969; Grotjahn, 1987). Even the inventor of the cloze procedure, Wilson Taylor, also mentioned the practicality and reliability of the tests after several pilot studies and experiments that were conducted with native speakers of English (Taylor, 1953). However, is a cloze test suitable for the assessment of academic language proficiency? What does it really measure? To further shed light on the use of cloze tests in this specific context, major literature related to the use of cloze tests in language testing will be first reviewed in this study.

### 3.1.2 What are Cloze Tests Said to Measure?

Given the extensive cloze test literature, argument about the construct measured by cloze tests has not been conclusive. As noted earlier, there have been shifts in the use of cloze tests. They were first used as a measure of readability (Taylor, 1953), as a measure of general language proficiency (Alderson, 1978; Ajideh, 2009; Bachman, 1982, 1985; Bormuth, 1969; Oller, 1972; Oller & Conrad, 1971), and as a measure of reading comprehension (Williams, Ari & Santamaria, 2011; Trembley, 2011; Yamashita, 2003).

When Taylor first introduced the cloze procedure in 1953, he presented it as a way of testing the readability of English prose. In his studies, he hypothesized that cloze tests would rank various test materials in the same order as the readability formula such as Flesch-Kincaid, and his findings showed that the mean scores of his participants corresponded with the Flesch-Kincaid indices.

There was a rapidly growing popularity in the use of cloze testing in the 1970s, and it was John Oller who first argued that the cloze test could be used as a measure of general language proficiency. Oller (1971, 1972, 1973, 1979), strongly argued in many of his publications that cloze tests could assess the "pragmatic expectancy grammar," or general language proficiency, which he considers the major mechanism underlying the learners' language abilities to create cohesion and coherence between sentences in a text. According to Oller (1979), pragmatic expectancy grammar involves both language-based knowledge and knowledge of the world. As argued by Oller (1979) and Babaii and Fatahi-Majd (2014), similar to communicative tests, successful completion of the cloze tasks involves the integrated multi-componential nature of language proficiency i.e., not only vocabulary and syntactic knowledge, but also semantic and pragmatic knowledge.

Consider the following example taken from Bailey (1998). If someone hears a person saying, "The cat ate the ___," but did not hear the last word. The most frequent answer that most people have to supply this blank would be *rat*, *chicken*, or *fish*. Indeed, in order to correctly guess what is missing, one needs to know not only English grammar but also something about cats. Because the phrase begins with the definite article *the*, only a noun can fill in this slot. Moreover, what most people would say first is *rat* because our experience tells us what thing is most likely to be eaten by cats. In short, at

least two kinds of language competence influence our abilities to process this sentence. One is syntactic knowledge and the other one is real world knowledge.

Arguments flavoring the validity of cloze testing for assessing reading comprehension have generally come from the correlation analyses between cloze test scores and reading tests. For example, Bormuth (1967, 1968, 1969) first conducted a study to determine whether cloze tests could measure reading comprehension by comparing the cloze tests to several reading comprehension tests. The results of his studies showed that the scores of the cloze test scores and reading comprehension tests were found to be relatively comparable. Oller and Conrad (1971) conducted a study to investigate the correlations between UCLA ESLPE test and other subtests, and the results also showed that the cloze test has a high correlation with a reading test ($r = .80$). Williams, Ari, and Santamaria (2011) compared the performance of two groups of one hundred post-secondary students on a reading comprehension measure and two types of cloze test (maze and open-ended). Their results suggest showed high correlations between both cloze measures and the reading test ($r = .68$ and $.52$, $p < .00$). Gellert and Elbro (2013) constructed a 10-minute cloze test with deletions that requires an understanding of ideas across the text. The results on the cloze test of 240 adult L2 learners were reported to have a significant relationship with their reading comprehension test, and the test also correlated with a 30-minute question-answering comprehension test ($r = .84$). Based on these results, Gellert and Elbro argued that cloze test could be used to measure reading comprehension.

However, when cloze tests are used as a measure of reading comprehension, researchers have not reached consensus on whether they can prompt text processing at

the inter-sentential level. Some researchers have made strong arguments that cloze test items are primarily measuring students' linguistic knowledge at the local or sentence level, as opposed to inter-sentential higher level (Alderson, 1979; Aborn, Rubenstein & Sterlig, 1959; Shanahan, Kamil & Tobin, 1982).

Alderson (1979) studied four different deletion rates (every 6th, 7th, 10th or 15th word) on different cloze passages by comparing the performance of students from different L1 backgrounds. Since his findings did not provide evidence that changes in deletion rates would affect the students' ability to comprehend the passages i.e., increasing the amount of context on the cloze test had no effect on the ease of the task, he concluded that a cloze test was not a suitable test of higher-order comprehension skills. However, one question can be raised about Alderson's conclusion. It is possible that the cloze passages used in Alderson's study were not sensitive to intersentential ties. Therefore, successful completion of those passages could still be achieved by using linguistic cues found in the intermediate environment. Therefore, using a text in which the order of the sentences plays an important role, might yield a different test result.

Aborn, Rubenstein, and Sterlig (1959) randomly selected more than 1,000 sentences from popular magazines to make cloze passages. These sentences varied in sentence length: 6 words, 11 words, and 25 words. In each sentence, one word was removed, and the word was deleted over four different sentence positions: initial, early medial, late medial, and final. Based on the test scores of 24 undergraduate students, the researchers did not find that increase in context beyond 10 words would affect predictability of the missing words and all positions except the final showed the same levels of predictability.

The study of Shanahan, Kamil, and Tobin (1982) also assessed the intersentential sensitivity of cloze passages as measures of the students' ability to use information beyond sentence boundaries. Six standard cloze passages were selected. The readability levels of all test passages were at the 7-8$^{th}$ grade, 11-12$^{th}$ grade, and college graduate level. The order of three passages was kept original; the sequences of the other three were randomly scrambled. The cloze tests were administered to a group of 125 undergraduate students. Each student was randomly assigned to one of the six cloze passages. However, the results did not appear to support the argument that cloze testing could measure language comprehension in the intersentential level because the mean scores on each cloze passage did not appear to be significantly different from one another.

Even though the results of these studies did not show that the sequence of sentences or an increased amount of context would affect the students' test performance, it would be wrong to simply conclude that cloze tests are not sensitive to the context beyond the intermediate environment. There were some limitations in these studies. First, the researchers did not consider the role of language proficiency as a moderator variable in assessing the participants' test performance. For example, Shanahan, Kamil, and Tobin (1982) recruited only undergraduate students to participate in their study. Therefore, it is possible that the differences in the participants' test performances were obscured when high and low ability groups were combined. Second, the researchers did not consider using different types of reading texts in their studies. Since sequences in some texts might not be of critical importance, it is possible that the passages used in those studies could be as comprehensible in their sequential form as they were in their scrambled form.

Scrambling sentences in the text or the increased amount of context beyond sentences might not have a significant effect on the participants' comprehension.

In contrast to the studies mentioned above, many researchers argue that cloze test is sensitive to several linguistic elements, and it can actually measure test takers' abilities to utilize information across sentence boundaries because they have found differences in the test scores when the passages were given to learners at different levels of proficiency.

Oller (1975) asserted that the cloze procedure could, to some degree, measure higher level processing abilities i.e., the abilities to understand overall text messages, and then retrieve basic syntactic and vocabulary information to create cohesion and coherence between sentences in a text. In his study, Oller scrambled the sentences of a cloze passage, and then gave both sequential and scrambled versions to native and non-native speakers of English. Results of his study showed that both groups experienced more difficulty with the reorganized scrambled cloze passages because they took much longer time to finish the task. Nonetheless, the native speakers of English scored better than non-native speakers on the test. Similar results were well supported by many studies conducted by Bachman (1985), Bailey (1998), Chavez-Oller, Chihara, Weaver, and Oller (1977), Chihara et al. (1977), Eckes and Grotjahn (2004), Gamarra and Jonz (1987), Gellert and Elbro (2013), Jonz (1990), McKenna and Layton (1990), Sasaki (2000), and Yamashita (2003). Unlike the results of the studies that found no differences between different test formats, a factor that was introduced in these studies is language proficiency. When scrambled passages were given to the learners at different proficiency levels, there were significant differences in the test performance among different groups of learners.

Bachman (1985) investigated two cloze tests with three deletion procedures: 1) syntactic, in which deletions occurred in the clausal level; 2) cohesive, in which deletions occurs in interclausal and intersentential levels; and 3) strategic, in which deletions occurs in the long-range patterns of coherence. The tests were administered to three groups of pre-university, university undergraduate, and graduate students during the fall and spring semesters of 1982 and 1983. These three groups varied in their levels of English proficiency. The findings demonstrated that the syntactic deletion type was the easiest for all groups; cohesive and strategic deletions were more difficult for learners in the lower level. Not surprisingly, the high group performed more successfully than the low groups in the tasks that required text-level information.

Chavez-Oller, Chihara, Weaver, and Oller (1977) specifically looked at item difficulty and item discrimination of scrambled and intact passages across four different proficiency levels: beginners, intermediate, advanced, and native speakers of English. Their results showed that the items in the scrambled version were more difficult than those in the intact version. More proficient subjects were able to take advantage of constraints across sentences. The authors offered this finding as an explanation of how successful cloze test completion could depend on intersentential contexts.

Similarly, Gamarra and Jonz (1987) and Ginther (1986) provided evidence that cloze testing is a sensitive measure of long-range textual constraints. They selected two expository reading passages. The organization of the first passage was loosely constructed, but that of the second passage was tightly structured. Before the test was given to a group of undergraduate and graduate students, the researchers reordered both passages, thereby creating four test passages: two passages in their original order and the

other two in scrambled order. The results of their study showed that textual sequence did have an impact on the test performance between two groups of students. The mean scores on the sequential tightly-structured passages were significantly higher than those on the scrambled version. This showed that there was a significant interaction effect between the test format (sequential vs. scrambled) and the test passage (loosely structured vs. tightly structured). The passage whose structure was tightly organized was somewhat more difficult. Besides the interaction effect between the test format and the characteristics of the texts, there was also an interaction between the test format and students' levels of proficiency. The graduate students scored higher on both types of passages.

In a more recent study, Yamashita (2003) used cloze tests to measure text-level processing abilities of two groups of Japanese students: skilled and less skilled readers. Based on the analysis of think-aloud protocol data, the researcher found that the test takers with higher reading skills were more likely to use information at the clausal, sentential, as well as the intersentential levels and even their background knowledge to complete the task. She further added that there were several participants in the higher group who used two or three sources of information in order to only answer an item.

Based on the above studies that investigated the intersentential sensitivity of the cloze procedure, the findings showed that cloze tests may measure the learners' comprehension of long-ranges linguistic ties and there seemed to be an interaction effect between language proficiency and test format when scrambled passages were given to the learners at different proficiency levels. As Bachman (1985), Chavez-Oller, Chihara, Weaver, and Oller (1977), Gamarra and Jonz (1987), Ginther (1986), and Yamashita (2003) reported, when a given text violated the long-range linguistic ties, it made the

texts more difficult to process. However, as Brown (1983) has argued, when it comes to students with low-level proficiency, a cloze test may only measure sentence or clause level processing because the test takers were not shown to be able to handle the language beyond sentences. In this case, Sigott (2004) and Abraham and Chapelle (2011) have argued that the construct of cloze testing may be fluid i.e., it could measure different constructs when it is used to assess the performance of the test takers from different levels of proficiency.

### 3.1.3 Concerns Related to Test Use

After studies of John Oller and his colleagues in the 1970s, the cloze technique started to receive extensive attention from a number of researchers. A number of studies have looked at the usefulness of the cloze procedure as a measure of general language proficiency (Abraham & Chapelle, 1990; Bachman, 1982, 1985; Brown, 1983; Oller, 1971, 1972, 1973; Oller & Conrad, 1971; Porter, 1978), and they showed moderate to strong correlations between scores on cloze tests and scores on other language measures. The studies of Bormuth (1968), Porter (1976), Rye (1979), and Taylor (1953) showed significant correlations between cloze test scores and reading comprehension tests, Katona and Dornyei (1993) and Porter (1976) with listening comprehension, Jonz (1976) with writing ability, Shohamy (1982) with speaking ability, Ajideh (2009), Bellert and Elbro (2013), Jonz (1976), and Katona and Dornyei (1993) with vocabulary, and Jonz (1976), Katona and Dornyei (1993), and Oller and Inal (1971) with grammar tests. According to Sigott (2004), the high correlations with the tests of different language skills are not surprising because, as can be seen, the processes required in the task involve

both receptive and productive language use. Based on strong correlations with other criterion measures, many researchers argue that the cloze test may be measuring a wide array of complex language skills. This phenomenon is what Oller (1971, 1972, 1973) calls "general language proficiency."

Even though many studies have shown the reliability and validity of cloze tests, the inconsistent results of the cloze procedure can still be found. Based on previous literature, there are two major issues related to cloze tests: deletion method and scoring procedure.

**Deletion Method.** There are two common types of deletion method for cloze tests: the fixed-rate deletion and the rational deletion. The fixed-rate deletion deletes every *nth* word (usually every fourth or seventh after the first or second sentences) in the text. By contrast, in the rational cloze test, deletions are purposeful and can be tailored to measure knowledge of specific grammatical points and vocabulary items (Oller & Inal, 1971). Regarding the different deletion methods, the cloze testing literature has shown that the choice of deletion method and deletion rate can make a difference in the results of cloze tests.

To start with, Bachman (1985) examined the impact of deletion method (fixed-ratio vs. rational deletion) on the reliability and difficulty of a college-level reading passage. He made different forms of the cloze test made from the same reading passage. With different deletion rates, he found that fixed-ratio deletion led to a significantly more difficult text than did rational deletion. Similarly, Alderson (1979) examined the effect of deletion rate (i.e., at every sixth, eighth, tenth, and twelfth word) on three different texts and concluded that changing the deletion frequencies could have a significant effect on

the levels of test difficulty and test validity, thereby producing different test results. Specifically, the deletion rate of every fourth word consistently yielded a significantly harder test. Changing the deletion rates from every sixth word to every 12th word also resulted in significant correlation changes with another language measure. Based on similar results from different studies, Brown (1983) suggested that longer texts, with a less frequent deletion rate are more suitable for low proficiency learners.

**Scoring Procedure**. Two scoring methods are commonly used for cloze tests: exact word scoring and semantically acceptable scoring method. The exact word method requires test takers to fill in the blanks with the exact same word as was in the original text. In contrast, the semantically acceptable scoring method gives partial or full credit to answers that are syntactically and semantically appropriate. With respect to scoring method, the results of previous studies showed to be in line with the findings that different scoring methods have a differential impact on the scores on and psychometric qualities (e.g., item facility, item discrimination, and reliability coefficients) of cloze tests (Alderson, 1979; Brown, 1980; Kobayashi, 2002; Porter, 1978).

For example, Brown (1980) investigated the use of two scoring methods on different texts, and his findings were in line with Alderson's study (1979) in the way that different methods yielded different means of item facility, item discrimination, and reliability coefficients, and the acceptable method yielded higher item facility and item discrimination, and internal consistency. He also examined the differences between exact-word and acceptable-word scoring methods in particular. Even though he found that the two methods could discriminate students between two groups quite well, acceptable-word scores can differentiate L2 learners of English better. Similar results

were also observed in a more recent study by Kobayashi (2002), whose study showed that the acceptable-word scoring method led to significantly easier test items when compared with the exact-word method.

Based on the findings of different studies, they show that cloze test can be reliable and strongly correlated with other tests, but there seems to be great variability in the effectiveness of cloze tests, which are associated with the deletion method and scoring procedure. Accordingly, in response to these issues, the variations of close testing procedure such as C-test and cloze-elide were developed (Bailey, 1998).

## 3.2 C-Tests and Cloze-Elide Tests

Growing out of the dissatisfaction with unpredictably inconsistent results caused by different deletion techniques and deletion frequencies used for constructing cloze test, and scoring methods, some scholars have proposed the uses of C-test and cloze-elide as alternatives to standard cloze tests (Anderson, 1979; Klein-Braley, 1983, 1998; Klein-Braley & Raatz, 1984). The rationale behind the C-test remains the same, i.e., the reduced redundancy principle. However, the major differences between C-tests and standard cloze tests lie in the deletion rate and clues that are given to the test takers.

### 3.2.1 What are C-Tests?

The C-test was first proposed by Carroll, Caton, and Wilds in 1959, and further developed by Klein-Braley and Raatz (1981). Even though C-test was proposed as an alternative to the cloze test procedure, it still owes the overall format and convention appearance of the standard cloze (Klein-Braley, 1985).

Unlike standard cloze tests in which the whole words are deleted at standard intervals and are replaced by blank spaces, only the first half of each word in C-test is deleted. This gives the test takers additional clues for the missing words. Typically, if the whole word has an even number of letters, then the exact half of the words is deleted. The words deleted can be grammatical or vocabulary words. The first sentence is normally left intact. Consider the following example taken from the practice test of the ACE-In.

> The Purdue University library system has a large number of books available online. These bo_ _ _ do n_ _ require y_ _ to ch_ _ _ them o_ _ and th _ _ do n_ _ have da_ _ _ for y_ _ to ret_ _ _ them to the lib_ _ _ _. The onl_ _ _ books a_ _ convenient b_ _ they ha_ _ limitations. F_ _ example, i_ is usu_ _ _ _ not poss_ _ _ _ to pr_ _ _ paper cop_ _ _ of t_ _ books o_ down_ _ _ _ them to your personal computer.
>
> The Purdue University library system has a large number of books available online. These books do not require you to check them out, and they do not have dates for you to return them to the library. The online books are convenient but they have limitations. For example, it is usually not possible to print paper copies of the books or download them to your personal computer.

### 3.2.2 What are Cloze-Elide Tests?

Another variation of standard cloze test is known as the cloze-elide technique. This technique was first proposed by Valette (1967) as a measure of reading ability (Manning, 1987). The characteristics of cloze-elide testing are actually different from those of other forms of cloze tests. Instead of requiring the test takers to fill in the blanks, cloze-elide tests require the test takers to delete extraneous words from a passage.

