THE TECHNICAL QUALITIES OF THE ELICITED IMITATION SUBSECTION OF THE ASSESSMENT OF COLLEGE ENGLISH, INTERNATIONAL (ACE-IN)

by

Xiaorui Li

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of English
West Lafayette, Indiana
August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair
College of Liberal Arts

Dr. Harry Denny
College of Liberal Arts

Dr. Tony Silva
College of Liberal Arts

Dr. Anne Traynor
College of Education

Approved by:
Dr. Dorsey Armstrong

Dedicated to Weiran, Beibei and Ollie.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. April Ginther who has provided me with valuable academic guidance and professional training throughout my Ph.D. study. She has inspired me in many different ways. I could not have completed my dissertation without her constructive feedback and continuous support. I also want to thank Dr. April Ginther for offering me the opportunity to work at the Oral English Proficiency Program (OEPP). I greatly benefit from working at the testing office and teaching courses at the OEPP.

Additionally, I would like to thank my committee members, Dr. Anne Traynor, Dr. Tony Silva, and Dr. Harry Denny for their insightful comments on this dissertation. I am grateful to Dr. Anne Traynor, for her valuable suggestions regarding the analysis of instructional sensitivity. My thanks also go to Dr. Xun Yan, for his help at the initial stage of my dissertation project. I would also like to acknowledge the Purdue Language Cultural and Exchange (PLaCE) program for providing me with the dissertation dataset and the course information of the program.

Last, but most importantly, I want to thank my loving parents, Li Xiang and Wang Lanping, who have always believed the importance of education. I could not have made it this far without their unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The present study investigated technical qualities of the elicited imitation (EI) items used by the Assessment of College English – International (ACE-In), a locally developed English language proficiency test used in the undergraduate English Academic Purpose Program at Purdue University. EI is a controversial language assessment tool that has been utilized and examined for decades. The simplicity of the test format and the ease of rating place EI in an advantageous position to be widely implemented in language assessment. On the other hand, EI has received a series of critiques, primarily questioning its validity. To offer insights into the quality of the EI subsection of the ACE-In and to provide guidance for continued test development and revision, the present study examined the measurement qualities of the items by analyzing the pre- and post-test performance of 100 examines on EI. The analyses consist of an item analysis that reports item difficulty, item discrimination, and total score reliability; an examination of pre-post changes in performance that reports a matched pairs t-test and item instructional sensitivity; and an analysis of the correlation patterns between EI scores and TOEFL iBT total and subsection scores.

The results of the item analysis indicated that the current EI task was slightly easy for the intended population, but test items functioned satisfactorily in terms of separating examinees of higher proficiency from those of lower proficiency. The EI task was also found to have high internal consistency across forms. As for the pre-post changes, a significant pair-wise difference was found between the pre- and post-performance after a semester of instruction. However, the results also reported that over half of the items were relatively insensitive to instruction. The last stage of the analysis indicated that while EI scores had a significant positive correlation with TOEFL iBT total scores and speaking subsection scores, EI scores were negatively correlated with TOEFL iBT reading subsection scores.

Findings of the present study provided evidence in favor of the use of EI as a measure of L2 proficiency, especially as a viable alternative to free-response items. EI is also argued to provide additional information regarding examinees' real-time language processing ability that standardized language tests are not intended to measure. Although the EI task used by the ACE-In is generally suitable for the targeted population and testing purposes, it can be further improved if test developers increase the number of difficult items and control the contents and the structures of sentence stimuli.

Examining the technical qualities of test items is fundamental but insufficient to build a validity argument for the test. The present EI test can benefit from test validation studies that exceed item analysis. Future research that focuses on improving item instructional sensitivity is also recommended.

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

International student enrollment enhances diversity at institutions of higher education in North America, but it also poses some challenges for the institutions. One of the challenges that universities often encounter is the English language proficiency of international students. Although most university applicants whose first language is not English provide English proficiency test scores, such as TOEFL or IELTS scores, meeting the required cut score set by universities or programs, especially of near the cut, does not always guarantee the English language proficiency required to comfortably meet academic demands (Haan, 2009; Ginther & Yan, 2018).

Ginther and Yan (2018) reported different correlational patterns between TOEFL iBT subsection scores and the first-year grade point average (GPA) among Chinese students at Purdue University. The negative correlations between GPA and TOEFL iBT reading and listening subsection scores suggested that developing a post-entry language test was beneficial for both the students and the institutions. Although a post-entry language test appears to be similar to standardized English proficiency tests such as TOEFL iBT and IELTS, it has its own unique purpose to the institution and is often developed to address the specific needs of the institutions' student population in relation to language proficiency (Read & Von Randow, 2013).

Furthermore, Purdue University, ranked 3rd for total international student enrollment among the U.S. public institutions in 2017 (Purdue University, 2017), has a large number of international students who are English second language (L2) speakers. The enrollment of international student, especially undergraduate students has increased dramatically in the past decade. Table 1.1 presents the 10-year enrollment trend at Purdue University. The table shows that the total number of international enrollments almost doubled in the past ten years. The surge in international student enrollment, especially at the undergraduate level, brought attention to the language proficiency of incoming international students and created a desire to better understand their needs.

Table 1.1 International Student 10-year Enrollment Trends

	2007	2017	Growth
Undergraduate	2042	4964	+143%
Graduate and Professional	2952	4169	+40%
Total	4994	9133	+83%

Therefore, the development of a post-entry exam to screen incoming international students' language proficiency was undertaken for placement into appropriate language support classes and tracking progress over time. Since the 2014 academic year, the Assessment of College English-International (ACE-In) has been administered at Purdue University to evaluate the language proficiency of international students. The ACE-In test is a locally-developed, internet-based, semi-direct English language test, which consists of three modules. The first module includes two tasks: a cloze-elide task and an elicited imitation task, the second module contains short-answer speaking tasks, and the third module is an essay writing task. The first module was designed to provide information about examinees' ability to process English in real time; the second and third modules were designed to provide speaking and writing performance data.

Once a language test has been developed and implemented, the next important step is to conduct test validation. Examining the validity of the test ensures that the test "employed is valid for the purpose for which it is administered" (p. 89, Henning, 1987). For example, a valid language test should provide useful information about test takers' abilities. This information can then be used to guide instruction. However, most post-entry language tests rarely go through professional validation because they were developed in-house (Knoch & Elder, 2013). As the ACE-In test is a recently developed post-entry language test, test developers and administrators agreed on the necessity and benefits of examining how well the test functions – quality control. What is also essential to build a validity argument for a test is to find evidence that the interpretation and uses of the test scores are appropriate (Kane, 2013). There are various approaches and multiple steps to establish a validity argument for an instrument. As Read (2015) suggests, investigating the technical quality should be the first step. Computing descriptive statistics, examining reliability coefficient and conducting item analysis are insufficient but fundamental to establish a validity argument.

Item analysis is often used as an approach to examine the technical quality of a test which presents the evidence of the reliability of the test and offers further guidance to the interpretation or use of the test scores. It also has been regarded as an important part of test piloting where trialing items often occurs (Fulcher, 2013). When conducting an item analysis, researchers often evaluate item difficulty, item discrimination, and score reliability. Knowing whether the difficulty level is suitable for the target population is important because the misfit between the item difficulty and the test population often leads to unreliable measure even if the items are carefully written (Henning 1987). Item difficulty alone is not sufficient to establish the quality of the item, nor it can provide enough evidence to keep or remove test items. Another index, item discrimination is often computed together with item difficulty. Item discrimination reveals the effectiveness of the items in terms of the response patterns across weak and strong examinees in the ability being tested. For example, both weak and strong examinees may score very low on an item that has poor discrimination. Test developers expect that a high score is an indication of a strong ability, while a low score is an indication of a low ability. Item discrimination influences test score interpretations and should be examined during item analysis. Lastly, instructional sensitivity, although less frequently discussed in item analysis, is another item statistic that demonstrates how well the item discriminates between examinees who have received instruction and those who have not (Crocker & Algina, 1986). This statistic can be used when the same group is pre-tested before and post-tested after instruction. In addition to item analysis, the validity evidence can also be obtained indirectly by correlating the EI test scores on another language test which is valid (Lado, 1965). In other words, if examinees who achieve high scores on an established test such as TOEFL iBT also score high on the examined test, the examined test is then likely to be a valid instrument to measure the same ability.

The present study examines one section of the ACE-In. To be more specific, I focus on the analysis of the performance on the elicited imitation (EI) task, which is the second task of the first module. This study analyzes the technical quality and reliability of the EI task through conducting a thorough item analysis of pre-post EI test scores.

EI is a controversial language assessment tool that has been utilized and examined for decades. In an EI task, examinees listen to a series of sentence stimuli embedded with target language structures and are asked to repeat the sentences as accurately as possible (Larsen-Freeman & Long, 1991). EI items are rated by human raters based on the accuracy of the repetition.

EI is argued to elicit examinees information processing procedures triggered by implicit linguistic knowledge; and thus, reflect the language ability of the examinees.

EI was originally designed for child language or first language (L1) development research. Since the early 1970s, it also has been widely used in the field of second language acquisition (SLA) to measure L2 proficiency. EI has gradually gained popularity as its administration and scoring processes are relatively easy. However, in the late 1970s, EI received a series of critiques, primarily questioning its validity (e.g. Hood & Lightbown, 1978; Hood & Schieffelin, 1978; McDade et al., 1982), which included the criticism that EI was a measure of rote repetition rather than of general language proficiency. In addition, the popularity of EI as an assessment tool declined considerably when the communicative competence approach gained increasing attention from SLA researchers and English as a Second Language (ESL) teachers because EI lacks face validity. Under the influence of communicative competence approach, SLA researchers argued that the development of language proficiency should be evaluated by the learners' ability to communicate as the focus of language study is language use (Savignon,1983). Scholars who favor communicative language testing believe that the closer the tasks resemble what happens in classrooms as in "real world" contexts of use, the better the tasks are. In other words, the language tasks that appear to simulate real-life situation and communication have higher face validity and are argued to be more appropriate to assess the language ability of language learners among some SLA researchers and language educators. EI, on the other hand, which has been often criticized as failing to capture the spontaneous speech and lacking authenticity (e.g. Hood & Lightbown, 1978; Hood & Schieffelin, 1978) has received less and less attention over the years. As a result, fewer studies related to EI were conducted in the 1990s.