However, as Sigott (2004) argued, the cloze-elide test is still similar to other kinds of reduced redundancy tests in many aspects, as it requires the learners to draw upon the same kind of language cognitive processes underlying general language proficiency.

Originally, cloze-elide was known as a speed reading test because the examinees were asked to read a long text and decide which words should be deleted within a limited time (Davies, 1975; Manning, 1987; Read, 2015). In order to complete the task successfully, test takers are required to have rapid skimming and scanning skills to understand a given text as quickly as possible, and then determine which semantic or grammatical information is erroneous (Manning, 1987). As Elder and Von Randow (2008) argue, the whole procedure of language processing required by the task can be viewed as an important part of the academic ability construct because it can reflect the abilities that the students are expected to perform in the academic domain. As argued by the ACE-In test developers, the test takers are specifically required to demonstrate some or all of the following language skills when completing the cloze-elide task:

- Display an understanding of overall content of the passages (e.g., authors' intention)

- Use context clues to determine unnecessary words

- Display control of grammar and vocabulary in context

The following passage is an example of a cloze-elide test taken from the practice test of the ACE-In.

> Living abroad is a unique experience that comes with a lot of responsibility and same independence. Along with the stress of

traveling and starting college in the US, of many struggle with culture shock. Culture shock more is something that is tough to see coming, but from the good news is there are ways to help you prepare. For example, some students experience culture the shock when they encounter university policies that are new and to them. Other students experience stress when it classroom activities require extensive group activities, or when an home emergency situation such as going to the hospital occurs. Planning more ahead can help you lower your chances of experiencing culture shock.

Living abroad is a unique experience that comes with a lot of responsibility and ~~same~~ independence. Along with the stress of traveling and starting college in the US, ~~of~~ many struggle with culture shock. Culture shock ~~more~~ is something that is tough to see coming, but ~~from~~ the good news is there are ways to help you prepare. For example, some students experience culture ~~the~~ shock when they encounter university policies that are new ~~and~~ to them. Other students experience stress when ~~it~~ classroom activities require extensive group activities, or when an ~~home~~ emergency situation such as going to the hospital occurs. Planning ~~more~~ ahead can help you lower your chances of experiencing culture shock.

The test developers selected the words for insertion from the K1, K2, and AWL words lists. K1 types refer to the number of words included in the list of first 1,000 most frequent words of English. K2 type is the number of words included in the list of second 1,000 most frequent words of English, and AWL type is the first 550 words of English that are frequent in academic texts across disciplines. All three word lists were generated from *Compleat Lexical Tutor v.6.2*, which is freely available on the Internet. Once the

test developers received the complete K1, K2, and AWL words lists, they were required to make sure that the words that would be inserted were of comparable difficulty as the target text. Then the words were randomly inserted into the texts line by line.

### 3.2.3 Research On C-Tests and Cloze Elide Tests

As mentioned, disagreement about the construct of cloze tests and dissatisfaction with contradictory results associated with varieties of cloze tests paved the way for the development of its variations such as C-tests and cloze elide tests. However, the C-test is the most common alternative to standard cloze tests. Many researchers have provided evidence which can support the use of C-test in the field of language testing. The major advantages of C-tests mentioned in the literature are twofold: practical and statistical.

**Practical Advantages**

First, C-tests are very easy to construct. It is quite simple to find a text and prepare it as a test when compared with the time required for the development of other kinds of tasks such as multiple-choice items (Dörnyei & Katona, 1993; Klein-Braley, 1997; Kniffka & Linnemann, 2014). Second, C-tests can have more test items with a shorter text because every other word is half-deleted. In this way, the probability of obtaining a representative sample of different word classes in the text is higher (Klein-Braley, 1997). Moreover, when a test contains more blanks in test passages, it is possible that the test items will require more intersentential context to answer. Third, scoring is exact and is not subject to graders' judgment (Eckes & Rudiger, 2006), and the tests can

be easily and quickly scored by computer. Last but not least, the test materials can be adapted to fit the context of test use (Ginther, 1986; Kniffka & Linnemann, 2014).

**Statistical Advantages**

In the field of language assessment, there is ample research on C-tests, and most of the researchers have argued that C-test can be used as a general proficiency measure (Baghaei, 2014; Chapelle, 1994; Eckes & Grotjahn, 2006; Katona & Dornyei, 1993; Klein-Braley, 1985, 1997; Saeedi et al., 2011). Many studies have also provided evidence for the criterion-related validation of C-tests. For example, Negishi (1987) reported strong correlation coefficients of .80 and .76 between C-tests and the reading subtest of English Language Battery (ELBA) and total ELBA, respectively. Katona and Dornyei (1993) demonstrated that C-test has a significant correlation with total TOEIC score (r = .63), TOEIC listening (r = .55), TOEIC reading (r = .54), and their own vocabulary and grammar test (r = .36). All correlation coefficients are significant at the p-level less than .001. Saeedi et al. (2011) conducted on 90 Iranian English majors. The participants' mean score on the TOEFL-CBT was 71.33 (SD = 14.24) and 71.98 (SD = 12.77) for the low and high groups, respectively. They have found a correlation of .92 between their C-tests and total TOEFL scores and correlations ranging from .72 to .88 with TOEFL sub-scores. The highest relationship between the C-test and TOEFL sub-section scores was found between the C-test and the structure sub-section of the TOEFL (r = .88). The correlation between the C-test and reading section is also high (r = .77). The weakest correlation was found between C-test and the listening section (r = .72).

Apart from the application of C-test in second language research that was particularly conducted to examine its reliability and validity, many studies showed that C-tests have been used in different contexts for different purposes. For example, Klein-Braley (1985) used a German C-test as part of preliminary selection and placement. She reported that the C-test generally functioned adequately in placing students in different courses. Moreover, she asserted that the C-test was generally accepted by many examinees as a legitimate testing procedure of their overall language proficiency.

A study of Kniffka and Linnemann (2014), conducted at the University of Cologne, Germany, also used a C-test to place L2 learners of German into appropriate class levels based on the levels of the Common European Framework of Reference (CEFR). The scores of the C-test were compared with another calibrated placement test called DIALANG[1], a web-based language diagnosis program that allows the learners of 14 European languages to assess their proficiency based on CEFR levels, and the results indicated evidence of medium correlations to all sub-sections of the DIALANG test. The adjusted correlation between C-test and listening is .37, writing .43, reading .43, grammar .59, and vocabulary .64.

A study conducted by Baghaei (2014) investigated to what extent the use of C-test could lend itself to be used as a measure general language proficiency of Persian of Iranian junior and senior high school students. The results indicated that C-test data could conform to the assumption of unidimensionality, and the Cronbach's alpha reliability of the test was shown vey high (r = .95). Moreover, the results from the ANOVA and post-

[1] To learn more about the DIALANG test, follow the link: http://www.lancaster.ac.uk/researchenterprise/dialang/about

hoc analyses suggested that the C-test could be used as a measure of general language proficiency because the mean differences between the two levels of students were shown to be significant.

While C-tests have been extensively researched in the field of language testing, there are far fewer studies focused on cloze-elide. According to Manning (1987), pioneering studies were conducted by Bowen in 1978 and Mullen in 1979. The first empirical study of cloze-elide by Bowen (1978) showed that cloze-elide was strongly correlated with the grammar (r = .77), reading (r = .65), and writing (r = .46) sub-sections of the Michigan Test of English Language Proficiency. Mullen's study in 1979 also reported a very high correlation between the cloze-elide and traditional cloze procedure and moderate correlation with sub-tests of TOEFL: writing (r = .40), grammar (r = .65), listening (r = .40), and reading (r = .40). Another two studies that looked at the use of cloze-elide tasks were conducted by Manning in 1986 and 1987. Even though his first study that were conducted on three groups of students (elementary, middle and high school students) showed a wide range of correlation coefficients, ranging between .33 to .89, Manning argued that cloze-elide could provide strong evidence in the use of cloze-elide tests as a measure of general language proficiency because it had a relatively strong correlations with other language sub-tests. For example, the correlations with graded essays ranged between .34 and .76 among the three groups of students, reading comprehension between .60 and .81, listening between .28 and .62, and speaking between .22 and .65. This study also showed that cloze-elide had moderate to strong correlations with other cloze test formats such as standard and multiple-choice cloze tests. The correlations were between .63 and .84. Moreover, the results of his multiple

regression analyses showed that cloze test score was one of the two best predictors of students' general English proficiency. Accordingly, Manning concluded that the cloze-elide test is a relatively more reliable and efficient measure of English language proficiency, when compared with other commonly used testing procedures such as graded essays.

In 1987, Manning conducted another study on the use of cloze-elide test. This one was a large-scale study focusing on the comparing the cloze-elide test scores with the TOEFL scores of 1,208 ESL students in the United States. The results of his factor analysis showed that the cloze-test task was a good predictor of general language proficiency because the correlation between the cloze-elide test and Factor 1 (TOEFL Reading, Writing, and Vocabulary) is .78 and Factor 2 TOEFL listening is .51.

Unlike the C-test, there are much fewer studies that particularly investigated the use of cloze elide tests in different contexts. Elder and Von Randow (2008) is one of the few studies. In their study, Elder and Von Randow (2008) mentioned that the cloze elide has been used as part of the post-entry test at the University of Auckland and the University of Melbourne. Their findings have shown that the cloze-elide test is an acceptably reliable screening tool with the correlation of .82 with the vocabulary sub-test. The overall scores of the cloze-elide and vocabulary tests can discriminate learners from two different levels of proficiency satisfactorily and can accurately predict the performance levels on a diagnostic test of listening, reading, and writing.

Even though the previous literature has proclaimed that c-test and cloze-elide tests can be an objective and reliable tool in measuring general proficiency and global comprehension (Baghaei, 2014; Chapelle, 1994; Eckes & Grotjahn, 2006; Katona &

Dornyei, 1993; Klein-Braley, 1985, 1997; Saeedi et al., 2011), there are still some criticisms of both tests. Both tests have been viewed as lacking authenticity, as the tasks do not represent the language activities that the students would engage in their everyday academic life. However, as argued earlier, many forms of communicative, authentic assessments can be expensive and require a great demand of human resources in administering and scoring. The time and cost cannot be justified when more reliable cost-effective measures exist. The major goal of the ACE-In is to reliably identify students who would benefit from language support. Therefore, it would make more sense for a testing tool to achieve this desired goal while at the same time avoiding the demand of unnecessary effort and cost. That is, the test can be administered to a large number of students and the test results could be processed and returned to students quickly by the use of computer scoring. Since C-tests and cloze-elide tests are efficient means of assessing the test takers' language proficiency, and they can be used to test many students in the most efficient way while still achieving high reliability, so an argument for the use of c-test and cloze-elide can be made with further investigation of this local context.

# CHAPTER 4. RESEARCH METHODOLOGY

## 4.1 Overview of the ACE-In

The ACE-In has been developed for identifying any international students who may benefit from language support. The ACE-In is modeled after the Diagnostic English Language Needs Assessment (DELNA) developed at the University of Auckland, New Zealand, in 2001 (Elder & Von Randow, 2008; Read, 2015). The DELNA is a post-entry test that consists of two phases: screening and diagnostic. The purpose of the screening phrase is to exempt students who are linguistically competent from further diagnosis, and that of the second phase is to identify specific language needs of the students who are recommended to take language support courses. The screening phase includes a cloze-elide test and a vocabulary test. The diagnostic phase consists of listening, reading, and writing tasks. The initial DELNA battery was first piloted and validated in 2001, and the test has been administered on a regular basis since then to all first-year undergraduate students upon their arrival at the university, regardless of their native language (Read, 2008). In 2011, the DELNA became mandatory for all incoming doctoral students (Read, 2015).

Similar to the DELNA, the ACE-In is also used to identify students who may need to develop academic language skills. Even though the DELNA has been successfully implemented and demonstrated to be reliable and valid for its intended

purposes for more than ten years of test use (Read, 2013), the ACE-In, which is a newly developed test, is used in a different academic institution; therefore, it must be validated within the local context. In other words, the ACE-In still needs its own validation procedure because the test contexts are different.

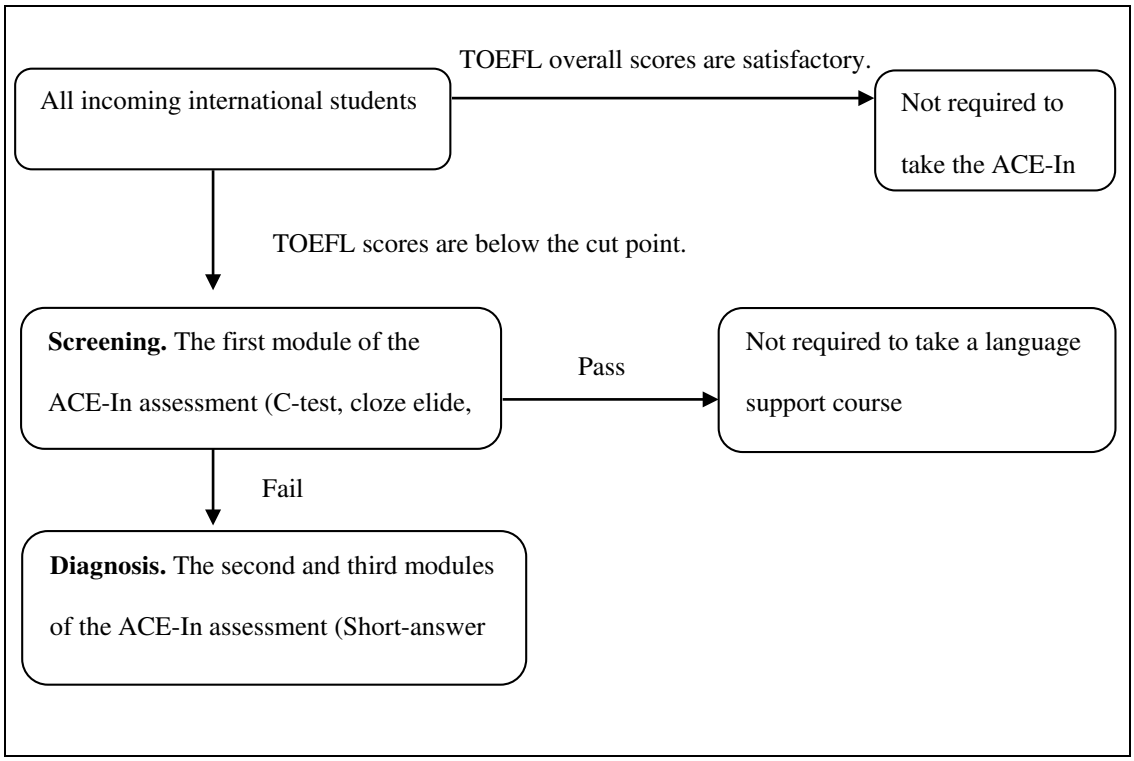


Figure 4.1. The Structure of the ACE-In. This flow chart shows the tentative ACE-In testing procedure. The ACE-In test consists of two major parts: screening and diagnostic. If the students' ACE-In scores are below the cut-off of the screening part, they may be required to take a support course. If they choose to register in an English class, they will be required to take the second and third modules of the ACE-In.

The ACE-In is developed for three main prospective purposes: 1) a post-entry screening test, 2) a diagnostic, and 3) a test to track progress over time, depending on instructional needs and opportunities. The test consists of three modules. The first module may be used as an initial screening, and the second and third modules as diagnostic tests. The scores from different sub-sections will be combined to divide the test takers into two different groups. All modules are administered in a computer lab. The first module is

made up of three tasks: 1) C-test or word completion task (four passages); 2) cloze elide or word deletion task (two passages); and 3) elicited imitation or listen and repeat task (twenty-four items). This screening process takes approximately 50 minutes to administer. In operation, the main purpose of the first module is to offer a quick and efficient means of identifying students who would benefit from extra language support. The second module is the short-answer speaking task (four items) and the third module is the writing essay task (one writing topic). The purpose of the second two modules is to provide additional information about the students whose scores are below the cut-point. This information may be used for a variety of purposes such as setting course objectives or tracking students' progress.

## 4.2 The Ace-In at the Piloting Stage

For the purpose of the pilot testing, all the participants are required to take all three modules. The data were collected during the fall semester of 2014 and spring semester of 2015. The ACE-In test is administered in computer labs on campus in group settings. The test materials are presented on-line on the screen in the same order to all participants: C-test, cloze elide, elicited-imitation, short-answer speaking task, and essay writing. The participants are assigned to one of the test forms randomly. After the test administrators explain the test instructions to the participants, the students work on the computer individually. The examinees are also asked to complete a survey at the end of the test. The entire ACE-In test takes approximately one and a half hours. The general descriptions of the ACE-In are given in Table 4-1.

Table 4.1

*The Overview of the ACE-In at the Piloting Stage*

| Module | Phase | Language skills | Test items | Time allotment |
|---|---|---|---|---|
| 1 | Screening | Reading and speaking | C-test<br>Cloze-elide<br>Elicited imitation | 50 minutes |
| 2 | Diagnosis | Speaking and listening | Read aloud<br>Express your opinion<br>Summarize a conversation | 30 minutes |
| 3 | Diagnosis | Writing | Essay writing | 40 minutes |

## 4.3 C-Test and Cloze-Elide Test Passages

Since this study focuses on the first two sub-sections of the ACE-In, only the detailed descriptions of the structure of these two sections and the procedure of the pilot testing will be described. There are four test forms on the ACE-In. Each form has four C-test and two cloze-elide passages. For the C-test texts, each text consists of four sentences with 25 blanks. Each passage is approximately 60-70 words long. Every other word in a passage is deleted, and the first half of the word is given as a clue. For cloze-elide texts, each text has approximately 300-350 words. Each passage has four paragraphs and 35 lines. Thirty unnecessary words were inserted into each line; the first and last lines of the first paragraph have been left intact, and the last line of the remaining paragraphs also has no insertions. The participants are given five minutes to read each text and complete the

task. The scores are the number of correctly answered items. The summary of the two tasks is given in Table 4-2.

Table 4.2

*Overview of Test Passages on Each Test Form*

| Task | Number of test passages | Length of each passage | Number of the test items in each passage |
|---|---|---|---|
| C-test | 4 | 60 - 70 words | 25 |
| Cloze-elide | 2 | 300 - 350 words | 30 |

The content of the ACE-In test passages is presented in academic contexts, rather than technical academic English per se. As suggested by Klein-Braley and Raatz (1984), the test materials should be selected from the target language domain because the language used in those materials represent the authentic language that the test takers will encounter on a daily basis. All ACE-In test passages were sampled from university materials and textbooks. The topics that were chosen can represent the content demand that all freshmen are going to be exposed to. They are intended to engage the test takers to the university setting. The topics are meaningful and valuable to undergraduate students who are going to start college. Every text is expository, for example, instructions to borrow books from other libraries and explanations of course policies. Considering the texts that were selected for the ACE-In, the examinees need not to have knowledge in any specific field of study. Even though it is true that students who will take the test are from

different disciplinary studies, the materials are generic and do not require discipline-specific knowledge. It would be difficult to determine or specify the test takers' field of study because many new students might not have yet entered their own field of study after first admitted to the university. Table 4-3 below summarizes the content of the test passages.

Table 4.3

*The Content of the Test Passages*

| Test form | Test | Topic |
|---|---|---|
| 1 | C-Test | Purdue library system (sample test passage) |
| | | Riding the bus |
| | | Interlibrary loan |
| | | Parking on-campus |
| | | Syllabus course policy |
| | Cloze-elide | Culture shock (sample test passage) |
| | | Research angle |
| | | Job descriptions of a manager |
| 2 | C-test | Purdue library system (sample test passage) |
| | | Credit hours |
| | | English 106-I |
| | | Group work |
| | | On-campus housing |

Table 4.3 Continued

| | | |
|---|---|---|
| | Cloze-elide | Culture shock (sample test passage) |
| | | E-books |
| | | Culture |
| 3 | C-test | Purdue library system (sample test passage) |
| | | Campus safety |
| | | Going to the gym |
| | | Undergraduate research |
| | | Writing lab |
| | Cloze-elide | Culture shock (sample test passage) |
| | | Plagiarism |
| | | Greek society |
| 4 | C-test | Purdue library system (sample test passage) |
| | | Alumni association |
| | | OEPT |
| | | Student club |
| | | University Hall |
| | Cloze-elide | Culture shock (sample test passage) |
| | | Writing purpose |
| | | Extracurricular activities |

## 4.4 Testing Procedure

For all test takers, the test session begins with sign-in and short orientation. The testing staff from the PLaCE program explains the test procedure and informs the test takers about the post-test survey that every test taker is required to complete. After that, the test takers are allowed to choose any test station that has been set up. They can

proceed each test section at their own pace. The test takers are allowed to take notes on paper provided by the test administrators. Most test takers take approximately two hours to complete both the test and survey. After the test takers finish the test and the survey, a test administrator will collect their notes. The test takers' responses are automatically saved in the database on the web server.

## 4.5 Participants of the Pilot Study

There are three groups of students who participated in the pilot study. Participants in the first group are the students who have enrolled in the GS-100 course during the fall semester of 2014 (General Studies 100: Reading, Writing & Speaking for International Students I), a support course designed to improve students' English language skills. Since there is no language requirement mandated by the university at the moment, all students in the first group are those who were recommended by their academic advisors to take the GS-100 course.

Participants in the second group are those who have overall TOEFL scores and speaking scores higher than those enrolled in the GS-100 course. The ACE-In was administered on a voluntary basis to international students in the higher group. An email was sent out by the director of the PLaCE program to recruit students whose TOEFL-iBT speaking scores are higher than 23. The participants in the first two groups are international undergraduate and graduate students, who study English as a second or foreign language. Regarding the third group of participants, they are native speakers of English. The recorded numbers of participants for each group are 200, 32, and 28, respectively.

In the future when the ACE-In is fully operationalized, the test developers may use the overall TOEFL-iBT total scores together with some sub-section scores as an additional reference when deciding who will be required to take the test.

## 4.6 Test Analysis

After the ACE-In has been developed, two issues are of major concern for the test developers: 1) whether the test can produce the scores that are reliable and 2) whether the test can assist in making decisions about students who could potentially benefit from the extra language program. Specifically, these two questions are directly related to the reliability of the test i.e., whether the test items correlate with one another and whether the test can discriminate between high and low groups of the test takers.

## 4.7 Classical Test Theory

In assessing the reliability of the ACE-In test, Classical Test Theory (CTT) will be used. CTT is chosen because this theory consists of a set of concepts and methods that provide a basis for score reliability, which is important for test development (Crocker & Algina, 1986; DeVellis, 2006; Kline, 2005). In this section, the major assumptions and concepts underlying CTT will be reviewed and discussed.

Classical Test Theory (CTT) is a based on the true score theory. According to Crocker and Algina (1986), CTT assumes that a person's observed test score is comprised of a true score and randomized error. The formula is illustrated below:

$$T \text{ (True score)} + E \text{ (Error)} = X \text{ (Observed score)}$$

According to Crocker and Algina (1986), DeVellis (2006), and Kline (2005), the true score (T) can be defined as the expected score that an individual should receive if he/she has been tested repeatedly over a number of times. It is the true score that reflects the true ability of the person. The error (E) is the difference between the true score and the observed test score (X). Based on this model, a true score can be obtained if there is no error. However, as Kline (2005) states, a measuring instrument is always imperfect. In other words, the result of any measurement or an observed score usually contains error which, for example, might come from the differences in testing environments between two test administrations or in the test takers' psychological or physical state on different days of testing. Therefore, the score achieved by a person on the same test can differ from one test administration to another. Because the true score cannot ever be obtained, it is important for test developers to estimate the variance of the error based on the observed scores of a group of examinees, either from one or more test administrations or from one or more test forms in order to determine how much the test can reliably reflect the true score. According to CTT, standard deviation of the observed scores from individuals is used to estimate the variance of the error, so-called the standard error of measurement. In principle, the smaller the standard error of measurement is, the more reliable the test can be in measuring the test takers' true ability.

The degree to which a test has can reliably measure the true ability can be indicated by the four aspects of item performance:

1. Descriptive statistics of the test scores
2. Test item difficulty (Test item facility)

3. Test item discrimination
4. Test reliability

### 4.7.1 Descriptive Statistics of the Test Scores

Typically, when a dataset of a test is examined, it is common practice for the test developers to first report descriptive statistics of test scores -- the mean and standard deviations are the basic features of descriptive statistics. The means and standard deviations are important because they can determine whether the test is relatively easy or difficult for the target population (Kline, 2005). Since the ACE-In is a norm-referenced test, the test developers would expect to see that the test is of medium difficulty. According to Crocker and Algina (1986), a test of medium difficulty can best maximize the variance among the test takers' scores.

### 4.7.2 Test Item Difficulty (Test Item Facility)

Based on CTT, test item difficulty or item facility is defined as the proportions of students who answered a particular item correctly (Brown, 2003; Crocker & Algina, 1986). Items with high values are considered easy items; items with low values are difficult items. It is commonly suggested that test items that are too easy or too difficult for the target population should either be revised or replaced because these items are not able to discriminate the abilities of examinees well enough (Bachman, 1990). According to Crocker and Algina (1986), test items should be of medium difficulty (between .3 and .7) in order to maximize greatest test variance or the test ability to discriminate the students, which is desirable when developing a norm-referenced test (Kline, 2005). In

other words, the test should be sensitive enough to demonstrate a hierarchy of language abilities among the test takers.

### 4.7.3 Test Item Discrimination

In addition to item difficulty, the point-biserial correlation, which is a common method for investigating item discrimination, will be reported. The point-biserial correlation can provide information on how well a particular item in a given test can separate between test takers who are relatively high and those who are relatively low. According to Brown (1988), the interpretation of this test index is similar to that of the Pearson's Product Moment Correlation, which ranges between -1.00 and +1.00. The higher the value of the index is, the better a particular item is in discriminating students as it indicates a stronger relationship between that item on the test and students' total scores. Even though there is no absolute benchmark for acceptable values of point-biserial correlation, the convention within the language-testing field suggests that a correlation of at least 0.3 is considered acceptable. If the point-biserial correlation index of an item is negative, it means that the students with lower ability received higher scores on that item than the higher ability students do. Therefore, an item with a negative item discrimination index should be discarded from the test. The point-biserial correlation is useful for a screening test because it can help test developers make a decision whether a particular test item on the test can sufficiently discriminate between those who have the necessary language skills and those who do not.

### 4.7.4 Test Reliability

Apart from the analyses of individual test items, this study also examines test reliability by applying the Cronbach's alpha coefficient and the Pearson product-moment correlation. In order to examine internal test consistency, Cronbach's alpha will be calculated. This reliability estimate provides an indication of "how responses to an item relate to the total test score" (Kline, 2005, p. 99). The items should have a high level of consistency because it indicates that the test items are measuring the same underlying construct. For the interpretation of test scores to be reliable, Nunnally (1978) recommends a minimum level of .7. Even though some researchers argue that the calculation of the internal consistency reliability coefficient is inappropriate for C-tests and cloze-elide tests because of the possible dependency of this kind of tests (Bachman, 1990; Faraday, 1983), it is still necessary for the test developers to show the evidence that the test scores can represent the same construct. Moreover, in order to determine to what extent C-test and cloze elide tasks are correlated with each another, the Pearson product-moment correlation analysis will be calculated.

In essence, an examination of test item performance allows the ACE-In test developers to establish reliability of the test via the assumptions about the consistency of the items and test scores (Crocker & Algina, 1986; Kline, 2005). These indices will provide useful information in improving the overall performance of the test items. If any test passages contain a high proportion of faulty items i.e., they do not reliably provide information for estimating the test takers' abilities, the test developers may need to flag those items for further review. The decision to revise or discard certain items should be made with caution, however. Because, in the case of C-test and cloze-elide passages, it is

not always easy to simply delete a certain item without having an impact on the performance of other items in the same passage. Moreover, some items on the test can be less difficult than others in nature i.e., we could expect grammatical items to be easier than lexical items (Jafapur, 1999). In this regard, should we delete those items from the test if they are shown to be too easy for the examinees? And what should we do with poor functioning items? One solution is to examine overall test performance of each test passage and try to select the ones that show overall satisfactory statistical results. As long as the statistical results of a certain passage are shown to be able to discriminate among the participants, the test can be argued to be suitable for test use in this context.

# CHAPTER 5. RESULTS

This study reports on reliability of the C-test and cloze-elide tasks. The responses of 260 test takers were analyzed through the use of Classical Test Theory. The results of 232 international examinees were validated against 28 English native speakers' test performance. This study first examined the variation of readability values across C-test and cloze elide text passages. Then the data were analyzed for four traditional item characteristics: 1) Descriptive statistics, 2) Item facility, 3) Item discrimination, and 4) Cronbach's alpha and Pearson's correlation reliability coefficients.

## 5.1 Demographics Data of the Participants

The majority of the ACE-In test takers are Chinese. They use either Mandarin, Cantonese, or Taiwanese as their native language. The test takers in our sample were mostly distributed across three programs: Science (26%), Exploratory Studies (25%), and Management (16%). Nineteen percent of the test takers did not indicate their majors. See Appendix A.

## 5.2 Readability Indices of Text Passages

This study examined whether readability varied across text passages in order to make sure that the passages in each test form are of comparable levels of text difficulty.

The researcher used two websites to calculate readability indices: online-Utility.org and texteval-pilot.ets.org. These two websites were chosen because they use different textual characters to calculate readability indices.

The website *online-Utility.org* provides the readability index estimate called the Flesch Kincaid Grade level, a commonly-used indication of text difficulty of contemporary academic English. The estimate of Flesch Kincaid Grade level calculates test difficulty using number of words per sentence (sentence length) and number of syllables per word. The result is an index that corresponds with a grade level of education in the United States. According to DuBay (2006), readability level scores between eight and twelve are appropriate for college students and general adults.

TextEvaluator™ Analysis Results which was developed by the Educational Testing Service (ETS) provides an overall text complexity score based on four textual dimensions: 1) syntactic complexity e.g., average sentence length, average number of modifiers per noun phrase, and average number of dependent clauses per sentence, 2) vocabulary difficulty, 3) lexical cohesion across sentences, and 4) prior knowledge required to understand a text. The result of overall text complexity scores is reported on a scale ranging from 100 (appropriate for young readers) to 2000 (appropriate for college graduate students). According to Sheehan (2015), a score between 970 and 1360 corresponds to common core 12$^{th}$ grade level.

Entries in Appendix B show Flesch Kincaid Grade Level readability index of each C-test and cloze-elide test passage. It is suggested that a Flesch Kincaid Grade level score of around 10-12 is the reading level on completion of high school and college students. Overall, the results show that all four forms of the ACE-In have similar Flesch Kincaid

grade level readability levels. C-test and cloze-elide passages were rated approximately between the tenth-grade and twelfth-grade readability levels, except for the third passage of Test Form 4 "*OEPT*" whose readability score is 15.4.

Passage 3, Test form 4: OEPT

> The Oral English Proficiency Program (OEPP) at Purdue University was established in 1987 under the support of the Office of the Provost. The OEPP was created to carry out the university policy which states that all international teaching assistants who do not speak English as their first language must demonstrate sufficient English speaking skills. Otherwise, these students cannot be assigned to duties that involve classroom teaching or direct interaction with undergraduate students.

A central question then is, what feature could be changed to make this text more comparable to the scores of other passages. Since one of the measures that the Flesch Kincaid Grade level measures is based on is sentence length, breaking up the long sentence like "The OEPP was created to carry out the university policy which states that all international teaching assistants who do not speak English as their first language must demonstrate sufficient English speaking skills." into separate sentences can lower the level of reading difficulty.

TextEvaluator is a text analysis tool that evaluates syntactic and vocabulary complexity, cohesion, and prior knowledge required to understand the text. Each of the component scores of the TextEvaluator tool is expressed on a scale that ranges from 1 to 100. Higher scores indicate higher levels of text difficulty.

The results of the analysis (See Appendix C) show that the overall complexity scores of both C-test and cloze-elide passages were approximately between 700 and 1100, which are roughly equivalent to the common core standard levels between Grade 9 and Grade 12 (Sheehan, 2015). However, some passages might be either more complex or easier than other passages. For example, based on the TextEvaluator complexity score, Passage "*Credit Hour*" might be too difficult to read for a person with reading skills of general undergraduate students. The syntactic and vocabulary complexity scores are relatively higher than those of other passages.

> Passage 2, Test form 2: Credit Hour
>
> If you are holding a student visa while studying in the US, there are certain requirements you have to fulfill in order to maintain your immigration status during your stay in the US. In addition to maintaining a valid passport and unexpired immigration documents, you have to enroll in your university as a fulltime student. This requires registering for a minimum of 12 credit hours per semester. This requirement applies to every international student for every fall or spring semester except during the last semester of your degree program.

Some recommendations to lower the overall complexity score include limiting the use of technical terms such as "immigration," using more familiar terms, for example, replacing the word "maintaining" with "having," and using shorter sentences in the text.

## 5.3 Analysis of the Data

The test takers' test responses were analyzed using the Statistical Package for Social Sciences (SPSS) and Excel spreadsheets. Any observations that had missing

responses were excluded from the analysis. After the first run of analysis, the responses of eight test takers were dropped because they were identified as extreme outliers. After the investigation of the nature of these outliers, it has been found that the test takers did not carefully consider their answers when completing the test. In this situation, the researcher believes that it is legitimate to simply drop these observations.

### 5.3.1 Descriptive Statistics of Test Scores

Tables 5-1 and 5-2 report the descriptive statistics of C-test and cloze elide test scores of all 232 international students separated by test form. Considering the score means and standard deviations of both tasks, the score means of C-test were very high, when compared with those of cloze-elide, and the scores of both tasks ranged very widely i.e., standard deviations were large. Given that the standard deviations of both tasks were relatively wide, the test developers may claim that the C-test and cloze-elide tasks could be served as a quick screening procedure to filter out students whose scores were far below the means across both tasks because these students were more likely to benefit from extra language support.

Table 5.1

*Descriptive Statistics of C-Test of International Students*

| | N | Min | Max | Mean | SD | Skewness |
|---|---|---|---|---|---|---|
| Overall | 232 | 51.00 | 98.00 | 77.80 | 9.98 | -0.32 |
| Form 1 | 66 | 51.00 | 97.00 | 76.73 | 10.73 | -0.31 |
| Form 2 | 58 | 53.00 | 97.00 | 79.19 | 10.14 | -0.28 |
| Form 3 | 55 | 56.00 | 94.00 | 77.62 | 9.45 | -0.20 |
| Form 4 | 53 | 51.00 | 98.00 | 77.79 | 9.47 | -0.51 |

Table 5.2

*Descriptive Statistics of Cloze-Elide Test of International Students*

| | N | Min | Max | Mean | SD | Skewness |
|---|---|---|---|---|---|---|
| Overall | 232 | 2.00 | 60.00 | 36.59 | 14.86 | -0.36 |
| Form 1 | 66 | 4.00 | 59.00 | 38.80 | 14.20 | -0.43 |
| Form 2 | 58 | 2.00 | 60.00 | 37.85 | 17.93 | -0.50 |
| Form 3 | 55 | 4.00 | 57.00 | 34.04 | 12.90 | -0.07 |
| Form 4 | 53 | 2.00 | 58.00 | 35.11 | 13.63 | -0.38 |

The test performances of 28 English native speakers are reported in Tables 5-3 and 5-4. A one-way ANOVA with repeated measures was run on the mean scores of both groups to see whether there were significant differences between native speakers of English and international students' test performance on both tasks. Tables 5-5 and 5-6

show the ANOVA results of each test task. The analyses showed that the native speakers of English performed significantly from non-native test takers on both C-test and cloze-elide tasks at p-values less than 0.000. Native speakers of English obtained higher scores than the group of non-native speakers. This can be interpreted that language proficiency is an obvious factor in the test takers' performance.