However, communicative competence approaches emphasize use over processing. In the past decade, a resurgence of interest in EI as a valid psycholinguistic technique has occurred in the field of SLA. EI has regained attention as many researchers have conducted studies to improve EI task design and to provide a validity argument for EI as a reliable assessment tool to measure general language proficiency (e.g. Erlam, 2006; Hsieh & Lee, 2014; Sarandi, 2015; Van Moere, 2012; Yan et al., 2016). More importantly, EI is capable of offering information regarding language learners' real-time language processing abilities – skills argued to be foundational to communicative competence.

Furthermore, language tasks purely relying on multiple choices such as the TOEFL iBT reading section might be problematic as excessive test preparation and cheating could considerably influence the test results (Ginther & Yan, 2018). Unlike multiple choice items, EI provides performance data which allows researchers to access the evidence of examinees' actual language abilities in real time. As one major section of the current ACE-In test, EI has been providing a large amount of performance data that facilitates the process of tracking students' progress over time.

1.2 Research Questions

The main goals of the present study are two-fold. First of all, by examining the item performance of the EI task, this study contributes to our understanding of the technical quality of EI items and provide suggestions for the current item trial and revising procedure as well as for further EI task design. In addition, the present study serves as a fundamental step in building a validity argument for the EI task used in the ACE-In test. Item analysis allows the test developers to decide whether the test difficulty is suitable for the current test population, whether items are successfully discriminating students at different proficiency levels and whether the instrument is sensitive to instruction. This procedure uncovers whether the present EI task is reliable and is the first step in determining whether the score interpretations are appropriate. In addition to item analysis, the present study also investigates the relationship between the EI test scores and the total and subsection scores of the Test of English as a Foreign Language (TOEFL) iBT test scores. Patterns of correlations between EI test scores and TOEFL iBT subsection/total scores are examined. The analysis reveals the reasons how a post-entry language test adds information to language proficiency test scores are provided by the students as part of their applications.

The present study addresses the following three questions: 1) What are the measurement qualities of the EI items of the ACE-In? 2) Does EI capture a difference between pre- and post-test scores? 3) What are the relationships between EI test scores and TOEFL iBT test scores?

1.3 Organization of the Study

This dissertation consists of five chapters. The present chapter, chapter one, introduces the background and the motivations of the present study. Chapter 2 reviews the relevant literature based on the theoretical discussion of three major aspects: explicit and implicit knowledge dichotomy, information processing models of SLA, and classical test theory. This chapter also provides a thorough review of the operationalization and the measures of implicit knowledge. By comparing EI with other instruments that were developed for similar purposes, chapter 2 argues for the advantages of EI as a measure of implicit knowledge and addresses central considerations in EI task design. Last but not least, chapter 2 reviews the methods used to investigate the technical quality of items. The four item parameters included in the review are item reliability, item difficulty, item discrimination and instructional sensitivity. Chapter 3 presents the methods of the present study including an overview of the current EI task design, the study sample, and the data analyses procedure based on three research questions. Chapter 3 also provides detailed information regarding rater training and rater performance. Chapter 4 reports descriptive statistics and the results of item analysis. Three research questions were addressed respectively based on the results. The last chapter, chapter 5, concludes the study and discusses limitations. This chapter also offers insight into the direction of future EI research and the implications of test development.

CHAPTER 2. LITERATURE REVIEW

The conceptual framework of this study has been influenced by the theoretical discussions of explicit and implicit knowledge, information processing models of SLA, and item analysis procedure. This chapter reviews the relevant literature in these three areas respectively.

2.1 Explicit Knowledge, Implicit Knowledge and the Measures

The discussion of the dichotomy of explicit and implicit knowledge explains why psycholinguistic assessment tools such as EI have been developed in addition to a variety of other language assessment methods to assess learners' linguistic competence. As both explicit knowledge and implicit knowledge, which are psychological constructs, can only be observed indirectly and their presence can only be inferred, it is beneficial to find appropriate measures of each type of knowledge. This section highlights the need to operationalize and the challenges to measurement of the implicit knowledge of second language learners.

2.1.1 Explicit and implicit dichotomy

It is generally understood that explicit knowledge is related to consciousness or awareness, whereas implicit knowledge is associated with intuition or automaticity. However, the scholarly discussion of explicit and implicit knowledge, in fact, often extends the differences introduced above. To gain a comprehensive understanding of these two types of knowledge, I started by reviewing the explicit and implicit paradigm in Second Language Acquisition (SLA). Dörnyei (2009) pointed out that when exploring the explicit/implicit paradigm, scholars had encountered a number of closely related but not identical terms that made their explorations complicated. To avoid confusion, I limited the discussion only to the literature that employed the terms "explicit" and "implicit".

The explicit and implicit paradigm has been mostly used on three levels in SLA literature. These three levels are explicit/implicit memory, explicit/implicit learning, and explicit/implicit knowledge (Dörynei, 2009). The relation among these three levels is commonly understood in the following manners: "explicit knowledge is acquired through explicit learning and is stored in explicit memory; and implicit knowledge is acquired through implicit learning and is stored in implicit memory" (Dörynei, 2009, p.135). These three aspects are certainly connected but needed to be addressed separately in research studies (Dörynei, 2009; Schmidt, 1994). The following discussion focuses on the distinction between explicit/implicit learning and explicit/implicit knowledge. Although these two juxtapositions can be simply understood as that explicit and implicit learning refers to the process of learning, and explicit and implicit knowledge refers to the end products of that learning (Schmidt, 1994), there are more subtle differences and connections that influence the operationalization and the measurement of the constructs.

2.1.2 Explicit learning vs. implicit learning

The distinction between explicit/implicit learning originates in cognitive psychology and is mirrored in SLA (R. Ellis, 2009a). According to Ellis (1994), explicit learning involves conscious and intentional operations through which an "individual makes and tests hypotheses in a search for structure" (p. 1), whereas implicit learning occurs naturally without conscious operation and allows the individual to acquire knowledge from the underlying structure of a complex stimulus environment. Ellis (1994)'s definition of explicit/implicit learning, to some extent, echoes Karshen's learning-acquisition hypothesis in his monitor model. Krashen (1982) argued that an adult second language learner employed two approaches to internalize the rules of a target language: one was through language learning and the other was via language acquisition, and both approaches account for the development of linguistic competence. Learning in Krashen's hypothesis, which is similar to explicit learning often takes place in formal classroom instruction. On the other hand, the acquisition device, which emphasizes the unconscious acquiring nature of language learning corresponds to implicit learning. "Consciousness" appears to be the key to the distinction between explicit and implicit learning; however, this concept has been often vaguely defined and used in the fields of philosophy, psychology neuroscience and cognitive science differently (Dörnyei, 2009). For example, McLaughlin (1978), a psychologist, questioned Krashen's unclear differentiation between conscious and subconscious and argued that Krashen's distinction between learning and acquisition was not supportable. The definition of consciousness became a premise of clear distinctions between implicit and explicit learning. Schmidt (1990, 1994) proposed four types of "consciousness": 1) consciousness as intentionality (e.g. intentional learning versus incidental learning), 2) consciousness as attention (e.g. attended learning versus unattended learning), 3) consciousness as awareness (e.g. explicit learning versus implicit learning), and 4) consciousness as control (e.g. controlled processing versus automatic process). As noted in the above classification, the third type, consciousness as awareness is the most commonly used sense of consciousness in distinguishing explicit learning from implicit learning. Furthermore, Schmidt (1990) operationalized awareness into three levels: 1) perception which "implies mental organization and the ability to create internal representations of external events" (p. 132), 2) noticing which is also known as focal awareness can be defined as "availability for verbal report" (p. 132), 3) understanding which involves the comparison between one's noticing and his/her prior experience. Based on Schmidt's arguments, language learners who have experienced explicit learning are likely to undergo the entire mental process associated with awareness, including perceiving information, noticing gaps and understanding the significance. On the contrary, implicit learning that occurs without consciousness does not require the complete mental process regardless of the learning outcome. However, some researchers believe that there is no completely implicit learning in which awareness is never present. For example, Schmidt (1994) argued that although metalinguistic awareness could be absent, the second level of awareness, noticing, was always involved in learning to some degree.

2.1.3 Explicit knowledge vs. implicit knowledge

As discussed in the previous section, there are two possible procedures, explicit learning and implicit learning for developing L2 proficiency. The end products of these two procedures, which are often referred as explicit knowledge and implicit knowledge, are also argued to be different from each other. Although both types of knowledge account for learners' linguistic knowledge, it is necessary to differentiate implicit knowledge from explicit knowledge. To distinguish explicit and implicit knowledge, Ellis (2009a) argued that the term "linguistic knowledge" should be examined first. The definition of linguistic knowledge primarily comes from two competing views: 1) the innatist's view, which attributes language learning to "a complex and biologically specified language module in the mind of the learner" (R. Ellis, 2009a, p.10); 2) the connectionist's view, which argues input as a primary reason that drives language learning (R. Ellis, 2005, R. Ellis, 2009a). Although these two views are divergent in terms of the driving force behind the development of linguistic knowledge, both views agree that linguistic knowledge comprises both explicit knowledge and implicit knowledge, and the latter is generally believed to play a primary role in developing linguistic competence (R. Ellis, 2005; R. Ellis, 2009a).

Ellis (2009a) also suggested that explicit knowledge and implicit knowledge were neurologically different from each other. Explicit and implicit knowledge were referred to as different outcomes gained from different learning processes. Although there is a controversy about whether these two types of knowledge can be viewed as dichotomy or continuum, more evidence favors the argument that they belong to two independent systems (R. Ellis, 2009a). Explicit knowledge, also known as conscious knowledge, is the outcome of explicit learning which often requires intentional and conscious operation. In contrast, implicit knowledge, also known as unconscious knowledge, results from implicit learning; implicit knowledge is "tacit and inaccessible to conscious introspection" (Rebuschat, 2013, p. 597). In addition, whether the knowledge can be verbalized is also regarded as a distinction between explicit and implicit knowledge. Explicit knowledge is argued something that can be verbalized whereas implicit knowledge cannot. For example, grammar knowledge that L2 learners develop through intentional learning conditions is argued to be explicit knowledge. As a result, L2 learners can articulate grammar rules that they have learned in classroom or from textbooks. On the other hand, learners are often observed to have difficulties in explaining idiomatic language use. Idiomatic language use is argued to be a part of learners' implicit knowledge developed through implicit learning processes that often occurs without conscious operation. However, Ellis (1993) also argues that explicit knowledge is available as a conscious representation, but, nevertheless, learners may still not yet be able to describe the rules using metalanguage. Implicit knowledge is perceived to be less overt and more difficult to discern, and implicit learning processes that occur during language acquisition interest many scholars and researchers in SLA and cognitive psychology (Rebuschat, 2013).