Table 5.3

*Descriptive Statistics of C-Test Performance of Native Speakers of English*

| | N | Min | Max | Mean | SD | Skewness |
|---|---|---|---|---|---|---|
| Overall | 28 | 89.00 | 99.00 | 95.50 | 2.53 | -0.893 |
| Form 1 | 6 | 92.00 | 98.00 | 94.83 | 2.56 | -0.60 |
| Form 2 | 5 | 90.00 | 99.00 | 95.40 | 3.29 | -1.29 |
| Form 3 | 9 | 89.00 | 97.00 | 94.89 | 2.57 | -1.72 |
| Form 4 | 8 | 93.00 | 99.00 | 96.75 | 1.91 | -1.01 |

Table 5.4

*Descriptive Statistics of Cloze-Elide Performance of Native Speakers of English*

| | N | Min | Max | Mean | SD | Skewness |
|---|---|---|---|---|---|---|
| Overall | 28 | 37.00 | 60.00 | 56.64 | 4.43 | -3.53 |
| Form 1 | 6 | 55.00 | 60.00 | 58.33 | 1.86 | -1.28 |
| Form 2 | 5 | 50.00 | 60.00 | 56.60 | 3.97 | -1.54 |
| Form 3 | 9 | 37.00 | 59.00 | 54.78 | 6.83 | -2.73 |
| Form 4 | 8 | 54.00 | 59.00 | 57.50 | 1.77 | -1.23 |

Table 5.5

*One Way ANOVA Repeated Measures for Comparing the C-test Performance of Native Speakers of English and International Students*

| Source of variation | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 7829.718 | 1 | 7829.718 | 87.100 | .000 |
| Within Groups | 23192.478 | 258 | 89.893 | | |
| Total | 31022.196 | 259 | | | |

*Table 5.6*

*One Way ANOVA Repeated Measures for Comparing the Cloze-Elide Performance of Native Speakers of English and International Students*

| Source of variation | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 10046.222 | 1 | 10046.222 | 50.322 | .000 |
| Within Groups | 51506.528 | 258 | 199.638 | | |
| Total | 61552.750 | 259 | | | |

### 5.3.2 Average Item Difficulty and Item Discrimination Values for Each Test Passage

For research question 1, the average values of item difficulty and item discrimination for all items of C-test and cloze-elide passages were calculated. The overall means of item difficulty across all C-test passages are higher than those of cloze-elide items, meaning that cloze-elide items are more difficult than c-test items. Overall, C-test items are considered easy (ID > 0.7), while cloze-elide items are of medium

difficulty (ID ≈ 0.6). Even though C-test items have acceptable discrimination i.e., point biserial correlation indices are between 0.3 and 0.4, cloze-elide items are shown to have much better discrimination values on average i.e., point biserial correlation indices are higher than 0.5. (See Appendix D)

### 5.3.3 Item Performance of Each Test Item and their Syntactic Property

Table 5-7 shows the examples of item difficulty and point biserial correlation values for each C-test item and cloze elide item on Passage 2 of Test Form 1. Items marked with an asterisk (*) are within the desirable range of item difficulty and item discrimination i.e. item difficulty values are between 0.3 and 0.7 and item discrimination values are higher than 0.3. Please see Appendix E for the entire results of item difficulty and point biserial correlation values for each C-test item and cloze elide item.

Table 5.7

*Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Item in the Pilot Data*

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 2: Riding the bus (The number of acceptable items is 9.) | | | | | |
| 1 | rides | Content | Noun | 0.94 | 0.16 |
| 2 | available | Content | Adj | 0.91 | 0.31 |
| 3* | all | Function | Det | 0.32 | 0.31 |
| 4* | and | Function | Conj | 0.55 | 0.42 |

Table 5.7 Continued

| | | | | | |
|---|---|---|---|---|---|
| 5 | city | Content | Noun | 0.97 | 0.27 |
| 6 | lines | Content | Noun | 0.79 | 0.52 |
| 7* | simply | Content | Adv | 0.39 | 0.40 |
| 8 | too | Function | Adv | 0.97 | 0.09 |
| 9 | their | Function | Det | 0.89 | 0.27 |
| 10 | university | Content | Noun | 0.97 | 0.04 |
| 11 | ID | Content | Noun | 0.89 | 0.24 |
| 12 | ride | Content | Verb | 0.98 | 0.34 |
| 13 | bus | Content | Noun | 0.98 | 0.18 |
| 14* | schedules | Content | Noun | 0.61 | 0.43 |
| 15 | daily | Content | Adj | 0.65 | 0.06 |
| 16 | lines | Content | Noun | 0.80 | 0.34 |
| 17 | available | Content | Adj | 0.86 | 0.55 |
| 18* | at | Function | Prep | 0.50 | 0.36 |
| 19* | office | Content | Noun | 0.55 | 0.44 |
| 20 | the | Function | Article | 0.85 | 0.60 |
| 21* | residence | Content | Noun | 0.45 | 0.46 |
| 22* | and | Function | Conj | 0.62 | 0.53 |
| 23 | the | Function | Article | 0.73 | 0.66 |
| 24 | information | Content | Noun | 0.73 | 0.66 |
| 25* | on | Function | Prep | 0.45 | 0.59 |

As illustrated in Table 5-7 and Appendix E, approximately, only one fourth of the C-test items have desirable item difficulty levels and discrimination values. Therefore, the C-test items were investigated further in order to find out whether there are any variables that can improve these item values. Two aspects of the items were examined:

word functions and word classes i.e., part-of-speech. The entries on Table 5-8 below and Appendix F show the item difficulty and item discrimination values of each item classified according to their word functions and part-of-speech. As illustrated, the first 25 easiest items of each test form are more likely to be function words rather than content words. And the finding of this study is in accordance with findings reported in several studies. According to Brown (1988), Klein-Braley (1981), and Perkins and German (1985), function words in cloze test passages tend to be easier because they can be guessed from a small number of words in a particular closed class. According to the results shown below, function words that seem to have higher values of item difficulty are pronouns and prepositions. Even though some of the easy items are content words, those easy items tend to be vocabulary words that most students encounter on a daily basis e.g., bus, university, project, course, and time, or words that repeatedly appeared throughout the texts e.g., campus, building, and parking.

Table 5.8

*The Items with the Highest Item Difficulty Values for Each Test Form in the Pilot Data*

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| C-Passage3Item2 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage3Item20 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage4Item13 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage5Item5 | to | Function | Prep | 1.00 | 0.00 |

Table 5.8 Continued

| | | | | | |
|---|---|---|---|---|---|
| C-Passage2Item12 | ride | Content | Verb | 0.98 | 0.34 |
| C-Passage2Item13 | bus | Content | Noun | 0.98 | 0.18 |
| C-Passage3Item4 | the | Function | Article | 0.98 | 0.20 |
| C-Passage3Item5 | from | Function | Prep | 0.98 | 0.43 |
| C-Passage3Item6 | library | Content | Noun | 0.98 | 0.15 |
| C-Passage3Item11 | it | Function | Pronoun | 0.98 | -0.02 |
| C-Passage3Item19 | system | Content | Noun | 0.98 | 0.09 |
| C-Passage3Item21 | to | Function | Prep | 0.98 | 0.09 |
| C-Passage5Item14 | you | Function | Pronoun | 0.98 | 0.13 |
| C-Passage5Item22 | you | Function | Pronoun | 0.98 | 0.26 |
| C-Passage2Item5 | city | Content | Noun | 0.97 | 0.27 |
| C-Passage2Item8 | too | Function | Adv | 0.97 | 0.09 |
| C-Passage2Item10 | university | Content | Noun | 0.97 | 0.04 |
| C-Passage3Item12 | actually | Content | Adv | 0.97 | -0.03 |
| C-Passage3Item22 | is | Function | Verb to be | 0.97 | 0.13 |
| C-Passage3Item24 | the | Function | Article | 0.97 | 0.25 |
| C-Passage4Item3 | for | Function | Prep | 0.97 | 0.09 |
| C-Passage5Item11 | in | Function | Prep | 0.97 | 0.19 |
| C-Passage5Item16 | attend | Content | Verb | 0.97 | 0.44 |
| C-Passage5Item21 | time | Content | Noun | 0.97 | 0.28 |
| C-Passage5Item19 | at | Function | Prep | 0.95 | 0.36 |

In contrast to the characteristics of C-test items, the item analyses of cloze elide items show that approximately 70% of the cloze elide items are within the good range of item difficulty (See Table 5-9). Even though some of the cloze-elide items are considered relatively easy i.e., item difficulty values are higher than 0.7, they do have good

discrimination values i.e., point-biserial correlation values are higher than 0.3. For example, even though 12 out of 30 items in Passage 2 Form 1 are considered easy, items # 1, 4, 10, 11, 13, 14, 15, and 17 have excellent item discrimination values. Note that items that have an asterisk (*) are within the desirable range of item difficulty and item discrimination i.e. item difficulty values are between 0.3 and 0.7 and item discrimination values are higher than 0.3. (Also see Appendix G)

Table 5.9

*Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Cloze-Elide Item in Pilot Data*

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 2: Research angle (The number of acceptable items is 15.) | | | | | |
| 1 | is | Function | Verb to be | 0.82 | 0.36 |
| 2 | were | Function | Verb to be | 0.92 | 0.27 |
| 3* | possible | Content | Adj | 0.44 | 0.63 |
| 4 | the | Function | Article | 0.94 | 0.42 |
| 5* | to | Function | Prep | 0.61 | 0.61 |
| 6 | fairly | Content | Adj/Adv | 0.14 | 0.24 |
| 7* | note | Content | Noun | 0.62 | 0.35 |
| 8 | to | Function | Prep | 0.94 | 0.09 |
| 9* | as | Function | Prep | 0.61 | 0.65 |
| 10 | fine | Content | Noun/Adj | 0.76 | 0.58 |
| 11 | such | Function | Det | 0.74 | 0.53 |

Table 5.9 Continued

| | | | | | |
|---|---|---|---|---|---|
| 12* | style | Content | Noun | 0.68 | 0.56 |
| 13 | your | Function | Det | 0.85 | 0.39 |
| 14 | find | Content | Verb | 0.80 | 0.60 |
| 15 | dismissive | Content | Adj | 0.74 | 0.56 |
| 16* | possible | Content | Adj | 0.45 | 0.67 |
| 17 | not | Function | Negation | 0.79 | 0.37 |
| 18* | sound | Content | Noun/Verb | 0.64 | 0.62 |
| 19* | debate | Content | Noun | 0.58 | 0.73 |
| 20 | as | Function | Prep | 0.82 | 0.29 |
| 21 | would | Function | Modal | 0.86 | 0.44 |
| 22* | what | Function | Pronoun | 0.30 | 0.51 |
| 23* | distinguish | Content | Verb | 0.35 | 0.65 |
| 24 | introduction | Content | Noun | 0.29 | 0.66 |
| 25* | their | Function | Det | 0.67 | 0.55 |
| 26* | and | Function | Conj | 0.70 | 0.54 |
| 27* | other | Function | Det | 0.64 | 0.58 |
| 28* | with | Function | Prep | 0.68 | 0.62 |
| 29* | missing | Content | Verb | 0.42 | 0.75 |
| 30 | pursued | Content | Verb | 0.27 | 0.69 |

In summary, all C-test items have similar item characteristics. The items are generally considered relatively easy for the participants in the pilot study, and their item discrimination values generally do not meet desired levels. Three fourths of the items have low discrimination ($r_{pb} < 0.3$) and/or levels of difficulty are outside the range considered acceptable for this form of item analysis ($0.30 < ID > 0.70$). On the contrary, more than 70% of the cloze elide items are within the acceptable range of item difficulty,

i.e. ID values are between 0.3 and 0.7, and more than 90% of them have their discrimination values greater than 0.3.

### 5.3.4 Reliability Analyses

For research question 2, reliability coefficients of both the C-test and cloze elide tests were analyzed by applying the Cronbach's alpha coefficient and the Pearson product-moment correlation. The Cronbach's alpha coefficients of C-test and cloze-elide are .88 and .96, respectively. The reliability coefficients of each form of the C-test range from .86 to .89, and cloze-elide from .94 to .98. As one might have noted, these values Cronbach's alpha coefficients are considered high, which can indicate excellent internal consistency of the items on the test. Moreover, the correlation analysis was conducted between the C-test and cloze-elide test scores. The Pearson product-moment correlation analysis revealed that the correlation between the C-test and cloze-elide is very high ($r = .66$), and it is significant at the .01 level. The results of these two analyses combined enable the ACE-In test developers to claim that both C-test and cloze-elide tasks could sufficiently serve their intended purpose of measuring an underlying construct in a uniform manner.

# CHAPTER 6. DISCUSSION AND CONCLUSIONS

## 6.1 Summary of the Study

The purpose of this study was to examine item characteristics of C-test and cloze-elide tasks i.e., item difficulty and item discrimination and their test characteristics i.e., internal consistency reliability and test correlation, and to evaluate whether the tasks are acceptable for their intended purposes.

The results of the item analyses indicate that C-test items were fairly easy for the test population, thus failing to discriminate between the individuals who had high level of proficiency and those who were in the low level. However, different results were found for the cloze elide task. The majority of cloze-elide items had desirable item difficulty and item discrimination values. Considering the test characteristics, both tasks were found to have very high internal consistency coefficients and correlation coefficients across forms. These analyses indicated that the test items were measuring the same underlying construct -- academic English language proficiency. According to Cummins (1979), the concept of academic language proficiency refers to the ability of students to quickly and efficiently function in an academic context regardless of previous language training. Following this line of reasoning, it is legitimate to claim that both C-test and cloze-elide tests measure the learners' automatic processing of the language, which can, in turn, help identify students who may benefit from English language support.

Even though the key results of the item analyses showed that C-test did not meet the acceptable standard of item difficulty and discrimination, does it necessarily mean that C-test cannot sufficiently serve its intended purpose as a preliminary sorting tool? Indeed, after examining the score distributions of both C-test and cloze-elide scores, the scores of both tasks range widely. With fairly wide standard deviations, the test developers could still use the scores of these two subtests combined to identify the students who had a uniformly low performance across both tasks. If the test developers decide to set the cut-scores at the means of both tasks, there is a probability of detecting 32% (75/232) of the test takers that had the scores of both C-test and cloze-elide test below the means. Considering the TOEFL scores of this group of students, approximately two thirds of them (44/75) had a TOEFL score lower than 90. Based on these considerations, the test developers may be able to identify the students who enter with the lowest levels of English proficiency.

Although the C-test may be deemed acceptable for its intended purpose, identifying the students who were unlikely to be able to cope with the demands of their academic course, the ACE-in test developers may still want to improve the item performance of the C-test by making the test passages more difficult. However, before the directions of test revision are proposed, the potential reasons of poor item performance should be identified.

1. Using dashes indicting the number of letters required for each test item could potentially make the C-test items overly easy, and the test takers may simply focus on guessing the vocabulary instead of comprehending the text and referring to grammar rules.

2. Deleting every other word in the text can produce a test with repeated items. Therefore, it is possible that the test takers simply copied the answers of the items they encountered before. The following test passage displays the problem of repeated items.

> Passage 5, Test form 4:
> Purdue has many buildings on campus that are named for famous alumni and past presidents of the university. One **buil_ _ _ _** that i_ not na_ _ _ after a per_ _ _ is Unive_ _ _ _ _ Hall. Th_ _ building i_ the on_ _ building rema_ _ _ _ _ from t_ _ original **buil_ _ _ _ _** that st_ _ _ on cam_ _ _ when t_ _ university w_ _ established i_ 1869. T_ _ university be_ _ _ construction o_ the **buil_ _ _ _** a f_ _ years af_ _ _ the unive_ _ _ _ _ was fou_ _ _ _. Back th_ _, the building was called the Main Building.

The above example shows three repeated words -- Items 1 (building), 11 (buildings), and 20 (building). Considered the results of the item analyses, these items produced poor item difficulty, and consequently, poor item discrimination. The item difficulty and discrimination values of these three items are ID = 1.00 and $r_{pb}$ = 0.00, ID = 0.92 and $r_{pb}$ = 0.17, and ID = 0.96 and $r_{pb}$ = 0.16, respectively.

3. Several texts contain the answers for some items in intact forms. Therefore, the test takers simply guessed the answer by scanning the texts. For example, in the following passage, the test takers could fill in the last two items of the sentence by reading the previous sentence, which contains the same noun phrase.

> Passage 2, Test form 3:
> The university attempts to provide a safe and secure environment for students, staff and visitors. Unfortunately how_ _ _ _, crime i_ a rea_ _ _ _ on mo_ _

university camp_ _ _ _. To ma_ _ the comm_ _ _ _ _ a sa_ _ place t_ live, a_ _ students sho _ _ keep i_ mind th _ _ **safety a_ _ security** are the respon _ _ _ _ _ _ _ _ of eve_ _ _ _e. It _s possible t_ maintain **saf_ _ _ and secu_ _ _ _** only wh_ _ every stu_ _ _ _ takes a_ active pa__ in the effort.

4. Topics that are very familiar to international undergraduate students e.g., interlibrary loan and syllabus course policy appeared to be very easy passages, when compared with those that are less familiar to them e.g., parking on-campus. As illustrated in Table 5-10, the means of item difficulty and discrimination of these passages are: Interlibrary loan, ID = 0.87, $r_{pb}$ = 0.26; Syllabus course policy, ID = 0.84, $r_{pb}$ = 0.34; Parking on-campus, ID = 0.66, $r_{pb}$ = 0.37.

## 6.2 Directions for Future Test Development

Given that this study revealed some shortcomings in the C-test procedure, it seems reasonable at this point to ask what can be done to improve the test performance of C-test. Can the item analyses be improved? Generally, the C-test items examined display marginal item facility and item discrimination. More than 70% of the test items are considered overly easy. If the ACE-In test developers want to make it more difficult for the test population, several previous studies have given detailed suggestions of how to do it.

1. Babaii and Ansary (2001) and Babaii and Moghaddam (2006) suggested that test developers may examine the characteristics of test passages to increase the level of text complexity e.g., readabiliy levels, vocabulary complexity, and syntactic complexity.

By changing these variables, it may be possible to lower the item difficulty values, which in turn increase the overall discrimination of the test.

2. To adjust the level of item difficulty, Sigott and Koberl (1996) suggested that increasing the number of letters deleted and changing deletion patterns could yield tests that were significantly more difficult than standard C-Test format. In this study, the researchers compared the mean score of standard C-test format with the following deletion patterns:

1) Two thirds of the letters were deleted (curious: cu_____)

2) Only the first letter was given (c_____)

3) The first half of the words was deleted (_____ious)

The results of this study showed that all three versions increased the level of test difficulty. The second version yielded the highest level of test difficulty (Mean = 51.7; SD = 9.0) when compared with the other three versions – Standard C-test (Mean = 81.9; SD = 8.6); Version 1 (Mean = 61.6; SD = 14.7); Version 3 (Mean = 67.0; SD = 10.9).

3. Cleary (1988) similarly proposed that using left-hand deletion rather than standard right-hand deletion or standard C-test could enhance the C-test items performance. In line with the results of this study, his item analyses rarely yielded discrimination values higher than 0.3 and the C-test items were shown to be very easy. However, by comparing the item difficulty means of the two versions of C-test (deletions on the right hand and deletions on the left hand), the left hand version was shown to be more difficult (Left version = 73.4; Right version = 84.8), having a higher discrimination index (Left version = 0.34; Right version = 0.21), and a higher reliability coefficient (Left version = 0.93; Right version = 0.83).

4. To remedy the issue with low item difficulty and item discrimination indices, Kamimoto (1992) suggested that leaving the items whose item difficulty higher than 0.70 and less than 0.2 for item discrimination intact can be a possible solution.

5. Using one long blank (kn_____) instead of dashes (kn _ _) to represent the number of deleted letters can increase the level of item difficulty because the test takers would not be able to automatically guess frequent words or phrases. A study of Babaii and Ansary (2001), for example, illustrated how the participants took a quick look at the phrase "of cou_ _ _," and had an automatic restore of "of course" without reliance on text comprehension.

6. Deleting every second word is likely to produce a C-test with repeated items. Therefore, Babaii and Moghaddam (2006) and Kamimoto (1992) recommended using systematic word deletion or tailored C-test. The test developers can avoid having repeated items by keeping words that appear more than once in the text intact.

## 6.3 Limitations of the Study

There are some limitations of this study. The first limitation is that the test was given to students who were currently enrolled in a program and had already started taking courses in their program. Therefore, item analyses might yield different results if the test is given to incoming students. Another limitation of this study involved the narrow range of the test takers' language proficiency. The majority of the participants had a TOFEL total score ranging between 75 and 103. Even though the PLaCE program tried to recruit higher-level students, only 32 students volunteered to participate in the pilot study and

the highest TOEFL score of this group is 107. The participants of this study were relatively homogeneous, and this may contribute to low discrimination values.

## 6.4 Implications of the Study

Even though determining the validity of the ACE-In is beyond the scope of this study, as part of the ongoing development of the ACE-In, the results of this study may be used to provide information that can be used to assess validity of the test and provide guidance to the test developers in revising test items for future ACE-In test administrations. Essentially, the development of a new post-entry English language assessment at Purdue University has the potential to be an effective means of assessing the college readiness of undergraduate international students. While those students who with scores below the threshold on both tasks can be identified as most in need of support. Finally, revising the C-test section is necessary before these items can be used effectively for identification a placement of students across the entire range of English language skills.

# LIST OF REFERENCES

# LIST OF REFERENCES

Aborn, M., Rubenstein, H., & Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology*, *57*(3), 171-180.

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, *76*(4), 468-479.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 219-227.

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning*, *30*(1), 59-76.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535-556.

Anderson, J. R. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, 165-180.

Babaii, E., & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle?. *System*, *29*(2), 209-219.

Babaii, E., & Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, *34*(4), 586-600.

Babaii, E., & Fatahi-Majd, M. (2014). Failed restorations in the C-test: Types, sources, and implications for C-test processing. In R. Grotjahn. *The C-Test: Current Trends*, 261-273.

Bachman, L. F. (1982). The trait structure of cloze test scores*. *TESOL Quarterly*, *16*(1), 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535-556.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Baghaei, P. (2014). Construction and validation of a C-test in Persian. In R. Grotjahn. *The C-Test: Current Trends*, 299-310.

Baker, B. A. (2011). Use of the Cloze-Elide Task in High-Stakes English Proficiency Testing. *Spaan Fellowship*, 1.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441-465.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2-20.

Blais, J. G., & Laurier, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language testing, 12*(1), 72-98.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives. Vol. 1: Cognitive Domain.* New York: McKay.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah: Lawrence Erlbaum Associates.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *10*(5), 291-299.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, *5*(3), 189-196.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly*, 358-365.

Bornstein, R. F. (2004). Face validity. In *The SAGE Encyclopedia of Social Science Research Methods*. California: Sage.

Bowen, J. D. (1978). The identification of irrelevant lexical distraction: An editing task. *TESL Reporter*, *12*(1), 1-3.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, *7*(1), 13-30.

Briere, E. J., Clausing, G., Senko, D., & Purcell, E. (1978). A look at cloze testing across languages and levels*. *The Modern Language Journal*, *62*(1-2), 23-26.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, *64*(3), 311-317.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing*, *5*(1), 19-31.

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing*, *10*(2), 93-116.

Brown, J. D. (2000). What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(2), 8-12.

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion. *Second Language Studies*, *21*(1), 79-125.

Brown, J. D. (2003). Norm-Reference Item Analysis (Item Facility And Item Discrimination). *Shiken: Jalt Testing & Evaluation Sig Newsletter*, *17*(2), 16-19.

Brown, J. D., Janssen, G., Trace, J., & Kozhevnikova, L. (2012). A preliminary study of cloze procedure as a tool for estimating English readability for Russian students. *Second Language Studies, 31*(1), 1-22.

Canale, M., & Swain, M. (1979). *Communicative Approach to Second Language Teaching and Testing*. Toronto: Ontario Ministry of Education.

Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In Allen, H.B., H.B., R.N. Campbell. *Testing the English proficiency of foreign students*, 313-320.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: what difference does it make?. *Language Testing*, *7*(2), 121-146.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research?. *Second Language Research*, *10*(2), 157-187.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference?. *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences?. *Language Learning*, *27*(1), 63-70.

Chihara, T., Sakurai, T., & Oller, J. W. (1989). Background and culture as factors in EFL reading comprehension. *Language Testing*, *6*(2), 143-149.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports*, *1*(4), 72-115.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cummins, J. (1999). BICS and CALP: Clarifying the Distinction. ERIC.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In *Encyclopedia of Language and Education.* New York: Springer.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, *1*(1), 50-69.

Des Brisay, M. (1994). Problems in developing an alternative to the TOEFL. *TESL Canada Journal*, *12*(1), 47-57.

DeVellis, R. F. (2006). Classical test theory. *Medical care*, *44*(11), S50-S59.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, *9*(2), 187-206.

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. Retrieved on January 4, 2016 from http://files.eric.ed.gov/fulltext/ED506403.pdf

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290-325.

Elder, C., & Von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly*, *5*(3), 173-194.

Farhady, H. (1982). Measures of Language Proficiency from the Learner's Perspective*. *TESOL Quarterly*, *16*(1), 43-59.

Farhady, H. (1996). Varieties of cloze procedure in EFL education. *Foreign Language Teaching Journal*, *12*(44), 217-229.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, *3*(4), 70-85.

Fleisher, L. S., Jenkins, J. R., & Pany, D. (1979). Effects on poor readers' comprehension of training in rapid decoding. *Reading Research Quarterly*, 30-48.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations?*. *Language Learning*, *41*(3), 313-336.

Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, *25*(4), 605-628.

Fox, J. (2005). Rethinking second language admission requirements: Problems with language-residency criteria and the need for language assessment and support. *Language Assessment Quarterly: An International Journal*, *2*(2), 85-115.

Fram, R. D. (1972). A review of the literature related to the cloze procedure. ERIC.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing*, *14*(2), 113-139.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Gamarra, A., & Jonz, J. (1987). Cloze procedure and the sequence of text. *Research in Literacy: Merging Perspectives*, 17-24.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *61*(3), 325-353.

Gellert, A. S., & Elbro, C. (2013). Do experimental measures of word learning predict vocabulary development over time? A study of children from grade 3 to 4. *Learning and Individual Differences*, *26*, 1-8.

Ginther, A. (1986). *Textual Sequence and Cloze Procedure* (Doctoral dissertation). East Texas State University, Texas.

Ginther, A. (2013). The use and interpretation of English proficiency test scores in the graduate admissions process. ESL Go! Newsletter, Purdue University.

Grabe, W. (2010). Reading in a second language. In R.B. Kaplan (Ed.) *Handbook of Applied Linguistics,* 2nd ed. (pp. 88-99). New York: Oxford University Press.

Green, A. B., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing*, *21*(4), 467-494.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley, & C. Stevenson (Eds.). *Taking their Measure: The Validity and Validation of Language Tests* (pp. 219-253). Bochum.

Haan, J. E. (2009). ESL and Internationalization at Purdue University: A History and Analysis. *ProQuest LLC*.

Hinofotis, F. B. (1980). Cloze testing: An overview. *CATESOL Occ. Papers*, *6*, 51-55.

Hymes, D. (1972). On communicative competence. *Sociolinguistics*, *269293*, 269-293.

Ikeguchi, C. B. (1995). Cloze testing options for the classroom. In J. D. Brown and S. Yamashita (Eds.). *Language Testing in Japan*, 166-178.

Ikeguchi, C. B. (1998). Do different C-tests discriminate proficiency levels of EL2 learners. *JALT Testing & Evaluation SIG Newsletter*, *2*(1), 3-8.

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

Jafarpur, A. (1999). Can the C-test be improved with classical item analysis?. *System*, *27*(1), 79-89.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure?. *TESOL Quarterly*, *24*(1), 61-83.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing*, *8*(1), 1-22.

Jordan, G. (2004). *Theory Construction in Second Language Acquisition*. Philadelphia: John Benjamins.

Kamimoto, T. (1992). An inquiry into what a C-test measures. *Fukuoka Women's Junior College Studies*, *44*, 67-79.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*(3), 135-170.

Kane, M.T. (2011). Validating score interpretations and uses: Messick lecture language testing research colloquium, Cambridge April 2010, *Language Testing, 29*(1), 3-17.

Katona, L., & Dörnyei, Z. (1993). The C-test: A teacher friendly way to test language proficiency. *Forum, 31*(2), 35.

Kniffka, G., & Linnemann, M. (2014). A German C-test for migrant children. In R. Grotjahn. *The C-Test: Current Trends*, 239-259.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, *3*, 85-108.

Klein-Braley, C. (1983). A cloze is a cloze is a question. *Issues in Language Testing Research*, 218-228.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test1. *Language Testing, 1*(2), 134-146.

Klein-Braley, C. (1984). Advance Prediction of Difficulty with C-Tests. ERIC.

Klein-Braley, C. (1985). A cloze-up on the C-test: a study in the construct validation of authentic tests. *Language Testing*, *2*(1), 76-104.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, *14*(1), 47-84.

Klein-Braley, C. (1998). *Learning about Language Assessment: Dilemmas, Decisions, and Directions*. New York: Heinle & Heinle.

Kline, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation.* London: Sage Publications.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessments, 2*(2), 48-66.

Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, *86*(4), 571-586.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing*, *29*(2), 291-308.

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing*, *30*(4), 467-489.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book*. London: Longman.

Lado, R. (1961). Linguistics and foreign language teaching. *Language Learning*, *11*(2), 29-52.

Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing*, *3*(2), 130-146.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33-51.

Light, R. L., Xu, M., & Mossop, J. (1987). English proficiency and academic performance of international students. *TESOL Quarterly*, *21*(2), 251-261.

Lin, W. Y., Yuan, H. C., & Feng, H. P. (2008). Language reduced redundancy tests: A reexamination of cloze test and C-Test. *Journal of Pan-Pacific Association of Applied Linguistics*, *12*(1), 61-79.

Manning, W. H. (1986). *Cloze-Elide: A Process Oriented Model of Language Proficiency*. ERIC

Manning, W. H. (1987). *Development of Cloze-Elide Tests of English as a Second Language*. (TOEFL Research Report 23). Educational Testing Service.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Mitchell, R., Myles, F., & Marsden, E. (2013). *Second Language Learning Theories*. London: Routledge.

Mullen, A. (2009). The Impact of Using a Proficiency Test as a Placement Tool: The Case of Test of English for International Communication (TOEIC) (doctoral dissertation). University of Laval, QC: Canada.

Negishi, M. (1987). The C-test: An integrative measure? *The IRLT Bulletin, 1*, 3-26.

Oller Jr, J. W., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly*, *5*(4), 315-326.

Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning*, *21*(2), 183-194.

Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *The Modern Language Journal, 56*(3), 151-158.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning*, *23*(1), 105-118.

Oller, J. W. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.

O'Loughlin, K. (1992). Cloze-What does it really tell us?. *TESOL in Context*, *2*(2), 21.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. New York: Cambridge University Press.

Pickering, M. (1976). Some Observations on Cloze Tests. ERIC: ED140602.

Plakans, L. (2013). Assessment of integrated skills. In *The Encyclopedia of Applied Linguistics*. Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9781405198431/titles.

Poel, C. J., & Weatherly, S. D. (1997). A cloze look at placement testing. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *1*(1), 4-10.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning*, *28*(2), 333-341.

Porter, D. (1988). Book review: Manning, WH 1987: Development of cloze-elide tests of English as a second language. TOEFL Research Report 23, April 1987, Princeton, New Jersey: Educational Testing Service. *Language Testing*, *5*(2), 250-252.

Purdue University. (2014). *English Proficiency & Other Standardized Tests*. Retrieved from http://www.iss.purdue.edu/admission/ugrad/tests.cfm.

Raatz, U. (1984). The factorial validity of C-Tests. In Culhane, T., Klein-Braley, C. and Stevenson, D.K., editors, *Practice and Problems in Language Testing 7*, Colchester: University of Essex.

Rankin, E. F., & Culhane, J. W. (1969). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *13*(3), 193-198.

Raymond, P. M. (1988). Close Procedure in the Teaching of Reading. *TESL Canada Journal*, *6*(1), 91-97.

Razi, S. (2005). A fresh look at the evaluation of ordering tasks in reading comprehension: weighted marking protocol. *Reading Matrix: An International Online Journal*, *5*(1), 1-14.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, *7*(3), 180-190.

Read, J. (2013). Issues in post-entry language assessment in English-medium universities. *Language Teaching*, *48*(02), 217-234.

Read, J., & Von Randow, J. (2013). A university post-entry English language assessment: Charting the changes. *International Journal of English Studies*, *13*(2), 89-110.

Read, J. (2015). *Assessing English Proficiency for University Study*. New York: Palgrave Macmillan.

Rye, J. (1979). A Closer Look At 'Cloze'. *English in Education*, *13*(3), 44-54.

Rye, J. (1982). Cloze Procedure and the Teaching of Reading. London: Heinemann.

Saeedi, M., Tavakoli, M., Rahimi Kazerooni, S., & Parvaresh, V. (2011). Do C-test and cloze procedure measure what they purport to be measuring?: A case of criterion-related validity. *International Journal of Human and Social Sciences*, *6*, 2-99.

Savignon, S. J. (1983). *Communicative Competence: Theory and Classroom Practice. Texts and Contexts in Second Language Learning*. MA: Addison-Wesley Publishing.

Sciarone, A. G., & Schoorl, J. J. (1989). The cloze test: Or why small isn't always beautiful.* *Language Learning*, *39*(3), 415-438.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*(2), 229-255.

Sheehan, K. M. (2015). Aligning TextEvaluator scores with the accelerated text complexity guidelines specified in the Common Core State Standards. (ETS Research Report) Princeton, NJ: Educational Testing Service.

Shohamy, E. (1982). Predicting speaking proficiency from cloze tests: theoretical and practical considerations for tests substitution. *Applied Linguistics*, *3*(2), 161-171.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

Sigott, G. (2004). *Towards Identifying the C-Test Construct*. New York: Peter Lang.

Sigott, G., & Köberl, J. (1996). Deletion patterns and C-test difficulty across languages. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*, *3*, 159-172.

Spolsky, B. (1969). *Reduced Redundancy as a Language Testing Tool*. Retrieved from ERIC database. (ED031702).

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31-40.

Soureshjani, K. H. (2012). Cognitive styles on c-test and cloze-elide test: Which style acts better?. *Language Testing in Asia*, *2*(2), 61.

Steinman, L. (2002). Considering the cloze. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *59*(2), 291-301.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*(2), 214-231.

Tardy, C. (2015, October). *Writing Assessment for Placement from an L2 Writing Perspective.* Plenary talk given at the Midwest Association of Language Testers Conference, Iowa City, IA.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.

Taylor, W. L. (1956). Recent Developments in the Use of "Cloze Procedure". *Journalism & Mass Communication Quarterly*, *33*(1), 42-99.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition*, *33*(3), 339-372.

Turner, C. E. (1989). The underlying factor structure of L2 close test performance in Francophone, university-level students: Causal modeling as an approach to construct validation. *Language Testing*, *6*(2), 172-197.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 321-344.

Wan, T. Y., Chapman, D. W., & Biggs, D. A. (1992). Academic stress of international students attending US universities. *Research in Higher Education*, *33*(5), 607-623.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Examinations, 37*, 2-10.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, *14*(5), 2.

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, *34*(2), 215-231.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, *45*(4), 628-660.

Wood, R. (1993). *Authenticity in Language Testing: Some Outstanding Questions*. Cambridge: Cambridge University Press.

Xi, X. (2008). Methods of test validation. *Encyclopedia of Language and Education*, *7*, 177-96.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267-293.