Then, what are some representations of implicit knowledge? Some scholars argue that implicit knowledge consists of formulated language and rule-based language (R. Ellis, 1993). Formulated language refers to lexicalized language units or "lexicalized sentence stems" (Pawley & Syder, 1983, p.192) that are available all times. For example, idiomatic expressions (e.g. *I don't know*) are lexicalized language units. Pawley & Syder (1983) attribute native-like fluency to lexicalized language units because lexicalized language units shorten the information processing time and allow speakers to attend to other tasks during conversation. On the other hand, rule-based language consists of generalized and internalized abstract structures that allow language learners to understand and produce grammatical sentences they have never heard before without exerting much conscious effort (R. Ellis, 1993). Both kinds are tacit, intuitive, and inaccessible to conscious introspection. However, both are available through automatic processing. As implicit knowledge accelerates language processing speed, implicit knowledge is also argued to be crucial to speakers' fluency and automaticity. Although L2 learners are often unable to verbalize the implicit knowledge that they have acquired, the knowledge is evident in their verbal behavior. For example, some L2 learners are well versed in using idiomatic expressions but unable to analyze the grammar structures of the lexicalized language that they use.

2.1.4 Explicit/Implicit knowledge interface

Although most researchers agree that explicit knowledge and implicit knowledge are different and both crucial to linguistic competence, the debate concerning the interface between explicit and implicit knowledge has continued for decades. The interface issue addresses the question whether implicit knowledge and explicit knowledge are transferrable between each other. This question has been answered in three different ways: the non-interface position, the strong interface position and the weak interface position (R. Ellis, 2009a). One of the proponents of non-interface position is Krashen. His Monitor Model argues that subconscious acquisition dominates language performance, and acquisition is the only vital means that contribute to gains in second language competence. Learning is restricted to serving as a monitor of output. Krashen's view on the distinction between learning and acquisition suggests the superiority of acquisition (Dörnyei, 2009). Furthermore, he claimed that learning and acquisition had no overlap with each other (N. N. Ellis, 1994). Krashen states that "learning does not become acquisition" (p. 83, Krashen, 1982). Hence, he held a very clear non-interface position regarding the learning and acquisition. The non-interface position further argues that explicit knowledge and implicit knowledge are independent of each other; explicit and implicit knowledge are distinct and disassociated as they are attained from different mechanisms, stored in the different parts of the brain, and retrieved through different processes.

On the other hand, interface positions include both weak and strong forms. The weak interface position argues that explicit knowledge can be utilized in the learning process, and it plays a significant role in the development of language proficiency. Rod Ellis is among the supporters of the weak interface position (Dörynei, 2009). Unlike Krashen, R. Ellis recognized the importance of both input and output because he believed that output practice could facilitate the automaticity of explicit and implicit knowledge (N. Ellis, 1994). Both non-interface and weak interface positions agree that SLA competence is substantially gained from implicit learning. However, weak interface recognizes the supporting role of explicit knowledge in language learning. A natural follow up is to ask to what extent explicit knowledge plays a role in language learning.

Hulstijn and Graaff (1994) proposed nine hypotheses for empirical research to evaluate how explicit knowledge may facilitate implicit language acquisition. They hypothesized that factors such as linguistic domain, complexity, scope and reliability, and learning types are potential conditions that may influence the interaction between implicit and explicit knowledge. However, Ellis (2005) also pointed out that the weak interface position was mostly based on the theoretical discussions and never gained sufficient empirical evidence in support of the position.

A strong interface position argues that explicit knowledge can be converted to implicit knowledge after a certain amount of practice or through other approaches. In other words, explicit knowledge and implicit knowledge are not dichotomous. Instead, they should be viewed as a continuum. This strong interface position derived from the skill acquisition literature in cognitive psychology, especially Anderson’s Adapted Control of Thought (ACT) theory (Dörynei, 2009; Hulstijn & Graaff, 1994). The ACT theory demonstrates three skill acquisition stages, and Adapted Control of Thought – Rational (ACT-R) specifically addresses the issue concerning the transformation from declarative facts or explicit knowledge to implicit representation. The strong interface position is further advanced and promoted by Sharwood Smith (1981) and DeKeyser (1998, 2007). The strong interface position argues that explicit knowledge can be derived from implicit knowledge, and implicit knowledge can be converted to explicit knowledge through practice (R. Ellis, 2009a). However, it remains to be seen the nature of the “practice” that could transform explicit knowledge into implicit knowledge.

In summary, explicit knowledge and implicit knowledge are primarily viewed as different representations of language learners’ linguistic competence that attained through explicit learning and implicit knowledge. Although debates regarding the relationship between explicit and implicit knowledge persist, scholars have recognized that the vital role that implicit knowledge plays in L2 development. Meanwhile, as implicit knowledge is less overt and observable than explicit knowledge, assessing implicit knowledge tends to be more challenging.

2.1.5 The measurement of implicit knowledge

Different operationalizations or underlying assumptions about constructs result in many problems when interpreting research results (R. Ellis, 2009a). In addition, as explicit and implicit knowledge play different roles in language acquisition and have different representations, it is helpful to operationalize these two psychological constructs respectively. Ellis (2009b) argued that separate measures of explicit and implicit knowledge can be provided by different language tests, but implicit knowledge was more challenging to measure than explicit knowledge. Because learners' linguistic competence consists of both implicit knowledge and explicit knowledge, learners' linguistic competence will be underrepresented if research studies and language assessments only focus on learners' explicit knowledge. Although more studies have targeted learners' explicit knowledge, in recent years, many researchers have indicated the possibility of tapping into implicit knowledge via psycholinguistic instruments (e.g. Erlam, 2006; Van Moere, 2012).

In early studies, explicit knowledge was often operationalized as the learners' explanation of specific linguistic features whereas implicit knowledge was operationalized as the learners' use of these features in oral or written language (R. Ellis, 2009b). For example, Seliger (1979) asked participants to perform a task with target grammar rules and then asked the participants to state the conscious rule related to the task. He then compared the participants' task performance and their abilities to state the rules. The task performance was regarded as the representation of participants' implicit knowledge; the ability to state the rule was regarded as the representation of participants' explicit knowledge. More recently, scholars have developed a more comprehensive view on the operationalization and measurement of implicit and explicit knowledge. Ellis (2005, 2009b) summarized seven criteria for the measures of implicit and explicit knowledge (Table 2.1) and argued that a valid instrument to measure either explicit or implicit knowledge must address these criteria directly.

Table 2.1 Task features of the measures of implicit and explicit knowledge

Criterion	*Implicit knowledge*	*Explicit knowledge*
Degree of awareness	The task requires the learner to respond according to ‘feel’	The task encourages the learner to respond using ‘rules’
Time available	The task is time-pressured	The task is performed without any time pressure
Focus of attention	The task calls for a primary focus on meaning	The task calls for a primary focus on form
Systematicity	The task results in consistent responses	The task results in variable responses
Certainty	The task results in responses that the learner is certain are correct/ incorrect	The task results in responses the correctness/incorrectness of which the learner is uncertain about
Utility of knowledge of metalanguage	The task does not require the learner to use metalinguistic knowledge	The task invites the learner to use metalinguistic knowledge
Learnability	The task favors learners who began learning as children	The task favors learners who have receive form-focused instruction

Source from Ellis (2009b), p.40

According to the above criteria, instruments measuring implicit knowledge should aim at eliciting spontaneous and subconscious performance from learners. To meet the requirements, the instruments measuring implicit knowledge often have two features: examinees are under time-pressure and examinees are asked to focus on meaning. Both features are designed to reduce the possibility that examinees consciously use their linguistic knowledge when completing the task. The instruments that are frequently employed to assess implicit knowledge include EI tasks, oral narrative tests, and timed grammaticality judgment tests (TGJT). These three instruments are employed in different research settings and have their own merits and weaknesses. Among these three instruments, EI is argued to provide a more accurate measure of implicit knowledge (R. R. Ellis, 2009b). Many studies have employed these instruments to measure L2 implicit knowledge, and researchers have been continuously working on the validation of EI (e.g. Bialystok, 1979; Erlam, 2006, Han & Ellis, 1998; Munnich et al., 1994).

2.1.6 Spontaneous speech, grammaticality judgment task and EI

To gain understanding of L2 speakers' implicit knowledge, researchers first need to collect speech data. Spontaneous speech, TGJT and EI tests are all commonly used approaches. The researchers who favor qualitative research methodologies tend to reject data elicited by instruments such as TGJT and EI tests (Larsen-Freeman & Long, 1991). In qualitative research, researchers tend to observe participants and record their spontaneous production without any interference, as using instruments to elicit language output, to some extent, may influence participants' performance. For those researchers, spontaneous or "natural" data is regarded as the most genuine and reliable type of data for investigating participants' linguistic competence.

While the use of spontaneous production data has merits, the constraints involved when collecting spontaneous speech prevents spontaneous speech from being widely used in SLA research. Collecting spontaneous production data is time-consuming, especially when the research involves target linguistic features. Furthermore, participants may create limitations on the collected data since they may avoid certain linguistic features that they have not fully mastered instead fully presenting an expansive picture of their linguistic repertoires which, of course, would include strengths as well as weaknesses (Larsen-Freeman & Long, 1991; Swain et al., 1974). Biased production poses a potential problem for researchers analyzing the spontaneous speech data. In this sense, spontaneous speech is not as "genuine" as many researchers may believe. On the other hand, SLA researchers who embrace quantitative research methodologies prefer using techniques to elicit speech data. Although an instrument is involved during data collection, the instrument allows researchers to collect relevant linguistic information regarding learners' linguistic competence. Among various data elicitation tools, TGJT and EI tests are the two most most-commonly utilized methods and are argued to elicit L2 learners' implicit knowledge (Munnich et al, 1994).

In a TGJT, participants listen to or read a sentence and then evaluate the grammaticality of the sentence. Participants are asked to indicate whether the stimuli sentences are "good" or "acceptable". This method is based on the assumption that in order to make a correct judgment, a test-taker needs to retrieve his/her implicit knowledge (Munnich et al., 1994). The correct judgment, in return, indicates that he/she has acquired the linguistic knowledge that embedded in the sentence stimuli, and he/she has access to that implicit knowledge. If a test-taker fails to make a correct judgment, he/she has possibly not yet acquired the linguistic knowledge associated with the target structure as he/she is unable to access the related implicit knowledge. One of the disadvantages of TGJT tests is that the method does not elicit any speech production data from participants. Rather, TGJT tests only elicits L2 learners' understanding of grammatical structures (Munnich et al., 1994). As TGJT tests do not elicit any performance data, the validity of the task measuring implicit L2 knowledge has been often questioned. The central concern is that as participants do not need to produce much output, they may simply use prescriptive grammar to answer judgement questions instead of retrieving and using their implicit L2 knowledge. Hence, TGJT tests are often considered to be insufficient to represent learners' actual underlying linguistic competence (Munnich et al, 1994).