# APPENDICES

Appendix A. Distributions of Test Takers across Their L1 and Study Programs

| | L1 background | | | | |
|---|---|---|---|---|---|
| Study programs | Chinese | Korean | English | Other | **Total** |
| Agriculture | 1 | 0 | 0 | 0 | 1 (0.3%) |
| Health | 6 | 0 | 0 | 0 | 6 (2%) |
| Liberal Arts | 12 | 1 | 3 | 1 | 17 (7%) |
| Management | 36 | 3 | 0 | 2 | 41 (16%) |
| Science | 62 | 1 | 1 | 3 | 67 (26%) |
| Technology | 3 | 0 | 1 | 2 | 6 (2%) |
| Engineering | 1 | 0 | 2 | 3 | 6 (2%) |
| Exploratory Studies | 66 | 0 | 0 | 0 | 66 (25%) |
| Unidentified | 18 | 4 | 21 | 7 | 50 (19%) |
| **Total** | 205 (79%) | 9 (3%) | 28 (11%) | 18 (7%) | 260 (100%) |

Appendix B. Flesch Kincaid Grade Level Readability Scores of C-test and Cloze-elide Passages

| Form | Passage | Topic | Flesch Kincaid |
|---|---|---|---|
| C-test | | | |
| 1 | 2 | Riding the bus | 11.83 |
| | 3 | Interlibrary loan | 9.59 |
| | 4 | Parking on-campus | 10.12 |
| | 5 | Syllabus course policy | 11.77 |
| 2 | 2 | Credit hours | 14.96 |
| | 3 | English 106-I | 10.93 |
| | 4 | Group work | 8.92 |
| | 5 | On-campus housing | 12.65 |
| 3 | 2 | Campus safety | 12.9 |
| | 3 | Going to the gym | 11.26 |
| | 4 | Undergraduate research | 12.73 |
| | 5 | Writing lab | 10.45 |
| 4 | 2 | Alumni association | 11.08 |
| | 3 | OEPT | 15.4 |
| | 4 | Student club | 13.97 |
| | 5 | University Hall | 10.99 |
| Cloze-elide | | | |
| 1 | 2 | Research angle | 9.21 |
| | 3 | Job descriptions of a manager | 11.78 |
| 2 | 2 | E-books | 9.65 |
| | 3 | Culture | 9.07 |
| 3 | 2 | Plagiarism | 11.67 |
| | 3 | Greek society | 11.83 |
| 4 | 2 | Writing purpose | 10.68 |
| | 3 | Extracurricular activities | 9.92 |

Appendix C. TextEvaluator Complexity Scores of C-test and Cloze-elide Passages

| Form | Passage | Topic | Syntactic complexity | Vocabulary difficulty | Lexical cohesion | Prior knowledge | Overall complexity |
|---|---|---|---|---|---|---|---|
| C-test | | | | | | | |
| 1 | 2 | Riding the bus | 49 | 73 | 61 | 54 | 790 |
| | 3 | Interlibrary loan | 68 | 44 | 51 | 4 | 869 |
| | 4 | Parking on-campus | 52 | 75 | 59 | 10 | 869 |
| | 5 | Syllabus course policy | 47 | 77 | 60 | 39 | 760 |
| 2 | 2 | Credit hours | 68 | 95 | 51 | 43 | 1360 |
| | 3 | English 106-I | 47 | 81 | 53 | 39 | 960 |
| | 4 | Group work | 58 | 69 | 60 | 38 | 730 |
| | 5 | On-campus housing | 67 | 75 | 55 | 39 | 880 |
| 3 | 2 | Campus safety | 50 | 70 | 53 | 4 | 780 |
| | 3 | Going to the gym | 72 | 61 | 54 | 45 | 1080 |
| | 4 | Undergraduate research | 52 | 83 | 53 | 4 | 740 |
| | 5 | Writing lab | 50 | 70 | 61 | 38 | 930 |
| 4 | 2 | Alumni association | 53 | 62 | 62 | 51 | 880 |
| | 3 | OEPT | 74 | 90 | 54 | 65 | 1230 |
| | 4 | Student club | 50 | 91 | 62 | 55 | 960 |
| | 5 | University Hall | 47 | 54 | 57 | 64 | 770 |
| Cloze-elide | | | | | | | |
| 1 | 2 | Research angle | 55 | 72 | 66 | 7 | 850 |
| | 3 | Job descriptions | 52 | 97 | 60 | 7 | 1110 |
| 2 | 2 | E-books | 51 | 70 | 73 | 21 | 860 |
| | 3 | Culture | 54 | 66 | 55 | 35 | 720 |

(Continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 2 | Plagiarism | 46 | 82 | 57 | 29 | 1100 |
| | 3 | Greek society | 48 | 83 | 71 | 28 | 1010 |
| 4 | 2 | Writing purpose | 60 | 60 | 63 | 24 | 940 |
| | 3 | Extracurricular activities | 50 | 79 | 60 | 7 | 869 |

Appendix D. Average Means and Standard Deviations of Item Difficulty and Item Discrimination of Each Test Passage

| Passage | Number of items | Topic | Item difficulty | | Point biserial correlation | |
|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD |
| Form 1 | | | | | | |
| C-test | | | | | | |
| Passage 2 | 25 | Riding the bus | 0.73 | 0.21 | 0.37 | 0.18 |
| Passage 3 | 25 | Interlibrary loan | 0.87 | 0.15 | 0.26 | 0.18 |
| Passage 4 | 25 | Parking on-campus | 0.66 | 0.21 | 0.37 | 0.18 |
| Passage 5 | 24 | Syllabus course policy | 0.84 | 0.16 | 0.34 | 0.16 |
| Cloze-elide | | | | | | |
| Passage 2 | 30 | Research angle | 0.64 | 0.22 | 0.52 | 0.16 |
| Passage 3 | 30 | Job descriptions of a manager | 0.66 | 0.17 | 0.57 | 0.14 |
| Form 2 | | | | | | |
| C-test | | | | | | |
| Passage 2 | 25 | Credit hours | 0.67 | 0.20 | 0.48 | 0.17 |
| Passage 3 | 24 | English 106-I | 0.85 | 0.15 | 0.31 | 0.17 |
| Passage 4 | 25 | Group work | 0.79 | 0.21 | 0.33 | 0.18 |
| Passage 5 | 25 | On-campus housing | 0.90 | 0.14 | 0.27 | 0.28 |
| Cloze-elide | | | | | | |
| Passage 2 | 30 | E-books | 0.60 | 0.14 | 0.63 | 0.11 |
| Passage 3 | 30 | Culture | 0.66 | 0.10 | 0.65 | 0.11 |
| Form 3 | | | | | | |
| C-test | | | | | | |
| Passage 2 | 24 | Campus safety | 0.75 | 0.18 | 0.39 | 0.19 |
| Passage 3 | 24 | Going to the gym | 0.83 | 0.16 | 0.34 | 0.23 |

(Continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Passage 4 | 25 | Undergraduate research | 0.83 | 0.14 | 0.26 | 0.18 |
| Passage 5 | 25 | Writing lab | 0.76 | 0.18 | 0.34 | 0.15 |
| Cloze-elide | | | | | | |
| Passage 2 | 30 | Plagiarism | 0.50 | 0.19 | 0.45 | 0.12 |
| Passage 3 | 30 | Greek society | 0.63 | 0.17 | 0.51 | 0.12 |
| Form 4 | | | | | | |
| C-test | | | | | | |
| Passage 2 | 25 | Alumni association | 0.88 | 0.15 | 0.32 | 0.18 |
| Passage 3 | 25 | OEPT | 0.69 | 0.28 | 0.33 | 0.23 |
| Passage 4 | 25 | Student club | 0.77 | 0.20 | 0.35 | 0.20 |
| Passage 5 | 25 | University Hall | 0.78 | 0.28 | 0.24 | 0.22 |
| Cloze-elide | | | | | | |
| Passage 2 | 30 | Writing purpose | 0.58 | 0.18 | 0.54 | 0.10 |
| Passage 3 | 29 | Extracurricular activities | 0.61 | 0.16 | 0.50 | 0.12 |

Appendix E. Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Item in the Pilot Data