The third method, EI, another psycholinguistic measurement, appears superior to TGJT in terms of capturing L2 learners' implicit knowledge. In an EI task, examinees listen to a series of sentence stimuli embedded with target language structures and are asked to repeat the sentences verbatim. An apparent merit of this method is that EI can elicit production data from examinees. Researchers can gain speech data from examinees for further analysis. However, this method is not flawless. The lack of authenticity and face validity has been an issue in the literature. What the proponents of face validity argue is that "a test which is to be used in a practical situation should, in addition to having pragmatic or statistical validity, appear practical, pertinent and related to the purpose of the purpose of the test as well" (p. 129, Mosier, 1947). In other words, a test with high face validity should not only be valid but also appear valid. Compared to language tasks developed under the influence of the communicative competence approach that promotes the task authenticity in teaching and assessment, the resemblance between EI tasks and real-life communication is weaker and fails to demonstrate strong face validity. Hood & Lightbown (1978) and Hood & Schieffelin (1978) argued that EI tasks were unable to represent spontaneous speech. Hood & Lightbown (1978) suggested three differences between EI and natural speech that could threaten the validity of EI. First, the central motive for speaking observed in spontaneous speech, the intention to communicate, is altered by EI because the oral production from participants is actually the repetition of sentence sentences; second, sentences presented to examinees lack contextual cues; each sentence stimulus are independent from each other and has no background information; third, sentences stimuli to be imitated include the target structures that are of particular interest to researchers, which do not necessarily reflect the language used in real communication. However, Van Moere (2012) argued that EI tasks did have a degree of authenticity in that repeating the words and phrases from your interlocutor's speech occured in everyday conversational interactions. To be more specific, to prepare or give a response during a conversation, drawing on the language used by conversational partner is often a necessary and important skill of the speaker (Yan, 2015). In addition, Gallimore and Tharp (1981) argued that although EI might not be 'an infallible mirror' (p. 391) of language competence, EI provided researchers with reliable evidence of linguistic competence under standardized conditions. Similarly, Naiman (1974) as well as Swain et al. (1974) suggested that EI could successfully elicit examinees' language production, and that examinees' performance on EI tasks in many respects closely approximates spontaneous speech. Many researchers have countered the argument that EI lacks authenticity and has low face validity. Utilizing EI tasks to collect valid speech production data from examinees to serve as evidence of L2 implicit knowledge and real-life speaking abilities is a more reasonable argument.

In brief, all three commonly used approaches to obtain speech data regarding implicit knowledge have their own advantages and disadvantages. Spontaneous speech requires a substantial amount of time and effort to gauge the comprehensive repertoire of participants. TGJT and EI tasks are both psycholinguistic oral test that can elicit data relevant to the target structures but may influence examinees' performance in problematic ways. An advantage of using EI is that EI elicits controlled production data from participants. Although EI is not infallible, compared to these other approaches, EI is a more reliable and practical way to elicit controlled language production data.

2.2 Information Processing Models and EI as a Measure of L2 Proficiency

As it is impossible to examine the brain's processing mechanisms during language learning, information processing models are the dominant approaches in cognitive psychology to explain this process (Yan, 2015). Information processing models represent how memory may store, retrieve and transform information as well as the process of automatizing and restructuring the information (Huitt, 2003). Although there are a variety of information processing models differ from each other, according to Huitt (2013), they share some fundamental assumptions: 1) the mental system has a limited capacity so that the amount of information can be actively processed is constrained; 2) the encoding, transformation, processing, storage, retrieval and utilization of information is overseen by a control mechanism; 3) there is a two-way flow of information which we gather information through the senses and then use information to construct meaning, 4) human cognition is generically prepared to process and organize information in specific way.

Some theories that are particularly useful to the field of second language learning were derived from the information processing approach (Yan, 2015). The present study draws on Levelt's (1989) processing competence model which is one of the models that built upon information processing models. Levelt's (1989) processing model demonstrates how implicit knowledge may be involved when people process information in real time. This model can be applied to L2 learners processing L2 language input and producing output. Levelt's (1989) processing competence model offers insights into the EI task design as an instrument that is argued to capture the information processing procedures may provide information regarding examinees' implicit knowledge.

2.2.1 Levelt's processing competence model

While the distinction between explicit knowledge and implicit knowledge suggests the need to assess implicit knowledge separately, Levelt's (1989) processing competence model provides theoretical foundation for developing measures of implicit knowledge. Most information processing models offer insight into the processing procedure of a learner from receiving information to producing speech. Although there are different theories regarding information and language processing models, the processing competence model developed by Levelt (1989)

(Figure 2.1) is important to EI because this model has been widely cited and remains influential in SLA research.

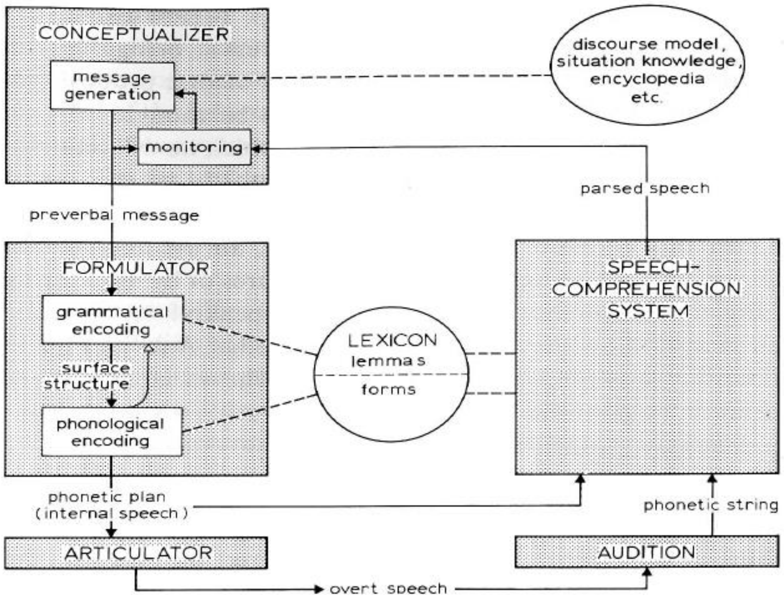

Figure 2.1. A blueprint of the speaker from Levelt (1989)

According to Levelt's (1989) processing competence model, in order to produce a meaningful utterance, a speaker experiences three stages of information processing: conceptualization, formulation, and articulation (Levelt, 1989). In the first stage of information processing, the speaker decodes the incoming message. A conceptual outline of the input is generated, and no lexical or grammatical structure is associated with this outline (Blake, 2006). During the second stage, the speaker translates a conceptual structure into a linguistic structure. To complete this translation, the speaker must first complete a grammatical encoding of the input including recognizing the structure, comprehending the meaning, and inferring intention. The speaker then completes the phonological encoding. During the second stage, an important mechanism associated with grammatical encoding is the speaker's 'lemma information' (p. 11, Levelt, 1989) stored in his/her mental lexicon, an independent module that maintains linguistic information. The speaker needs to retrieve his/her lemma information from the lexicon to facilitate grammatical encoding. The amount of information stocked in lexicon may determine whether the speaker is able to encode information and proceed to the next stage. If the speaker is capable of completing grammatical encoding, he/she then moves to the last stage. In the last stage, the speaker articulates a meaningful utterance and completes his/her performance. The product of the last stage is called 'overt speech' (p. 13, Levelt, 1989). The psycholinguistic processing that involves the use of lemma knowledge during the second stage has a crucial impact on the overt speech produced in the third stage as failing to complete grammatical encoding during the second stage prevents the speaker from moving to the last stage of information processing. Based on this model, the overt speech reflects one's linguistic information stored in mental lexicon and retrieved during the second stage of information processing. In fact, Naiman (1974) suggested a roughly similar model. He suggested that accurate imitation of the syntactic structure required the decoding and the encoding of the structure according to child's productive system.

As for L2 speakers, based on Levelt's (1989) model, to articulate meaningful L2 utterance, L2 speakers also must experience the three stages of information processing: meaning conceptualizing input, conducting grammatical and phonological encoding, and articulating meaningful utterance. The 'lemma knowledge' retrieved from mental lexicon in the second processing stage is regarded as the speaker's L2 implicit knowledge stored as resources (Van Moere, 2012). Because the information stored in the mental lexicon may vary substantially depending on learners' L2 language proficiency, the difference of 'overt speech' produced during the third stage is often observable. The difference of the performance in the last stage, in return, manifests one's L2 linguistic competence. An instrument that is capable of capturing the different performance during last stage of information processing might tap into L2 speakers' implicit knowledge. Based on Levelt's information processing model, a number of psycholinguistic assessment tools have been developed. EI is one of the instruments that is believed to measure L2 implicit knowledge according to Level's model. Yan et al. (2015) offered an account of the information processing elicited by the EI task:

> In the case of EI tasks, when learners hear a sentence, the linguistic input is first stored temporarily as acoustic images in their short-term memory (STM). Based on assumptions of the statistical models of language acquisition, learners need to rely on their implicit linguistic knowledge to decode the linguistic input of the sentence (i.e., comprehension) and retrieve matching lexical and syntactic structures to reconstruct the meaning of the sentence (i.e. repetition). (p. 12)

Based on the above illustration, speakers with different levels of L2 proficiency are expected to produce 'overt speech' differently when completing EI tasks. Advanced learners who have stored sufficient linguistic knowledge in the mental lexicon should be able to successfully encode the input, reconstruct the sentence and produce accurate ‘overt speech’, and their information processing should be more efficient and require less time. In contrast, learners with low proficiency whose linguistic competence fails to afford them with enough structural and lexical support to conduct grammatical encoding may not be able to complete all information processing stages. As a result, imitating sentence stimuli accurately is much more challenging for low proficiency speakers. All in all, the essential step of EI is whether the target structure in the stimuli sentence can be matched and retrieved from one’s mental lexicon, in other words, one’s L2 implicit knowledge.

The above interpretation provides the basis for the argument that EI associated with information processing procedure can tap into one’s L2 implicit knowledge. The underlying assumption is that EI triggers the occurrence and elicits the outcome of the information processing procedure. This assumption implies that the response to the EI task is in fact the product of comprehending the information input during the second stage (i.e. formulation stage of the Levelt’s (1989) model of information processing procedures where implicit linguistic knowledge is involved. However, some researchers found this assumption unwarranted as it was difficult to demonstrate that the processing procedures actually occur, and the information input is fully comprehended.

For example, dating back to the early 1960s, Fraser et al. (1963) made a distinction between imitation and production in their child language development research. In their study, the researchers defined imitation as “the controlling stimuli could be model performances of the utterances under investigation” (p. 123), whereas production was “the controlling stimuli could be reference conditions appropriate to the emission of the utterance” (p. 123). The research results suggested that comprehension preceded production, but imitation preceded comprehension. In other words, whether the speaker comprehends input information plays a role in distinguishing these two concepts. Based on this finding, the researchers further argued that “imitation is a perceptual-motor skill that does not depend on comprehension” (p. 133, Fraser et al., 1963). In this sense, the responses elicited by EI might not be the result of the processing competence model nor the representation of L2 implicit knowledge if examinees repeat without the need to comprehend the meaning of the sentence stimuli.

Whether examinees have to comprehend meaning prior to imitation has always been a central issue for EI research. As for the current EI task utilized in the ACE-In, what is crucial to the researchers is to ensure that the item responses elicited from the task are productions of meaning comprehension not a mere rote imitation. In the past decade, scholars have conducted extensive research to improve EI task design and to provide validity evidence (e.g. Bley-Vroman & Chaudron, 1994; Erlam, 2006; Yan et al., 2015). However, as Hood & Lightbown (1977) have pointed out, a complication is that the definition of "comprehension", "imitation" and "production" varies across research studies. Thus, to avoid confusion, it is worthwhile to clarify the definitions of the terms and present the evidence that examinees' responses involve meaning comprehension in EI research studies. The next chapter provides a literature review on the operationalization of implicit knowledge and the ways EI researchers have addressed the problems with EI task design.

2.2.2 EI as a psycholinguistic tool to measure L2 proficiency

EI is a language task that replicates the information processing procedure and elicits oral production that is argued to represent examinees' L2 implicit linguistic knowledge. Based on the Levelt's (1989) model, in order to repeat each sentence stimulus verbatim, examinees should experience three stages: conceptualizing sentence stimuli, conducting grammatical and phonological encoding, and articulating the repetition. The lemma knowledge retrieved from mental lexicon during the second processing stage manifests examinees' implicit linguistic knowledge (Van Moere, 2012). As the implicit knowledge that is available in mental lexicon may vary substantially among examinees with different L2 proficiency levels, the overt speech produced in the third stage, in other words, the repetition elicited by EI sentence stimuli, is expected to be different. Whether the repetition is complete and accurate largely depends on the implicit knowledge that examinees stored in their mental lexicon. Therefore, the performance of EI is argued to be able to represent examinees' underlying L2 linguistic competence.

Several decades ago, EI was primarily used in first language development studies, and it was later introduced to the field of SLA. In the late 1970s, EI underwent a series of critiques which mainly resided in the validity of the instrument as a measure of L2 proficiency (e.g. Hood & Lightbown, 1978; Hood & Schieffelin, 1978; McDade et al., 1982). In addition to the critique of face validity discussed earlier in this chapter, a more serious criticism involves the construct validity of EI, in other words, what EI actually measures. Some researchers have argued that completing an EI task does not require information retrieval from one's L2 implicit knowledge because the task can be accomplished by solely relying on short-term retention. In response to the critiques, researchers conducted several studies to show the evidence of the reconstructive nature of EI and to further improve EI task design (e.g. Erlam, 2006; Hsieh, & Lee, 2014; Van Moere, 2012; Yan et al., 2016). To determine whether EI elicits simple rote repetition or implicit knowledge retrieval, we need to start with examining the relationship between EI and memory system.

From a traditional view of cognitive psychology, short-term retention is a function solely related to an individual's short-term memory (STM) span and is not influenced by long-term memory (LTM) (Erlam, 2006). Based on this view, if completing an EI task only involves in the use of STM, EI is then unable to elicit examinees' usage of implicit linguistic knowledge. However, more recent studies in cognitive psychology have supplied evidence that EI task responses exceed the use of STM, and these responses do manifest the use of linguistic knowledge stored in lexicon (Erlam, 2006). For example, Yan et al. (2016) recognized the function of STM in EI, and they attributed saving an acoustic image of input during the first stage of conceptualization to STM; however, Yan et al. (2016) argued that the process did not end at STM. Yan (2015) also argued that parroting is more likely to occur with short sentence stimuli because parroting only requires repeating a string of words without decoding the sentence meaning. However, in order to complete longer EI tasks, examinees have to proceed to the next stage where examinees have to retrieve information from the implicit linguistic knowledge and reconstruct the meaning of the sentence.

To further clarify the roles of STM, it is also necessary to reexamine the discussion of the famous "magical number seven" proposed by Miller (1956). Miller (1956) presented evidence for the existence of a clear and definite limit to the accuracy of immediate memory. The limit was called the "span of absolute judgment" (p. 90, Miller, 1965). He further proposed that the maximum of the "span of absolute judgment" of an adult was near the number seven, so the maximum amount of information that one can retain is approximately seven units (+/- 2). What makes this limit even more magical is that the "unit" of information does not refer to any unit of length. According to Miller (1956), a "unit," also referred to as a "chunk", is the largest meaningful unit of information to a speaker. Each chunk may contain multiple bits of information. Based on his assumptions, although the "span of absolute judgment" has a limit of seven chunks, the number of bits of information in each chunk varies among speakers. The reason why sentences with more than seven words may also be stored in STM and be accurately imitated is that each word is only counted as a bit of information, and each chunk may contain more than one bit of information, in other words, more than one word. Therefore, to increase the number of words saved in STM, one needs to retain larger chunks that contain as many bits of information as possible.

A follow-up question to the above argument is that since a chunk is not necessarily a word, then what is a chunk? Pawley & Syder (1983) referred to those chunks as “lexicalized sentence stems” (p. 191) whose grammatical form and lexical content were fixed as a whole unit. Lexicalized sentence stems often comprise more than one word. Based on the explicit and implicit dichotomy discussion, lexicalized sentence stems constitute as implicit knowledge stored in lexicon. Lexicalized sentence stems may be idioms, collocations, as well as “regular form-meaning pairings” (p. 192). To process an EI sentence stimulus, examinees need to retrieve information from lexicon where lexicalized sentence stems are stored and to decompose the entire sentence stimulus into several lexicalized sentence stems. Due to the decomposition and reconstruction, the number of chunks will be reduced so that the complete information can fit in one’s STM. Pawley & Syder (1983) further argued that those stored lexicalized sentence stems available at all times in lexicon might accelerate speakers’ information encoding and free speakers to attend to other tasks in conversation. This argument explains why higher proficiency examinees who possess a large number of lexicalized sentence stems respond to the sentence stimuli more efficiently and accurately whereas low proficiency examinees who have a limited number of lexicalized sentence stems often encounter difficulty reproducing the complete or accurate information (Yan, 2015).

In summary, while STM plays a role in EI, the accuracy of the repetition is largely determined by the implicit linguistic knowledge that examinees mastered. In addition, the implicit knowledge may determine examinee’s processing speed. When sentence stimuli exceed one’s STM span, an examinee must automatically employ his/her implicit knowledge to decompose and reconstruct the information that he/she could retain. Examinees with sufficient linguistic competence are able to use lexicalized sentence stems to maximize the length of each chunk and repeat longer sentences accurately. On the contrary, low proficiency examinees are likely to lose information because without recourse to the lexicalized sentence stems, the chunks they reconstruct are expected to be much shorter, which means that less information can be reproduced. Therefore, a well-designed EI task is argued to be capable of eliciting examinees’ use of implicit linguistic knowledge and providing evidence of L2 speakers’ language proficiency.

2.3 Reconstructive EI Task Design

In the past decade, research has been conducted to refine EI task design so that the possibility of a mere rote repetition in EI task performance is minimized (e.g. Eralm, 2006; Yan et al., 2016). Erlam (2006) suggested that the design of EI tasks could considerably affect the nature of the task as it determined to what extent the task was a measure of L2 implicit knowledge or a measure of his/her ability to imitate given stimuli verbatim. Erlam (2006) distinguished reconstructive EI from rote repetition based on several design features as shown in Table 2.2.

Table 2.2 Features of an elicited imitation test adapted from Erlam (2006)

	Reconstructive: **Likely a measure of implicit knowledge**	**Rote repetition:** **Likely a measure of explicit knowledge**
Design	1. Task design requires primary **focus on meaning**	1. Task design requires primary **focus on form**
	2. **Delay** between presentation of sentence & repetition	2. **No delay** between presentation of stimuli & repetition
Results	1. Ungrammatical sentences corrected	1. Ungrammatical sentences repeated verbatim
	2. No correlation between length of sentence & success	2. Correlation between length of sentence & success

For the purpose of shedding light on the implications of task design, this section elaborates on four widely accepted distinctive features of reconstructive EI tasks: the length of sentence stimuli, grammatical features of sentence stimuli, repetition delay, and focus on meaning.

2.3.1 Length of stimuli sentence

Based on Miller's theory (1956), if the length of sentence stimuli is shorter than one's "span of absolute judgment" which is the maximum amount of information held by STM, the examinee will be able to rely solely on short-term retention and imitate the sentence stimuli verbatim without retrieving information from lexicon. The influence of the length of sentence stimuli has drawn researchers' attention. For example, Bley-Vroman and Chaudron (1994) argued that by manipulating the length of the stimuli sentence, researchers could reduce the possibility of examinees' use of short-term retention in EI. However, what is still debatable is the appropriate length cut-off point for an EI sentence stimulus. The bottom line is that the length of sentence stimuli should at least exceed the "span of absolute judgment". For EI task design, the majority of researchers have chosen to use syllable rather than word count as the unit to measure the length of sentence stimuli. If we treat one syllable as one bit of information, to exceed the "span of absolute judgment", each sentence stimulus needs to contain at least seven or more syllables. According to this standard, some earlier studies using EI may violate the rule of thumb which may influence the interpretation of the research results. For example, Fraser et al.'s (1963) used sentence stimuli that did not quite meet the requirement. Some examples of the stimuli used in that study were "the girl is cooking," "his wagon," and "the deer is running." Apparently, the length of the above stimuli is within the "span of absolute judgment", which considerably enhances the possibility of mere rote repetition. Due to the short sentence stimuli used in the study, some researchers questioned the conclusion that imitation proceeded comprehension because the use of short sentence stimuli may account for participants' performance (e.g. Swain et al., 1974). Therefore, the length of sentence stimuli is an important consideration when designing an EI task. Whether the length of sentence stimuli is reasonable may influence the interpretation of test results and the use of test scores.