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 3: Interlibrary loan (The number of acceptable items is 2.) | | | | | |
| 1 | this | Function | Det | 0.55 | 0.22 |
| 2 | you | Function | Pronoun | 1.00 | 0 |
| 3* | consider | Content | Verb | 0.56 | 0.43 |
| 4 | the | Function | Article | 0.98 | 0.20 |
| 5 | from | Function | Prep | 0.98 | 0.43 |
| 6 | library | Content | Noun | 0.98 | 0.15 |
| 7 | may | Function | Modal | 0.95 | 0.42 |
| 8* | like | Function | Prep | 0.70 | 0.40 |
| 9 | lot | Function | Pronoun | 0.71 | 0.63 |
| 10 | trouble | Content | Noun | 0.80 | 0.37 |
| 11 | it | Function | Pronoun | 0.98 | -0.02 |
| 12 | actually | Content | Adv | 0.97 | -0.03 |
| 13 | simple | Content | Adj | 0.76 | 0.49 |
| 14 | there | Function | Pronoun | 0.80 | 0.32 |
| 15 | many | Function | Det | 0.94 | 0.28 |
| 16 | that | Function | Det | 0.77 | 0.41 |
| 17 | in | Function | Prep | 0.85 | 0.41 |
| 18 | kind | Content | Noun | 0.91 | 0.26 |
| 19 | system | Content | Noun | 0.98 | 0.09 |
| 20 | you | Function | Pronoun | 1.00 | 0 |
| 21 | to | Function | Prep | 0.98 | 0.09 |
| 22 | is | Function | Verb to be | 0.97 | 0.13 |
| | | | | (Continue) | |
| 23 | request | Content | Noun | 0.59 | 0.21 |
| 24 | the | Function | Article | 0.97 | 0.25 |
| 25 | you | Function | Pronoun | 0.95 | 0.46 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 4: Parking on-campus (The number of acceptable items is 10.) | | | | | |
| 1 | must | Function | Modal | 0.95 | 0.00 |
| 2 | parking | Content | Noun | 0.94 | 0.26 |
| 3 | for | Function | Prep | 0.97 | 0.09 |
| 4* | vehicle | Content | Noun | 0.70 | 0.46 |
| 5 | some | Function | Det | 0.73 | 0.18 |
| 6 | halls | Content | Noun | 0.67 | 0.23 |
| 7 | parking | Content | Noun | 0.91 | 0.26 |
| 8 | in | Function | Prep | 0.79 | 0.31 |
| 9* | fees | Content | Noun | 0.52 | 0.38 |
| 10* | check | Content | Verb | 0.38 | 0.57 |
| 11 | housing | Content | Noun | 0.59 | 0.48 |
| 12 | carefully | Content | Adv | 0.76 | 0.34 |
| 13 | you | Function | Pronoun | 1.00 | 0 |
| 14* | guests | Content | Noun | 0.53 | 0.48 |
| 15 | can | Function | Modal | 0.70 | 0.29 |
| 16 | at | Function | Prep | 0.71 | 0.50 |
| 17* | visitor | Content | Noun | 0.35 | 0.35 |
| 18* | garage | Content | Noun | 0.33 | 0.53 |
| 19 | at | Function | Prep | 0.76 | 0.49 |
| 20 | residence | Content | Noun | 0.71 | 0.21 |
| 21* | guest | Content | Noun | 0.56 | 0.47 |
| 22* | permits | Content | Noun | 0.59 | 0.64 |
| 23* | available | Content | Adj | 0.68 | 0.57 |
| 24 | anyone | Function | Pronoun | 0.17 | 0.48 |
| 25* | student | Content | Noun | 0.61 | 0.67 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 5: Syllabus course policy (The number of acceptable items is 3.) | | | | | |
| 1 | attendence | Content | Noun | 0.38 | 0.29 |
| 2 | participation | Content | Noun | 0.80 | 0.32 |
| 3 | class | Content | Noun | 0.92 | 0.49 |
| 4* | are | Function | Verb to be | 0.70 | 0.64 |
| 5 | to | Function | Prep | 1.00 | 0 |
| 6 | success | Content | Noun | 0.89 | 0.23 |
| 7* | are | Function | Verb to be | 0.62 | 0.54 |
| 8* | included | Content | Verb | 0.58 | 0.35 |
| 9 | part | Content | Noun | 0.95 | -0.02 |
| 10 | course | Content | Noun | 0.73 | 0.45 |
| 11 | in | Function | Prep | 0.97 | 0.19 |
| 12 | as | Function | Prep | 0.86 | 0.41 |
| 13 | college | Content | Noun | 0.83 | 0.34 |
| 14 | you | Function | Pronoun | 0.98 | 0.13 |
| 15 | expected | Content | Verb | 0.91 | 0.25 |
| 16 | attend | Content | Verb | 0.97 | 0.44 |
| 17 | class | Content | Noun | 0.92 | 0.35 |
| 18 | and | Function | Conj | 0.88 | 0.33 |
| 19 | at | Function | Prep | 0.95 | 0.36 |
| 20 | classroom | Content | Noun | 0.73 | 0.56 |
| 21 | time | Content | Noun | 0.97 | 0.28 |
| 22 | you | Function | Pronoun | 0.98 | 0.26 |
| 23 | inform | Content | Verb | 0.83 | 0.37 |
| 24 | instructor | Content | Noun | 0.71 | 0.52 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 2: Credit hours (The number of acceptable items is 12.) | | | | | |
| 1 | addition | Function | Noun | 0.93 | 0.25 |
| 2* | maintaining | Content | Verb | 0.52 | 0.52 |
| 3* | valid | Content | Adj | 0.60 | 0.50 |
| 4* | and | Function | Conj | 0.71 | 0.37 |
| 5 | immigration | Content | Noun | 0.74 | 0.21 |
| 6 | you | Function | Pronoun | 0.97 | 0.33 |
| 7 | to | Function | Prep | 0.97 | 0.33 |
| 8 | in | Function | Prep | 0.84 | 0.35 |
| 9 | university | Content | Noun | 0.90 | 0.30 |
| 10 | time | Content | Noun | 0.78 | 0.47 |
| 11* | this | Function | Det | 0.48 | 0.39 |
| 12 | registering | Content | Verb | 0.19 | 0.27 |
| 13* | minimum | Content | Noun | 0.55 | 0.45 |
| 14 | credit | Content | Noun | 0.83 | 0.47 |
| 15 | per | Function | Prep | 0.88 | 0.51 |
| 16* | this | Function | Det | 0.43 | 0.56 |
| 17* | applies | Content | Verb | 0.34 | 0.60 |
| 18 | every | Function | Det | 0.72 | 0.62 |
| 19 | student | Content | Noun | 0.74 | 0.57 |
| 20 | every | Function | Det | 0.72 | 0.66 |
| 21* | or | Function | Conj | 0.50 | 0.67 |
| 22* | semester | Content | Noun | 0.67 | 0.77 |
| 23* | during | Function | Prep | 0.67 | 0.72 |
| 24* | last | Function | Det | 0.60 | 0.69 |
| 25* | of | Function | Prep | 0.43 | 0.50 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 3: English 106-I (The number of acceptable items is 2.) | | | | | |
| 1 | course | Content | Noun | 1.00 | 0 |
| 2 | designed | Content | Verb | 0.72 | 0.61 |
| 3 | for | Function | Prep | 1.00 | 0 |
| 4 | students | Content | Noun | 0.88 | 0.22 |
| 5 | more | Function | Det | 0.79 | 0.20 |
| 6 | on | Function | Prep | 0.97 | 0.02 |
| 7 | and | Function | Conj | 0.93 | 0.31 |
| 8 | it | Function | Pronoun | 0.97 | 0.24 |
| 9 | typically | Content | Adv | 0.90 | 0.47 |
| 10 | by | Function | Prep | 0.88 | 0.38 |
| 11 | with | Function | Prep | 0.74 | 0.47 |
| 12 | in | Function | Prep | 0.90 | 0.32 |
| 13 | composition | Content | Noun | 0.74 | 0.50 |
| 14 | an | Function | Article | 0.91 | 0.01 |
| 15* | of | Function | Prep | 0.55 | 0.39 |
| 16 | it | Function | Pronoun | 0.91 | 0.08 |
| 17* | like | Content | Adj | 0.34 | 0.40 |
| 18 | write | Content | Verb | 0.83 | 0.48 |
| 19 | second | Content | Number | 0.93 | 0.39 |
| 20 | course | Content | Noun | 0.95 | 0.11 |
| 21 | an | Function | Article | 0.97 | 0.20 |
| 22 | limit | Content | Noun | 0.71 | 0.60 |
| 23 | students | Content | Noun | 0.88 | 0.28 |
| 24 | class | Content | Noun | 0.90 | 0.47 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 4: Group work (The number of acceptable items is 6.) | | | | | |
| 1* | at | Function | Prep | 0.47 | 0.50 |
| 2* | point | Content | Noun | 0.41 | 0.43 |
| 3 | your | Function | Det | 0.95 | 0.28 |
| 4* | life | Content | Noun | 0.53 | 0.44 |
| 5 | on | Function | Prep | 0.93 | 0.26 |
| 6* | size | Content | Noun | 0.34 | 0.41 |
| 7 | your | Function | Det | 1.00 | 0 |
| 8 | you | Function | Pronoun | 1.00 | 0 |
| 9 | be | Function | Verb to be | 0.97 | 0.26 |
| 10 | to | Function | Prep | 0.98 | 0.12 |
| 11 | on | Function | Prep | 0.84 | 0.41 |
| 12* | or | Function | Conj | 0.36 | 0.27 |
| 13 | projects | Content | Noun | 1.00 | 0 |
| 14 | part | Content | Noun | 0.78 | 0.00 |
| 15* | team | Content | Noun | 0.60 | 0.49 |
| 16 | fact | Function | Noun | 0.72 | 0.34 |
| 17 | work | Content | Noun | 0.90 | 0.48 |
| 18 | make | Content | Verb | 0.90 | 0.43 |
| 19 | more | Function | Det | 0.95 | 0.53 |
| 20 | and | Function | Conj | 0.97 | 0.35 |
| 21 | group | Content | Noun | 0.72 | 0.41 |
| 22 | gives | Content | Verb | 0.83 | 0.51 |
| 23 | great | Content | Adj | 0.78 | 0.42 |
| 24 | to | Function | Prep | 0.93 | 0.46 |
| 25 | from | Function | Prep | 0.88 | 0.53 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 5: On-campus housing (The number of acceptable items is 3.) | | | | | |
| 1 | you | Function | Pronoun | 0.98 | 0.08 |
| 2 | on | Function | Prep | 0.98 | -0.08 |
| 3 | you | Function | Pronoun | 1.00 | 0 |
| 4 | not | Function | Negation | 1.00 | 0 |
| 5 | to | Function | Prep | 1.00 | 0 |
| 6 | about | Function | Prep | 1.00 | 0 |
| 7 | or | Function | Conj | 0.74 | 0.36 |
| 8 | to | Function | Prep | 1.00 | 0 |
| 9 | on | Function | Prep | 0.98 | 0.13 |
| 10 | because | Function | Conj | 0.97 | 0.23 |
| 11 | university | Content | Noun | 0.93 | 0.55 |
| 12 | are | Function | Verb to be | 0.84 | 0.68 |
| 13* | walking | Content | Verb | 0.52 | 0.51 |
| 14 | from | Function | Prep | 0.97 | 0.03 |
| 15* | residence | Content | Noun | 0.59 | 0.46 |
| 16 | living | Content | Verb | 1.00 | 0 |
| 17 | campus | Content | Noun | 0.98 | 0.18 |
| 18 | also | Function | Conj | 0.72 | 0.64 |
| 19 | you | Function | Pronoun | 1.00 | 0 |
| 20 | opportunities | Content | Noun | 0.88 | 0.60 |
| 21 | make | Content | Verb | 0.91 | 0.53 |
| 22 | and | Function | Conj | 1.00 | 0 |
| 23* | involved | Content | Verb | 0.62 | 0.64 |
| 24 | the | Function | Article | 0.88 | 0.58 |
| 25 | university | Content | Noun | 0.93 | 0.68 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 2: Campus safety (The number of acceptable items is 6.) | | | | | |
| 1 | however | Function | Conj | 0.96 | 0.20 |
| 2 | is | Function | Verb to be | 0.98 | 0.00 |
| 3* | reality | Content | Noun | 0.58 | 0.30 |
| 4* | most | Function | Det | 0.65 | 0.32 |
| 5 | campuses | Content | Noun | 0.67 | 0.23 |
| 6* | make | Content | Verb | 0.51 | 0.57 |
| 7* | community | Content | Noun | 0.67 | 0.31 |
| 8 | safe | Content | Adj | 0.82 | 0.44 |
| 9 | to | Function | Prep | 0.98 | -0.04 |
| 10 | all | Function | Det | 0.42 | 0.28 |
| 11 | should | Function | Modal | 0.95 | 0.19 |
| 12 | in | Function | Prep | 0.76 | 0.30 |
| 13 | that | Function | Det | 0.75 | 0.24 |
| 14* | and | Function | Conj | 0.69 | 0.44 |
| 15 | responsibility | Content | N | 0.78 | 0.51 |
| 16 | everyone | Function | Pronoun | 0.82 | 0.66 |
| 17 | is | Function | Verb to be | 0.91 | 0.62 |
| 18 | to | Function | Prep | 0.87 | 0.61 |
| 19 | safety | Content | Noun | 0.84 | 0.51 |
| 20 | security | Content | Noun | 0.80 | 0.57 |
| 21* | when | Function | Conj | 0.67 | 0.46 |
| 22 | student | Content | Noun | 0.84 | 0.49 |
| 23 | an | Function | Article | 0.76 | 0.62 |
| 24 | part | Content | Noun | 0.27 | 0.45 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 3: Going to the gym (The number of acceptable items is 5.) | | | | | |
| 1 | to | Function | Prep | 1.00 | 0 |
| 2 | from | Function | Prep | 0.98 | 0.25 |
| 3 | students | Content | Noun | 0.96 | 0.21 |
| 4* | worked | Content | Verb | 0.58 | 0.35 |
| 5 | campus | Content | Noun | 0.95 | -0.09 |
| 6 | at | Function | Prep | 1.00 | 0 |
| 7 | once | Function | Adv | 0.82 | 0.54 |
| 8 | week | Content | Noun | 0.85 | 0.59 |
| 9 | more | Function | Det | 0.78 | 0.00 |
| 10 | to | Function | Prep | 0.98 | 0.25 |
| 11 | higher | Content | Adj | 0.91 | 0.32 |
| 12* | point | Content | Noun | 0.53 | 0.64 |
| 13* | than | Function | Conj | 0.69 | 0.50 |
| 14 | who | Function | Pronoun | 0.84 | 0.59 |
| 15 | less | Function | Det | 0.78 | 0.67 |
| 16 | not | Function | Negation | 0.80 | 0.46 |
| 17* | all | Function | Det | 0.62 | 0.61 |
| 18 | is | Function | Verb to be | 0.96 | 0.05 |
| 19 | students | Content | Noun | 0.95 | 0.22 |
| 20 | by | Function | Prep | 0.76 | 0.39 |
| 21 | and | Function | Conj | 0.82 | 0.27 |
| 22* | tend | Content | Verb | 0.42 | 0.53 |
| 23 | have | Content | Verb | 0.95 | 0.37 |
| 24 | time | Content | Noun | 0.89 | 0.33 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 4: Undergraduate research (The number of acceptable items is 3.) | | | | | |
| 1* | in | Function | Prep | 0.56 | 0.45 |
| 2 | research | Content | Noun | 0.80 | 0.41 |
| 3 | gives | Content | Verb | 0.84 | 0.34 |
| 4* | considerable | Content | Adj | 0.51 | 0.56 |
| 5 | skills | Content | Noun | 0.87 | 0.26 |
| 6 | to | Function | Prep | 1.00 | 0 |
| 7 | apply | Content | Verb | 0.82 | 0.19 |
| 8 | jobs | Content | Noun | 0.87 | 0.20 |
| 9 | graduate | Content | Noun | 1.00 | 0 |
| 10 | many | Function | Det | 0.75 | 0.21 |
| 11 | programs | Content | Noun | 0.89 | 0.06 |
| 12 | campus | Content | Noun | 1.00 | 0 |
| 13 | you | Function | Pronoun | 0.96 | 0.24 |
| 14 | opportunities | Content | Noun | 0.98 | -0.06 |
| 15 | work | Content | Verb | 0.93 | 0.32 |
| 16 | professors | Content | Noun | 0.76 | 0.33 |
| 17 | other | Function | Det | 0.82 | 0.48 |
| 18* | more | Function | Det | 0.49 | 0.35 |
| 19 | it | Function | Pronoun | 0.89 | 0.31 |
| 20 | you | Function | Pronoun | 1.00 | 0 |
| 21 | the | Function | Article | 0.89 | 0.42 |
| 22 | of | Function | Prep | 0.73 | 0.29 |
| 23 | that | Function | Pronoun | 0.75 | 0.40 |
| 24 | your | Function | Det | 0.87 | 0.32 |
| 25 | and | Function | Conj | 0.80 | 0.46 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 5: Writing lab (The number of acceptable items is 6.) | | | | | |
| 1* | find | Content | Verb | 0.65 | 0.31 |
| 2 | tutorials | Content | Noun | 0.71 | 0.29 |
| 3 | because | Function | Conj | 0.96 | 0.17 |
| 4 | tutors | Content | Noun | 0.82 | 0.16 |
| 5 | help | Content | Verb | 0.96 | 0.17 |
| 6* | select | Content | Verb | 0.64 | 0.43 |
| 7 | and | Function | Conj | 0.91 | 0.13 |
| 8 | feedback | Content | Noun | 0.91 | 0.39 |
| 9 | their | Function | Det | 0.80 | 0.43 |
| 10 | students | Content | Noun | 0.95 | 0.03 |
| 11 | bring | Content | Verb | 0.95 | 0.38 |
| 12 | papers | Content | Noun | 0.82 | 0.17 |
| 13 | their | Function | Det | 0.84 | 0.38 |
| 14 | they | Function | Pronoun | 0.84 | 0.28 |
| 15 | also | Function | Conj | 0.71 | 0.21 |
| 16 | tutorials | Content | Noun | 0.76 | 0.43 |
| 17 | work | Content | Verb | 0.91 | 0.26 |
| 18 | resumes | Content | Noun | 0.24 | 0.38 |
| 19 | applications | Content | Noun | 0.89 | 0.41 |
| 20* | any | Function | Det | 0.44 | 0.35 |
| 21 | writing | Content | Noun | 0.82 | 0.43 |
| 22* | they | Function | Pronoun | 0.51 | 0.71 |
| 23* | working | Content | Verb | 0.65 | 0.55 |
| 24* | including | Content | Verb | 0.64 | 0.48 |
| 25 | for | Function | Prep | 0.75 | 0.52 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 2: Alumni association (The number of acceptable items is 3.) | | | | | |
| 1 | are | Function | Verb to be | 0.96 | 0.11 |
| 2 | for | Function | Prep | 1.00 | 0 |
| 3 | to | Function | Prep | 1.00 | 0 |
| 4* | contact | Content | Noun | 0.43 | 0.48 |
| 5 | your | Function | Det | 0.98 | 0.43 |
| 6 | and | Function | Conj | 0.92 | 0.38 |
| 7 | friends | Content | Noun | 0.98 | 0.00 |
| 8 | your | Function | Det | 0.98 | 0.43 |
| 9 | are | Function | Verb to be | 0.87 | 0.25 |
| 10 | to | Function | Prep | 1.00 | 0 |
| 11 | in | Function | Prep | 0.96 | 0.30 |
| 12 | is | Function | Verb to be | 0.83 | 0.42 |
| 13 | the | Function | Article | 0.94 | 0.31 |
| 14 | association | Content | Noun | 0.89 | 0.43 |
| 15* | group | Content | Noun | 0.47 | 0.53 |
| 16 | annual | Content | Conj | 0.79 | 0.42 |
| 17 | and | Function | Conj | 0.81 | 0.55 |
| 18 | monthly | Content | Adj | 0.81 | 0.38 |
| 19 | to | Function | Prep | 0.98 | 0.43 |
| 20* | alumni | Content | Noun | 0.68 | 0.52 |
| 21 | in | Function | Prep | 0.92 | 0.55 |
| 22 | with | Function | Prep | 0.96 | 0.46 |
| 23 | other | Function | Det | 0.92 | 0.22 |
| 24 | with | Function | Prep | 0.87 | 0.36 |
| 25 | university | Content | Noun | 0.96 | 0.15 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 3: OEPT (The number of acceptable items is 8.) | | | | | |
| 1 | was | Function | Verb to be | 0.94 | 0.13 |
| 2 | to | Function | Prep | 0.98 | -0.11 |
| 3* | out | Function | Prep | 0.47 | 0.43 |
| 4 | university | Content | Noun | 0.96 | 0.21 |
| 5* | which | Function | Pronoun | 0.68 | 0.42 |
| 6 | that | Function | Comp | 0.57 | 0.53 |
| 7 | international | Content | Adj | 0.94 | -0.05 |
| 8* | assistants | Content | Noun | 0.45 | 0.35 |
| 9 | do | Function | Verb to be | 0.98 | 0.01 |
| 10 | speak | Content | Verb | 0.96 | 0.01 |
| 11 | as | Function | Prep | 0.79 | 0.53 |
| 12 | first | Content | Number | 0.77 | 0.57 |
| 13* | must | Function | Modal | 0.64 | 0.53 |
| 14 | sufficient | Content | Adj | 0.81 | 0.31 |
| 15 | speaking | Content | Verb | 0.98 | -0.03 |
| 16 | otherwise | Function | Conj | 0.87 | 0.31 |
| 17 | students | Content | Noun | 0.83 | 0.14 |
| 18 | cannot | Function | Modal | 0.96 | 0.24 |
| 19 | assigned | Content | Verb | 0.04 | 0.44 |
| 20 | duties | Content | Noun | 0.17 | 0.43 |
| 21 | involve | Content | Verb | 0.09 | 0.39 |
| 22* | teaching | Content | Verb | 0.53 | 0.57 |
| 23* | direct | Content | Adj | 0.47 | 0.47 |
| 24* | with | Function | Prep | 0.68 | 0.67 |
| 25* | students | Content | Noun | 0.64 | 0.61 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 4: Student club (The number of acceptable items is 4.) | | | | | |
| 1* | fact | Content | Noun | 0.75 | 0.17 |
| 2 | has | Function | Verb | 0.91 | -0.11 |
| 3 | of | Function | Prep | 0.83 | 0.56 |
| 4 | largest | Content | Adj | 0.87 | 0.28 |
| 5 | and | Function | Conj | 0.92 | 0.32 |
| 6 | community | Content | Noun | 0.85 | 0.49 |
| 7 | all | Function | Det | 0.32 | 0.13 |
| 8* | clubs | Content | Noun | 0.43 | 0.41 |
| 9 | students | Content | Noun | 0.98 | 0.22 |
| 10 | opportunities | Content | Noun | 0.94 | 0.16 |
| 11 | build | Content | Verb | 0.68 | 0.43 |
| 12 | relationships | Content | Noun | 0.87 | 0.25 |
| 13 | get | Content | Verb | 0.89 | 0.08 |
| 14 | at | Function | Prep | 0.85 | 0.47 |
| 15 | students | Content | Noun | 0.94 | 0.19 |
| 16 | members | Content | Noun | 0.83 | 0.29 |
| 17 | and | Function | Conj | 0.79 | 0.39 |
| 18 | together | Content | Adv | 0.92 | 0.32 |
| 19 | are | Function | Verb to be | 0.32 | 0.18 |
| 20 | in | Function | Prep | 0.92 | 0.49 |
| 21* | variety | Content | Noun | 0.34 | 0.46 |
| 22 | activities | Content | Noun | 0.87 | 0.60 |
| 23 | enhance | Content | Verb | 0.79 | 0.61 |
| 24 | experiences | Content | Noun | 0.75 | 0.71 |
| 25* | college | Content | Noun | 0.66 | 0.61 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 5: University Hall (The number of acceptable items is 6.) | | | | | |
| 1 | building | Content | Noun | 1.00 | 0 |
| 2 | is | Function | Verb to be | 1.00 | 0 |
| 3 | named | Content | Verb | 0.85 | 0.36 |
| 4* | person | Content | Noun | 0.62 | 0.44 |
| 5 | university | Content | Noun | 1.00 | 0 |
| 6 | this | Function | Det | 0.77 | -0.12 |
| 7 | is | Function | Verb to be | 0.98 | -0.04 |
| 8* | only | Function | Adj | 0.60 | 0.54 |
| 9* | remaining | Content | Verb | 0.34 | 0.44 |
| 10 | the | Function | Article | 0.96 | 0.20 |
| 11 | buildings | Content | Noun | 0.92 | 0.17 |
| 12 | stood | Content | Verb | 0.08 | 0.29 |
| 13 | campus | Content | Noun | 0.96 | 0.24 |
| 14 | the | Function | Article | 1.00 | 0 |
| 15 | was | Function | Verb to be | 1.00 | 0 |
| 16 | in | Function | Prep | 0.98 | 0.14 |
| 17 | the | Function | Article | 1.00 | 0 |
| 18 | began | Content | Verb | 0.25 | 0.41 |
| 19 | on | Function | Prep | 0.74 | 0.17 |
| 20 | building | Content | Noun | 0.96 | 0.16 |
| 21* | few | Function | Det | 0.49 | 0.53 |
| 22 | after | Function | Conj | 0.91 | 0.44 |
| 23 | university | Content | Noun | 0.96 | 0.41 |
| 24* | founded | Content | Verb | 0.58 | 0.59 |
| 25* | then | Function | Conj | 0.42 | 0.55 |

Appendix F. The Items with the Highest Item Difficulty Values for Each Test Form in the Pilot Data

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| C-Passage3Item1 | course | Content | Noun | 1.00 | 0.00 |
| C-Passage3Item3 | for | Function | Prep | 1.00 | 0.00 |
| C-Passage4Item7 | your | Function | Det | 1.00 | 0.00 |
| C-Passage4Item8 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage4Item13 | projects | Content | Noun | 1.00 | 0.00 |
| C-Passage5Item3 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage5Item4 | not | Function | Negation | 1.00 | 0.00 |
| C-Passage5Item5 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage5Item6 | about | Function | Prep | 1.00 | 0.00 |
| C-Passage5Item8 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage5Item16 | living | Content | Verb | 1.00 | 0.00 |
| C-Passage5Item19 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage5Item22 | and | Function | Conj | 1.00 | 0.00 |
| C-Passage4Item10 | to | Function | Prep | 0.98 | 0.12 |
| C-Passage5Item1 | you | Function | Pronoun | 0.98 | 0.08 |
| C-Passage5Item2 | on | Function | Prep | 0.98 | -0.08 |
| C-Passage5Item9 | on | Function | Prep | 0.98 | 0.13 |
| C-Passage5Item17 | campus | Content | Noun | 0.98 | 0.18 |
| C-Passage3Item6 | on | Function | Prep | 0.97 | 0.02 |
| C-Passage3Item8 | it | Function | Pronoun | 0.97 | 0.24 |
| C-Passage3Item21 | an | Function | Article | 0.97 | 0.20 |
| C-Passage4Item9 | be | Function | Verb to be | 0.97 | 0.26 |
| C-Passage4Item20 | and | Function | Conj | 0.97 | 0.35 |
| C-Passage5Item10 | because | Function | Conj | 0.97 | 0.23 |
| C-Passage5Item14 | from | Function | Prep | 0.97 | 0.03 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| C-Passage3Item1 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage3Item6 | at | Function | Prep | 1.00 | 0.00 |
| C-Passage4Item6 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage4Item9 | graduate | Content | Noun | 1.00 | 0.00 |
| C-Passage4Item12 | campus | Content | Noun | 1.00 | 0.00 |
| C-Passage4Item20 | you | Function | Pronoun | 1.00 | 0.00 |
| C-Passage2Item2 | is | Function | Verb to be | 0.98 | 0.00 |
| C-Passage2Item9 | to | Function | Prep | 0.98 | -0.04 |
| C-Passage3Item2 | from | Function | Prep | 0.98 | 0.25 |
| C-Passage3Item10 | to | Function | Prep | 0.98 | 0.25 |
| C-Passage4Item14 | opportunities | Content | Noun | 0.98 | -0.06 |
| C-Passage2Item1 | however | Function | Conj | 0.96 | 0.20 |
| C-Passage3Item3 | students | Content | Noun | 0.96 | 0.21 |
| C-Passage3Item18 | is | Function | Verb to be | 0.96 | 0.05 |
| C-Passage4Item13 | you | Function | Pronoun | 0.96 | 0.24 |
| C-Passage5Item3 | because | Function | Conj | 0.96 | 0.17 |
| C-Passage5Item5 | help | Content | Verb | 0.96 | 0.17 |
| C-Passage2Item11 | should | Function | Modal | 0.95 | 0.19 |
| C-Passage3Item5 | campus | Content | Noun | 0.95 | -0.09 |
| C-Passage3Item19 | students | Content | Noun | 0.95 | 0.22 |
| C-Passage3Item23 | have | Content | Verb | 0.95 | 0.37 |
| C-Passage5Item10 | students | Content | Noun | 0.95 | 0.03 |
| C-Passage5Item11 | bring | Content | Verb | 0.95 | 0.38 |
| C-Passage4Item15 | work | Content | Verb | 0.93 | 0.32 |
| C-Passage5Item17 | work | Content | Verb | 0.91 | 0.26 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| C-Passage2Item2 | for | Function | Prep | 1.00 | 0.00 |
| C-Passage2Item3 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage2Item10 | to | Function | Prep | 1.00 | 0.00 |
| C-Passage5Item1 | building | Content | Noun | 1.00 | 0.00 |
| C-Passage5Item2 | is | Function | Verb to be | 1.00 | 0.00 |
| C-Passage5Item5 | university | Content | Noun | 1.00 | 0.00 |
| C-Passage5Item14 | the | Function | Article | 1.00 | 0.00 |
| C-Passage5Item15 | was | Function | Verb to be | 1.00 | 0.00 |
| C-Passage5Item17 | the | Function | Article | 1.00 | 0.00 |
| C-Passage2Item5 | your | Function | Det | 0.98 | 0.43 |
| C-Passage2Item7 | friends | Content | Noun | 0.98 | 0.00 |
| C-Passage2Item8 | your | Function | Det | 0.98 | 0.43 |
| C-Passage2Item19 | to | Function | Prep | 0.98 | 0.43 |
| C-Passage3Item2 | to | Function | Prep | 0.98 | -0.11 |
| C-Passage3Item9 | do | Function | Verb to be | 0.98 | 0.01 |
| C-Passage3Item15 | speaking | Content | Verb | 0.98 | -0.03 |
| C-Passage4Item9 | students | Content | Noun | 0.98 | 0.22 |
| C-Passage5Item7 | is | Function | Verb to be | 0.98 | -0.04 |
| C-Passage5Item16 | in | Function | Prep | 0.98 | 0.14 |
| C-Passage3Item10 | speak | Content | Verb | 0.96 | 0.01 |
| C-Passage3Item18 | cannot | Function | Modal | 0.96 | 0.24 |
| C-Passage5Item10 | the | Function | Article | 0.96 | 0.20 |
| C-Passage5Item13 | campus | Content | Noun | 0.96 | 0.24 |
| C-Passage5Item20 | building | Content | Noun | 0.96 | 0.16 |
| C-Passage5Item23 | university | Content | Noun | 0.96 | 0.41 |