While the magical number seven is the traditional view of cognitive psychology with respect to length, SLA researchers have not yet reached agreement on the cutoffs for the length of EI sentence stimuli (Yan et al., 2015). Yan et al.'s (2015) meta-analysis constructed three length bands: short (< 8 syllables), medium (8-15 syllables) and long (>15 syllables). According to the results of the meta-analysis, about one third of the EI studies examined used stimuli sentences of medium length, and many studies used sentence stimuli across all three length bands. For example, Erlam (2006) used sentences between 8 to 18 syllables in length. In addition, some researchers selected sentence stimuli that were much longer than the magical seven; for example, Naiman (1974) and Munnich, et al. (1994) used sentences that contain 15 or more syllables. Similar to short sentence stimuli, extremely long sentence stimuli are also problematics, especially if they are used for child language development research. Lastly, the task difficulty may be manipulated by the length of sentence stimuli; when the length sentence stimuli pass a certain threshold, the test becomes too difficult that may cause a floor effect and contaminate the speech data (Naiman, 1974).

In brief, the length of sentence stimuli may vary depending on specific research purposes and participant pool, but what researchers need to bear in mind is that the length of sentence stimuli affects research results as well as the interpretations and implications of the results. For the sake of future research, it is also necessary for researchers to report the length of the sentence stimuli used in their study as well as their rationale for choosing that particular sentence length.

2.3.2 Grammatical structure of stimuli sentence

In addition to the length of sentence stimuli, the grammatical structures embedded in sentence stimuli are another important consideration for EI task design. As most studies include target grammatical structures, researchers usually have their own rationales when designing EI tasks. Therefore, the grammatical structures embedded in sentence stimuli may vary considerably across EI tasks. Due to the variability of grammar structures, the inauthenticity of sentence stimuli and the usage of ungrammatical sentence stimuli have been and continue to be controversial topics in EI design.

Inauthentic sentence stimuli refer to sentences with target grammatical structures embedded that sound odd and unnatural (Vinther, 2002). Inauthentic sentence stimuli are problematic because inauthentic sentences tend to be very difficult to reconstruct, even for first language (L1) speakers. For example, two EI sentence stimuli used in Fraser et al. (1963) were "the woman gives the bunny the teddy" and "the woman gives the teddy the bunny." This pair of sentence stimuli is extremely similar and does not sound natural. Slobin (1973, cited by Vinther, 2002) criticized the use of inauthentic sentence stimuli in this study as these inauthentic sentence stimuli strain examinees' energy and affect the validity of the research results. Hood & Lightbown (1978) also criticized the EI task design with inauthentic sentence stimuli because inauthentic sentence stimuli failed to represent natural speech, a desirable feature of EI task design. When researchers have target grammatical structures in mind, it is likely that they may make up some sentences that fulfil research purposes but compromise validity. One possible remedial design is to provide examinees along with the context if inauthentic sentence stimuli are to be utilized (Vinther, 2002). Nowadays, researchers have been much more attentive to the authenticity of sentence stimuli. A trend of incorporating an increasing number of conversational words and phrases into EI tasks has been observed in more recent studies (Van Moere, 2012).

Some other researchers have been concerned about the difficulty level of sentence stimuli (e.g. Hamayan et al., 1977; Smith, 1973). They argued that some grammatical structures such as conjunctions and complements might be easier to repeat, while others such as relative clauses and verb auxiliaries were much more difficult. If the range of difficulty level may influence whether learners are able to complete the task successfully, EI may only reflect learners' knowledge of the target grammar and linguistic complexity rather than leaners' language proficiency (Jessop et al., 2007).

Whether sentence stimuli should include ungrammatical sentences is another unsettled debate in the field. Many recent studies provided examinees with both grammatical and ungrammatical sentence stimuli and asked examinees to repeat all sentence stimuli in a grammatical way (Erlam, 2006; Sarandi, 2015). Ungrammatical sentences were included because researchers hypothesized that the ability to identify grammatical and ungrammatical sentences and the ability to fix ungrammatical sentences spontaneously are indications of one's language automaticity (Erlam, 2006). In other words, the ability to correct ungrammatical sentences demonstrates one's advanced storage of L2 implicit knowledge. However, the presence of ungrammatical sentences may introduce confounding variables. For example, examinees may respond differently to different types of grammatical errors. It is unclear whether different types of grammatical errors elicit the same information regarding examinees' linguistic competence. The instructions for EI tasks containing both grammatical and ungrammatical sentences also need careful examination. Researchers may have to decide whether task instructions should inform examinees of the presence of ungrammatical sentences, and whether examinees are expected to correct ungrammatical sentences in their responses. As a subtle change in the instruction may produce different results (Gallimore & Tharp, 1981), researchers should report their research design and clarify the rationale. Unfortunately, many studies only offer very vague descriptions of task design and instructions which may leave their conclusions problematic, even unwarranted.

In terms of selecting grammatical structures, researchers also have to monitor the change of item difficulty level when inserting different grammatical structures. Sentence stimuli that are too difficult or too easy are problematic because the floor and ceiling effect will contaminate the research data. Besides inserting target grammar structures, other possible factors that can influence the difficulty of sentence stimuli such as the presence of ungrammatical sentence stimuli, the complexity of syntactical structure and the number of lexical strings (Yan et al., 2016) may interact with each other and lead to different results. Therefore, researchers have to be cautious about their choice of grammatical structures and offer relevant information about those selected grammatical structures in their research design.

2.3.3 Repetition delay

A third approach that has been widely taken to improve the reconstructive EI task design is repetition delay. In general, there are two major approaches to delay the repetition: one is to add a silent period, usually lasting about three to five seconds before the repetition, and the other is to insert an interruptive task between sentence stimuli and examinees' repetition (Yan et al., 2016). The rationale behind delaying repetition is that not allowing examinees to repeat immediately can reduce the possibility of rote repetition. From the cognitive psychology viewpoint, STM fades rapidly within about five seconds (Vinther, 2002). To retain the input information over a five-second silent period, one has to comprehend the sentence stimulus and utilize his/her linguistic knowledge to reconstruct the sentence. By including a silent period, a reconstructive EI task is designed to elicit performance that cannot depend solely on rote repetition. This assumption is also supported by McDade et al. (1982)'s study. The researchers reported that if the subjects did not comprehend the information, their ability to imitate sentences declined dramatically when the repetition did not occur immediately after the stimulus. However, when the subjects comprehended the information, their repetition was accurate regardless of the presence of the repetition delay.

Similarly, inserting an interruptive task is another approach to delay repetition. The design of interruptive tasks may vary from study to study, but the function of the interruptive tasks is the same: to prevent examines from producing responses that are rote repetitions. For example, Erlam (2006) utilized a beliefs questionnaire as the interruptive task for her study. The beliefs questionnaire was given after the examinees hear the sentence stimuli and before their repetition. The examinees were asked to indicate their attitudes towards the sentence stimuli in the beliefs questionnaire. Although examinees' answers to the questionnaire did not influence the test score, the questionnaire successfully delayed the repetition. Another interesting example of an interruptive task is used by Fraser et al.'s (1963) study. In this study, after hearing each sentence, the participants were asked to pick a picture that correctly illustrated the sentence they heard. This task was primarily designed to check participants' comprehension, but it can also be used as an interruptive task.

The underlying assumption of delayed repetition is that with immediate repetition, examinees may rehearse the sound chain to maintain the acoustic image in STM until the moment of repetition (Yan et al., 2016). The information retained after a silent period or an interruptive task tends to be the information comprehended by examinees rather than mere acoustic images saved in STM (Swain et al., 1974, Vinther, 2002). In other words, with repetition delay, examinees have to experience the four stages of the information processing procedure and utilize their implicit knowledge to support delayed repetition. Therefore, the imitation occurred after the silent period or interruptive task is less likely to be a mere rote repetition. Although repetition delay may improve the reconstructive nature of EI, inserting a silent period or an interruptive task also introduces other concerns. In general, a silent period of three to five seconds is a more accepted approach to delay repetition. Interruptive tasks, on the other hand, have received mixed reviews from researchers because interruptive tasks may introduce confounding variables and interfere with examinees' information processing procedures. For example, Vinther (2002) pointed out that the insertion of an interruptive task might interfere with the initial information processing procedure and accelerate the fading of sentence stimuli. Twenty-one of the 58 studies analyzed by Yan et al. (2016) employed delayed repetition. Repetition delay is a more recent approach to improve EI task design which needs carful design and requires further research.

2.3.4 Focus on meaning

In L1 child development research, young children tend to interpret EI on a message level, which is also the researchers' intention in EI task design (Bley-Vroman & Chaudron, 1994). Focusing on meaning supports the reconstructive nature of EI as examinees have to decode the information and comprehend the meaning before repetition. However, L2 adult speakers' perception of EI tasks might be very different from L1 children. They may attempt to articulate each word verbatim as accurately as possible. Focusing on form contradicts the reconstructive nature of EI because what researchers want to measure is the information processing procedure that triggers the use of implicit linguistic knowledge, and this information processing procedure is based on speakers' attention to meaning. Therefore, to ensure the reconstructive nature of EI, focusing on meaning is another crucial factor (Erlam, 2006).

Examinees may pay attention to both the form and the meaning of sentence stimuli, but what researchers certainly hope is to guide examinees to focus on meaning. This is because the memory capacity of a speaker includes informational and structural aspects when receiving a message. If examinees pay more attention to the structural aspect, or the form of the message, examinees may suffer from a trade-off of the informational aspect (Hulstijn & Hulstijn, 1984). In addition, examinees tend to attend to the aspect of the test that is more relevant to the task goal. In other words, memory of form may be improved when task demands attention to form (Erlam, 2006). Hence, attention to form is likely to either provoke examinees' use of explicit knowledge or use of short-term retention. As EI is designed to elicit close to spontaneous speech and aims to tap into learners' implicit knowledge, to avoid mere rote repetition, EI task instructions should guide examinees to focus on meaning rather than form. If examinees pay more attention to the message and information comprehension, they are more likely to experience the three hypothesized stages of information processing (i.e. conceptualization, formulation, and articulation) which is desirable for the research purpose. If the entire information processing procedure occurs, the repetitions elicited by EI, then, could be more accurately reflect examinees' L2 implicit knowledge.

Compared to the other features of the reconstructive EI, attention to meaning has been less frequently discussed and reviewed in literature. Although this feature was mentioned in a few articles (i.e. Bley-Vroman & Chaudron, 1994; Erlam, 2006), there is relatively limited research investigating the means of guiding examinees to focus on meaning. Erlam (2006) stated that the belief questionnaire was an interruptive task that can also draw examinees' attention to meaning. Although the answers to the belief questionnaire do not count toward the test score, examinees tend to comprehend and focus on the meaning of the sentence stimuli when they are asked to indicate their attitudes. A more recent EI redesign, the "syntactic priming comprehension and production" task (Hsieh & Lee, 2014) incorporated a comprehension task that may also draw examinees' attention to meaning. Unlike many other comprehension tasks, this task occurs before examinees' even hear the sentence stimuli. The examinees are first given several pictures and asked to choose the picture that correspond to the sentence stimuli that later they hear. Since the examinees know that they need to associate pictures to sentence stimuli at the beginning of the EI task, they tend to pay more attention to the meaning of the sentence stimuli. Although the primary purpose of this task is to check examinees comprehension, this task also serves as an approach to implicitly draw participants' attention to meaning. Focusing on meaning is an important feature of a reconstructive EI task. However, more discussions and empirical evidence are needed to bolster this argument. Meanwhile, exploring the methods that could guide the examinees to focus on meaning is also necessary.

At this point, this chapter reviews the relevant literature regarding 1) the construct that we are interested to measure and the operationalization of the construct; 2) the instrument that has developed to measure the construct and the theoretical model that supports the validity of the instrument; and lastly 3) the design features of the instrument that contribute to the validity of the instrument. The next section reviews the method that used to examine the performance of the instrument.

2.4 Item Analysis Procedure

2.4.1 Classical Test Theory

The analysis of the current EI task employs Classical Test Theory (CTT). CTT provides researchers with a general framework for reviewing the development of an instrument (Crocker & Algina, 1986). Knowledge of test theory is beneficial to researchers who develop and evaluate a psychological test to measure a variable of interest, as it provides guidance to evaluate the accuracy and sensitivity of the instruments (Crocker & Algina, 1986). CTT is based on the classical true score model proposed by British psychologist Charles Spearman. The classical true score model argues that any observed test score could be envisioned as the composition of a true score and a random error component (Crocker & Algina, 1986, Fulcher, 2013), which can be expressed as

$$\mathrm{X} = T + E$$

where X = observed test score,

T = the individual's true score,

E = a random error component.

Both T and E are hypothetical components that cannot be directly observed or measured. As language ability is a psychological construct, we can never directly observe or measure the "true" language abilities or in other word, the "true score" of an examinee. The true language abilities can only be estimated based on the true score model which specifies the hypothesized relationship between the true score and the observed test score (Bachman, 1990). What a language test can provide is only an observed test score, and the individual's true score is to be interpreted from the observed test score. Test developers aim to design instruments that can gain an observed test score as close to its true score as possible. Because the discrepancy between the true score and the observed test score cannot be eliminated, to gain a more accurate estimation of one's true ability means to minimize the discrepancy. Based on the equation of the true score model, an observed test score that is very close to one's true score should be obtained under the circumstance where very few random errors are present.

As random errors pose threat to the reliability and validity of the test, knowing what brings random errors and how random errors influence the test use is crucial to test development. A random error is a source of concern when interpreting the observed score and reduce the consistency and the usefulness of the test scores (Crocker & Algina, 1986). A random error could result from "guessing, distractions in the testing situation administration errors, content sampling, scoring errors and fluctuations in the individual examinee's state" (p. 106, Crocker & Algina, 1986). The above sources of random errors should be always taken into consideration when constructing and evaluating an instrument as well as interpreting test scores.

Examining the technical quality of test items allows test developers to consider possible random error sources associated with test items. Classical item analysis, a process of computing and examining any statistical property of examinees' response to a test item is typically used in test construction to select the most useful and reliable items (Crocker & Algina, 1986, Moses, 2017). This statistical procedure is used to better understand the characteristics of individual test item and identify items that fail to function properly (Bachman, 2004). Item parameters that are commonly examined in an item analysis study fall into three categories:

1) Indices that describe the distribution of responses to a single item (i.e., the mean and variance of the item responses)
2) Indices that describe the degree of relationship between response to the item and some criterion of interest
3) Indices that are a function of both item variance and relationship to a criterion.

(p. 311, Crocker and Algina (1986))

The most popular indices and frequently reported of each category are item difficulty and item discrimination. Item analysis is a typical procedure during test construction and revision, especially for dichotomously scored items. The item analysis for EI, which is scored based on a five-point scale is slightly different from dichotomous items in terms of selecting indices but follows the same procedure. Ortega et al. (1999) conducted item analysis to examine four parallel forms of the same EI test in four languages. The study reported that the EI data yielded high reliability and good discrimination. For their study, the item analysis facilitated the item selection procedure and provided validity evidence for paralleled forms of EI in foreign languages other than English.

Based on CTT, the present study examines item difficulty, item discrimination and total score reliability which are the three most useful indices commonly reported in item analysis. In addition, the present study also further investigates the instructional sensitivity of the EI items. Although instructional sensitivity is less frequently reported in item analysis, the information could be helpful to identify items that are sensitive to instruction.

2.4.2 Item difficulty

The first item technical quality examined in the present study is item difficulty. An item difficulty index is applied depending on the type of the items. For polytomous items such as EI, the widely acknowledged item difficulty index is the average item score (Moses, 2017).

When an item can contribute to more item variance, it can provide more useful information to the test (Fulcher, 2013). In most cases, when an item is not too difficult or too easy for the target examinees, the total test score variance will be maximized. However, depending on the purpose of the test, test developers may have different expectations of the range of item difficulty. For example, the appropriate average difficulty and the spread of item difficulties differs for norm- and criterion-referenced tests; criterion-referenced tests are often constructed to be easier than norm-referenced tests (Thorndike & Thorndike, 2009). In addition, although maximizing variability of item difficulty might be an important consideration for some tests, other tests such as diagnostic tests and pre-tests administered prior to instruction can still gain intended information even if everyone gets a perfect score or everyone gets a zero score.

Moreover, the item format may also influence the item difficulty level and how the index will be calculated. For constructed-response items such as short answer questions and essays, examinees often cannot supply answers by guessing. When it comes to selected-response items, guessing becomes a parameter that needs to be taken into consideration.

Lastly, as a test often consists of multiple individual items, some researchers may be interested to know the difficulty level of an average item on the test. It can be gained by $\mu_p = \frac{\sum_i p_i}{k}$, where k is the number of items on the test and p_i is the item difficulty of item i (Crocker & Algina, 1986).

2.4.3 Item discrimination

The second class of item parameters often examined during an item analysis study is item discrimination. Many tests are developed to provide information about individual differences on the construct measured by the test. To serve for this purpose, test items must be capable of effectively distinguishing examinees who are relatively high on the construct of interest from those are relatively low. When examining item discrimination index, the total score on all items is often used as an operational definition of the examinee's relative standing on the construct of interest (Crocker & Algina, 1986). The index assists test developers to identify items that high-scoring examinees have a high possibility of scoring in higher category and low-scoring examinees have a low possibility of scoring in higher category (Crocker & Algina, 1986). In contrast, the items that both high-scoring and low-scoring examinees can answer correctly or fail to answer are less desirable. Furthermore, the items that high-scoring examinees have lower possibility of answering correctly than low-scoring examinees are even more problematic. The items that have the same item difficulty level can show different abilities to differentiate examinees. Item discrimination index helps test developers to discover items that are problematic or unsuitable for certain test purposes.

Several discrimination parameters can serve the purpose. One of the most popular methods of reporting item discrimination, the index of discrimination (D) is applied to dichotomous items. As for polytomous items such as EI, estimating the correlation between score on the single item and score on the set of items is an alternative to examining item discrimination (Thorndike & Thorndike, 2009). There are several correlation coefficients that can be used for this purpose such as point biserial correlation, biserial correlation coefficient, polyserial correlation coefficient, phi coefficient and tetrachoric correlation coefficient. Olsson et al. (1982) offered a guide to select correlation coefficient based on the type of variable.

Table 2.3 Types of correlation coefficients based on the type of variables

	Scale of X		
Scale of Y	Dichotomous	Polychotomous-Ordinal Categories	Continuous-Interval
Dichotomous	Observed: Phi Inferred: Tetrachoric	Observed: No special term Inferred: Polychoric (Special case)	Observed: Point Biserial Inferred: Biserial
Polychotomous-Ordinal Categories		Observed: No special term Inferred: Polychoric	Observed: Point polyserial Inferred: Polyserial
Continuous-Interval			Observed and Inferred: Product Moment

Different correlation coefficients are selected as the discrimination index based on the different types of variables involved in the study. For example, polytomous items often generate ordinal variables, and the total scores of a set of these items generate continuous variables. Under such circumstances, based on Olsson et al (1982), point polyserial or polyserial correlation are suitable coefficients to estimate the correlation between the item score and total score and serves as the item discrimination index.

In addition to variable types, the range of item difficulty and the purpose of the study could also influence the decision. Crocker and Algina (1986) pointed out that when items were of moderate difficulty, there was little difference among discrimination statistics; however, when the items were at extreme difficult levels, discrepancies among different discrimination indices started to become evident. Therefore, if the study involves extremely difficult items, item discrimination index should be selected mostly based on the purpose of the study. If the goal is to select items at one extreme of the difficulty range, and the test developer suspects that future samples will differ in ability from the present item analysis group, the use of biserial correlation coefficient or the polyserial correlation is recommended. On the other hand, if test developers are confident that future samples will be similar in ability to the item analysis sample, and the goal is to select items that will have high internal consistency, the point-biserial or point-polyserial correlation is more desirable (Crocker & Algina, 1986).

Lastly, after obtaining the item discrimination index of each item, test developers can then decide whether revision or deletion is necessary for each item. Although the interpretation of the *D* value may vary, the general guidelines for the interpretation offered by Ebel (1965) are as follows:

1. If $D \geq .40$, the items is functioning quite satisfactorily
2. If $.30 \leq D \leq .39$, little or no revision is required.
3. If $.20 \leq D \leq .29$, the item is marginal and needs revision.
4. If $D \leq .19$, the item should be eliminated or completely revised

While a *D* value, which is generally used for binary items, will not be computed in the present study, the above criterion proposed by Ebel (1965) will be applied to the polyserial correlations which were reported as the item discrimination index for this study.