Appendix G. Syntactic Classification and the Values of Item Difficulty and Item Discrimination for Each Cloze-Elide Item in Pilot Data

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 1 | | | | | |
| Passage 3: Job descriptions of a manager (The number of acceptable items is 13.) | | | | | |
| 1 | the | Function | Article | 0.77 | 0.27 |
| 2 | waste | Content | Noun/Verb | 0.77 | 0.67 |
| 3* | when | Function | Det | 0.70 | 0.59 |
| 4 | when | Function | Det | 0.76 | 0.51 |
| 5* | and | Function | Conj | 0.50 | 0.60 |
| 6 | food | Content | Noun | 0.74 | 0.74 |
| 7 | to | Function | Prep | 0.82 | 0.41 |
| 8* | room | Content | Noun | 0.47 | 0.70 |
| 9 | be | Function | Verb to be | 0.92 | 0.33 |
| 10* | service | Content | Noun | 0.47 | 0.43 |
| 11 | size | Content | Noun | 0.74 | 0.65 |
| 12* | restaurant | Content | Noun | 0.56 | 0.76 |
| 13* | of | Function | Prep | 0.62 | 0.59 |
| 14* | considerably | Content | Adv | 0.52 | 0.72 |
| 15* | prevent | Content | Verb | 0.44 | 0.70 |
| 16 | size | Content | Noun | 0.74 | 0.65 |
| 17 | of | Function | Prep | 0.83 | 0.46 |
| 18 | is | Function | Verb to be | 0.91 | 0.40 |
| 19 | open | Content | Verb | 0.24 | 0.50 |
| 20* | employees | Content | Noun | 0.61 | 0.39 |
| 21 | arrive | Content | Verb | 0.76 | 0.75 |
| 22* | special | Content | Adj | 0.47 | 0.69 |
| 23 | for | Function | Prep | 0.94 | 0.48 |
| 24 | to | Function | Prep | 0.71 | 0.50 |
| 25 | the | Function | Article | 0.71 | 0.73 |
| 26* | events | Content | Noun | 0.39 | 0.59 |
| 27* | where | Function | Det | 0.59 | 0.60 |
| 28 | many | Function | Det | 0.80 | 0.55 |
| 29 | more | Function | Det | 0.76 | 0.41 |
| 30* | various | Content | Adj | 0.47 | 0.74 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 2: E-books (The number of acceptable items is 24.) | | | | | |
| 1 | showing | Content | Verb | 0.78 | 0.62 |
| 2* | reading | Content | Verb | 0.64 | 0.76 |
| 3 | with | Function | Prep | 0.84 | 0.29 |
| 4 | on | Function | Prep | 0.84 | 0.47 |
| 5* | an | Function | Article | 0.60 | 0.43 |
| 6* | about | Function | Prep | 0.40 | 0.54 |
| 7 | return | Content | Verb | 0.72 | 0.73 |
| 8* | giant | Content | Noun | 0.41 | 0.57 |
| 9* | to | Function | Prep | 0.66 | 0.69 |
| 10* | material | Content | Noun | 0.53 | 0.75 |
| 11* | return | Content | Verb | 0.60 | 0.75 |
| 12 | topic | Content | Noun | 0.76 | 0.67 |
| 13* | costs | Content | Noun | 0.55 | 0.68 |
| 14* | book | Content | Noun | 0.69 | 0.65 |
| 15 | however | Function | Conj | 0.90 | 0.53 |
| 16* | a | Function | Article | 0.38 | 0.60 |
| 17* | prone | Content | Adj | 0.50 | 0.65 |
| 18* | to | Function | Prep | 0.57 | 0.69 |
| 19* | isolation | Content | Noun | 0.52 | 0.76 |
| 20* | possible | Content | Adj | 0.60 | 0.67 |
| 21* | entire | Content | Adj | 0.41 | 0.52 |
| 22* | electronic | Content | Adj | 0.69 | 0.71 |
| 23* | average | Content | Adj | 0.64 | 0.73 |
| 24* | long | Content | Adj | 0.55 | 0.76 |
| 25* | and | Function | Conj | 0.59 | 0.59 |
| 26* | forget | Content | Verb | 0.57 | 0.64 |
| 27* | to | Function | Prep | 0.57 | 0.55 |
| 28* | reread | Content | Verb | 0.48 | 0.64 |
| 29* | every | Function | Det | 0.53 | 0.54 |
| 30* | unless | Function | Conj | 0.38 | 0.66 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 2 | | | | | |
| Passage 3: Culture (The number of acceptable items is 17.) | | | | | |
| 1* | world | Content | Noun | 0.67 | 0.73 |
| 2 | self | Function | Pronoun | 0.72 | 0.78 |
| 3 | those | Function | Det | 0.72 | 0.45 |
| 4 | of | Function | Prep | 0.74 | 0.38 |
| 5 | a | Function | Article | 0.81 | 0.62 |
| 6 | society | Content | Noun | 0.76 | 0.64 |
| 7 | conversation | Content | Noun | 0.76 | 0.66 |
| 8 | people | Content | Noun | 0.88 | 0.44 |
| 9 | society | Content | Noun | 0.74 | 0.73 |
| 10 | theory | Content | Noun | 0.74 | 0.73 |
| 11 | of | Function | Prep | 0.74 | 0.58 |
| 12* | and | Function | Conj | 0.64 | 0.67 |
| 13 | way | Content | Noun | 0.71 | 0.52 |
| 14* | definition | Content | Noun | 0.67 | 0.56 |
| 15* | society | Content | Noun | 0.57 | 0.61 |
| 16 | to | Function | Prep | 0.78 | 0.62 |
| 17* | in | Function | Prep | 0.53 | 0.63 |
| 18 | how | Function | Pronoun | 0.71 | 0.73 |
| 19* | willingness | Content | Noun | 0.50 | 0.78 |
| 20* | do | Content | Verb | 0.64 | 0.63 |
| 21* | result | Content | Noun | 0.60 | 0.76 |
| 22* | field | Content | Noun | 0.53 | 0.75 |
| 23* | understanding | Content | Noun | 0.67 | 0.71 |
| 24* | willingness | Content | Noun | 0.57 | 0.75 |
| 25* | changed | Content | Verb | 0.52 | 0.53 |
| 26* | continue | Content | Verb | 0.53 | 0.69 |
| 27* | interaction | Content | Noun | 0.67 | 0.80 |
| 28* | way | Content | Noun | 0.64 | 0.76 |
| 29* | looking | Content | Verb | 0.52 | 0.75 |
| 30* | as | Function | Prep | 0.64 | 0.49 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 2: Plagiarism (The number of acceptable items is 22.) | | | | | |
| 1 | citing | Content | Verb | 0.60 | 0.21 |
| 2* | least | Content | Det | 0.78 | 0.33 |
| 3 | if | Function | Conj | 0.96 | 0.38 |
| 4* | on | Function | Prep | 0.67 | 0.41 |
| 5 | to | Function | Prep | 0.75 | 0.44 |
| 6 | between | Content | Adverb | 0.75 | 0.46 |
| 7* | teachers | Content | Noun | 0.49 | 0.47 |
| 8* | forget | Content | Verb | 0.42 | 0.07 |
| 9 | lengthy | Content | Adj | 0.18 | 0.44 |
| 10* | at | Function | Prep | 0.55 | 0.59 |
| 11* | include | Content | Verb | 0.44 | 0.55 |
| 12* | busy | Content | Adj | 0.45 | 0.50 |
| 13* | policies | Content | Noun | 0.35 | 0.65 |
| 14* | a | Function | Article | 0.64 | 0.55 |
| 15* | source | Content | Noun | 0.45 | 0.55 |
| 16* | editor | Content | Noun | 0.31 | 0.46 |
| 17* | you | Function | Pronoun | 0.67 | 0.40 |
| 18* | undocumented | Content | Adj | 0.45 | 0.58 |
| 19 | organization | Content | Noun | 0.20 | 0.43 |
| 20* | distinguish | Content | Verb | 0.38 | 0.48 |
| 21* | be | Function | Verb to be | 0.69 | 0.54 |
| 22 | you | Function | Pronoun | 0.76 | 0.43 |
| 23* | places | Content | Noun | 0.38 | 0.66 |
| 24* | teachers | Content | Noun | 0.35 | 0.47 |
| 25* | the | Function | Article | 0.47 | 0.51 |
| 26* | it | Function | Pronoun | 0.51 | 0.41 |
| 27* | sure | Content | Adj | 0.45 | 0.39 |
| 28* | might | Function | Modal | 0.36 | 0.52 |
| 29 | student | Content | Noun | 0.20 | 0.38 |
| 30* | do | Content | Verb | 0.40 | 0.36 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 3 | | | | | |
| Passage 3: Greek society (The number of acceptable items is 14.) | | | | | |
| 1 | in | Function | Prep | 0.71 | 0.58 |
| 2 | specific | Content | Adj | 0.71 | 0.62 |
| 3 | information | Content | Noun | 0.80 | 0.44 |
| 4 | a | Function | Article | 0.87 | 0.43 |
| 5* | signing | Content | Verb | 0.67 | 0.63 |
| 6 | nothing | Function | Pronoun | 0.80 | 0.36 |
| 7 | it | Function | Pronoun | 0.78 | 0.42 |
| 8 | ask | Content | Verb | 0.82 | 0.29 |
| 9* | to | Function | Prep | 0.64 | 0.43 |
| 10 | are | Function | Verb to be | 0.75 | 0.31 |
| 11 | you | Function | Pronoun | 0.82 | 0.42 |
| 12 | looking | Content | Verb | 0.71 | 0.64 |
| 13* | form | Content | Verb | 0.64 | 0.54 |
| 14* | offer | Content | Verb | 0.33 | 0.50 |
| 15* | more | Function | Det | 0.40 | 0.33 |
| 16* | can | Function | Modal | 0.69 | 0.56 |
| 17* | time | Content | Noun | 0.64 | 0.72 |
| 18* | paid | Content | Verb | 0.51 | 0.65 |
| 19* | for | Function | Prep | 0.36 | 0.47 |
| 20 | strong | Content | Adj | 0.71 | 0.61 |
| 21 | on | Function | Prep | 0.71 | 0.55 |
| 22 | wants | Content | Verb | 0.78 | 0.68 |
| 23* | week | Content | Noun | 0.60 | 0.66 |
| 24 | provide | Content | Verb | 0.22 | 0.42 |
| 25* | dues | Content | Noun | 0.49 | 0.62 |
| 26* | benefit | Content | Noun | 0.49 | 0.68 |
| 27 | is | Function | Verb to be | 0.78 | 0.34 |
| 28* | at | Function | Prep | 0.67 | 0.51 |
| 29 | get | Content | Verb | 0.27 | 0.53 |
| 30* | must | Function | Modal | 0.60 | 0.46 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 2: Writing purpose (The number of acceptable items is 20.) | | | | | |
| 1 | large | Content | Adj | 0.85 | 0.48 |
| 2 | to | Function | Prep | 0.75 | 0.40 |
| 3* | handle | Content | Verb | 0.60 | 0.53 |
| 4 | there | Function | Det | 0.89 | 0.40 |
| 5* | section | Content | Noun | 0.64 | 0.62 |
| 6* | information | Content | Noun | 0.70 | 0.63 |
| 7 | with | Function | Prep | 0.87 | 0.54 |
| 8 | in | Function | Prep | 0.75 | 0.50 |
| 9 | the | Function | Article | 0.87 | 0.55 |
| 10* | available | Content | Adj | 0.64 | 0.71 |
| 11 | down | Content | Adv | 0.72 | 0.71 |
| 12* | report | Content | Verb/noun | 0.40 | 0.41 |
| 13* | particularly | Content | Adv | 0.64 | 0.50 |
| 14* | available | Content | Adj | 0.47 | 0.51 |
| 15* | with | Function | Prep | 0.70 | 0.64 |
| 16* | difficult | Content | Adv | 0.47 | 0.51 |
| 17* | verify | Content | Verb | 0.40 | 0.41 |
| 18* | highlight | Content | Verb/noun | 0.58 | 0.48 |
| 19* | those | Function | Det | 0.70 | 0.57 |
| 20 | information | Content | Noun | 0.74 | 0.67 |
| 21* | report | Content | Verb/noun | 0.32 | 0.55 |
| 22* | reach | Content | Verb | 0.57 | 0.68 |
| 23* | they | Function | Pronoun | 0.55 | 0.52 |
| 24* | written | Content | Verb | 0.34 | 0.48 |
| 25* | other | Function | Det | 0.55 | 0.69 |
| 26* | find | Content | Verb | 0.36 | 0.64 |
| 27 | broad | Content | Adj | 0.26 | 0.40 |
| 28* | can | Function | Modal | 0.47 | 0.43 |
| 29 | well | Content | Adv | 0.28 | 0.60 |
| 30* | a | Function | Article | 0.47 | 0.47 |

(Continued)

| Item number | Key | Word type | Part of speech | Item difficulty | Point biserial correlation |
|---|---|---|---|---|---|
| Form 4 | | | | | |
| Passage 3: Extracurricular activities (The number of acceptable items is 22.) | | | | | |
| 1 | upon | Function | Prep | 0.72 | 0.45 |
| 2 | wish | Content | Prep | 0.87 | 0.46 |
| 3* | of | Function | Prep | 0.60 | 0.47 |
| 4 | it | Function | Pronoun | 0.89 | 0.46 |
| 5* | to | Function | Prep | 0.64 | 0.55 |
| 6* | people | Content | Noun | 0.70 | 0.44 |
| 7* | way | Content | Noun | 0.68 | 0.47 |
| 8* | gift | Content | Noun | 0.55 | 0.61 |
| 9* | life | Content | Noun | 0.55 | 0.63 |
| 10* | create | Content | Verb | 0.30 | 0.38 |
| 11* | you | Function | Pronoun | 0.70 | 0.19 |
| 12* | also | Function | Conj | 0.64 | 0.59 |
| 13 | pastime | Content | Noun | 0.85 | 0.54 |
| 14* | just | Function | Adj/adv | 0.68 | 0.43 |
| 15* | required | Content | Verb | 0.58 | 0.36 |
| 16* | spend | Content | Verb | 0.32 | 0.31 |
| 17* | is | Function | Verb to be | 0.55 | 0.41 |
| 18* | being | Function | Verb to be | 0.51 | 0.45 |
| 19 | who | Function | Det | 0.74 | 0.51 |
| 20* | job | Content | Noun | 0.40 | 0.70 |
| 21* | pleasure | Content | Noun | 0.58 | 0.53 |
| 22* | of | Function | Prep | 0.68 | 0.52 |
| 23 | perhaps | Content | Adv | 0.77 | 0.39 |
| 24* | are | Function | Verb to be | 0.66 | 0.68 |
| 25* | as | Function | Prep | 0.62 | 0.39 |
| 26* | not | Function | Negation | 0.53 | 0.63 |
| 27* | gives | Content | Verb | 0.64 | 0.62 |
| 28* | for | Function | Prep | 0.40 | 0.53 |
| 29 | since | Function | Conj | 0.23 | 0.63 |

VITA

# VITA

Suthathip (Ploy) Thirakunkovit was originally from Bangkok, Thailand. She earned a B.A. in Linguistics from Thammasat University, Thailand in 2003. In 2005, she received her M.A. from Southern Illinois University at Carbondale. Upon her completion, she has been employed as a full-time lecturer at Mahidol University, Thailand.

In August 2011, she started her Ph.D. program in the Second Language Studies at Purdue University. Her research interests broadly focus on test validation and writing assessments. After she completed her study, she will return home and resume her job at Mahidol University.