2.4.4 Total score reliability

The third item technical quality parameter examined in the present study is total score reliability. Scales and instruments that comprise multiple individual measurements such as different items are particularly in need of the evaluation of reliability. This procedure provides information regarding the reliability of a composite in terms of the statistical properties of its internal components (Crocker & Algina, 1986). As one set of EI task often consists of multiple sentence stimuli, examining the reliability of the composite scores can demonstrate whether the test designer was correct in expecting the collection of EI items to yield interpretable statements about individual differences in language proficiency. Two methods are often considered to examine the reliability of a composite: The Spearman Brown Prophecy and Cronbach’s alpha (Crocker & Algina, 1986). To use the Spearman Brown Prophecy, all the components should be parallel tests and the reliability of one of the components is known, while Cronbach’s alpha is preferred when different forms of the tests are not perfectly parallel, but the composite score variance and the covariances among all its components are available. According to Crocker and Algina (1986), because it is often unlikely that all tests in composite are strictly parallel in real testing situations, using the reliability coefficient of precision which is the correlation that would be obtained between two perfectly parallel forms of the test might not be always ideal. Therefore, Cronbach’s alpha, proposed by Lee Cronbach in 1951, is more widely utilized as an index to measure the internal consistency of a test and is the most frequently reported reliability coefficient (Fulcher, 2013). Cronbach’s alpha estimates a lower bound of reliability coefficient that must be smaller than the coefficient of precision. The lower bound of coefficient is also commonly known as coefficient alpha which can be expressed as $\frac{k}{k-1}(1-\frac{\Sigma\sigma_i^2}{\sigma_X^2})$, where k is the number of items, σ_i^2 is the variance of item i, and σ_X^2 is the total test variance on the test. For the above reasons, Cronbach’s alpha is selected as an index for this study to present the reliability of the composite scores of the EI task.

2.4.5 Instructional sensitivity

The last class of item parameter evaluated in this study is instructional sensitivity. Instructional sensitivity or instructional validity refers to the ability of the test to detect differences in student performance before and after instruction. In other words, an instructionally sensitive test should be capable of measuring the differences in instruction received by students (Polikoff, 2010). The measure of tests’ instructional sensitivity is argued to provide empirical support for inferences on instruction based on test scores (Naumann et al., 2017). To be more specific, a measure of the instructional sensitivity of an item provides the information regarding the effectiveness of the item to discriminate between examinees who have received instruction and those who have not.

Instructional sensitivity as an item property was brought into attention when criterion-referenced testing was developed. Language assessment scholars (e.g. Haladyna & Roid, 1981) argued that the traditional item statistics used norm-referenced tests such as item difficulty, item discrimination and item-total correlation did not often meet the demands of criterion-referenced tests (Polikoff, 2010). The poor items identified by traditional techniques may not entirely useless when used in criterion-referenced tests. To address the shortcoming of traditional techniques, psychometricians developed instructional sensitivity that can specifically serve the need to evaluate item quality of criterion-referenced tests. A finding of low or no sensitivity may be due to a poor-quality test that is insensitive to instruction or to poor quality instruction; however, a finding of high sensitivity indicates both a good instruction and also a high-quality test sensitive to that instruction (Polikoff, 2010).

In the past several decades, a variety of approaches to measure instructional sensitivity have been developed based on the different views on instructional sensitivity. The first wave of instructional sensitivity indices was based on item responses as traditional item statistics. To calculate the indices, pre-test and post-test data is often required. Early in the 1980s, another set of instructional sensitivity method that link to information about instructional emphasis in classroom were developed. The techniques include reform-oriented instructional practice surveys, proportional or time measures of curriculum coverage, teacher-rated yeas-or-no topic coverage, a content-by-cognitive demand taxonomy and observations or expert-rated alignment (Polikoff, 2010). The analytical methods used also different based on the data yield from the different techniques. These methods were developed because with the increasing weight and consequences attached to assessments, the instructional sensitivity that purely relied on statistical methods which only touch on instruction indirectly may be insufficient (Baker & Herman, 1983; Linn, 1983). The last type of instructional sensitivity measure is based on expert judgment. This method emphasizes the importance of the judgement of content specialists and data bearing on the content validity of test scores (Popham, 2007). However, Polikoff (2007) argued that this type of measure of instruction sensitivity required careful empirical work that could support the claim that "educators and content experts could discern instructionally sensitive items from those that are not instructionally sensitive" (p. 11).

Although there is a shared understanding of instructional sensitivity among the above methods, the common approach to operationalization is absent (Naumann et al., 2016). For the present study, the instructional sensitivity is operationalized as traditional item statistics which is based on item responses. There are two major categories of indicators widely used in instructional sensitivity studies based on item responses: Pretest-Posttest Difference Index (PPDI), which is the proportion of students passing the item on the post-test minus the proportion of students passing the item on the pre-test, and contingency table sensitivity indices, which use the patterns in item response across pre- and post-tests (Polikoff, 2010). However, previous research on instructional sensitivity has mainly focused on dichotomous items (Naumann et al., 2016; Polikoff, 2010), and proposed instructional sensitivity indices are exclusively used for dichotomous items. On the other hand, the discussion regarding evaluating instructional sensitivity of polytomous items is limited in literature.

Naumann et al. (2016) recently proposed that instructional sensitivity index was appropriate for pre-/post- test designs. Based on Naumann et al. (2016)'s framework, one possible approach to the instructional sensitivity of polytomous items is to conceptualize the instructional sensitivity index as the standardized mean difference. If item difficulty is conceptualized as the mean observed item score at a given time, the standardized mean difference can demonstrate the change of the performance over time. In this case, both effect sizes measures, Cohen's *d* and Hedge's *g* (Hedges, 1987) can be used as the instructional sensitivity index for polytomous items. While both Cohen's *d* and Hedge's *g* are widely used as the effect sizes based on means, Hedge's *g* is the standardized mean difference with an adjustment for sample size. With small sample sizes, especially a sample size smaller than 20, Hedge's *g* is favored.

CHAPTER 3. RESEARCH METHODS

The present study examines the performance of the EI task, and the measurement quality of the items used in the EI task of the ACE-In. The study provides an overview of the characteristics of the EI items through a variety of aspects and analyze the effectiveness of the items used in different testing stages. This chapter introduces the present EI task used in the ACE-In test and the research methodology employed.

3.1 Overview of the ACE-In test and EI task

The Assessment of College English-International (ACE-In), a locally developed, internet-based, semi-direct post-entry English language test, is administered by the Purdue Language Culture Exchange (PlaCE) Program. International undergraduate students who speak English as a second language with TOEFL iBT total score lower than 101 or IELTS score less than 7.5 are required to enroll in the PLaCE program courses ENGL 110 and ENGL 111; all PLaCE students are required to take the ACE-In. The ACE-In is recently developed but has gone through the piloting stage and several rounds of revision. The current ACE-In has three modules: the first module consists of two tasks: a cloze elide task and an elicited imitation task; the second module contains a short-answer speaking task, and the third module has a timed essay writing task. According to the PLaCE program, the ACE-In serves for five major purposes: (1) to identify English language learning needs of international students; (2) to establish sub-skill profiles; (3) to provide a baseline for language instruction; (4) to examine language development; (5) to contribute to language program evaluation. All items within each module are timed to evaluate examinees' real-time abilities to comprehend, speak and write English. As the ACE-In test is an internet-based test, all tests are administered in the university computer labs. Students' responses are uploaded and saved in a web platform and then rated by the PLaCE program instructors and staff members.

The present study only focuses on the evaluation of the EI task, which is the second task of the first module of the ACE-In. Test developers developed four sets of EI task as the ACE-In has four comparable forms designed under the same criteria. Each set of EI task contains 12 items, or in other words, 12 sentence stimuli. The test form is randomly assigned to each examinee during the pre- and post-test.

When completing the EI task, examinees are asked to listen to 12 sentence stimuli, and each sentence is played only once. After each sentence is played, examinees are asked to complete an interruptive task. The computer screen displays two words. One of the two words is mentioned in the sentence stimulus, and examinees are then asked to choose the word mentioned within eight seconds. After completing this short question, examinees are asked to repeat the sentence they hear verbatim within 20 seconds. Examinees' responses are recorded and saved for rating.

The design of the EI task has strictly followed the reconstructive features suggested by literature (See APPENDIX for the task instruction and a sample EI task) to increase the likelihood that the items elicit the use of implicit knowledge and the repetition occurs after meaning comprehension. The task developers have taken into consideration of the following three reconstructive features of EI: the length of the stimuli, repetition delay and focus on meaning. First, the test developers controlled the length of sentence stimuli. The sentence stimuli utilized in the present EI task has two bands of sentence length: the medium length sentence (15 or 16 syllables) and the long sentence (20 or 21 syllables). The length of all sentence stimuli exceeds the amount of information that can be usually held by STM to reduce possible rote repetition. In addition, the question of selecting the mentioned word before repetition is purposely designed by the test developers to delay the repetition (Yan, 2015), so that the possibility of examinees rehearsing the sentence stimuli before repetition is greatly reduced. Lastly, the present EI tasks are time sensitive. Test developers controlled the time allowed for examinees to respond. Examinees are allowed 20 seconds to complete the repetition. When under time pressure, examinees are argued to be more likely to use their implicit knowledge to process the meaning of the sentence stimuli as one entity to complete the task as opposed to using explicit knowledge to analyze the structure of the sentence (Yan, 2015).

Table 3.1 Elicited Imitation: Rating scale

Level	Description	Example Prompt: *Purdue students have free access to printing on campus*.
4 Exact repetition	Repeating the prompt sentence word for word; however, contractions, substitutions for proper noun (e.g., names) are allowed.	*a. Purdue students have free access to printing on campus.* *b. University students have free access to printing on campus.*
3 Appropriate paraphrasing	Did not repeat word for word but paraphrase the prompt sentence in such a way that the response is grammatical and remains the same meaning as the prompt.	*a. Purdue students can print their documents for free on campus.* *b. Purdue students enjoy free printing service on campus.* *c. You can have free access to printing on campus.* *d. All students have free access to printing on campus.*
2 Minor deviation	Repeating the prompt sentence with minor grammatical errors which do not distort meaning; or missing minor information which does not change the main idea of the prompt sentence; or a combination of both.	*a. Purdue students have free access to printers on campus.* *b. Purdue students have free access to printing.* *c. Student has free access to printing on campus.* *d. Students print for free on campus.*
1 Inadequate response	Repeating the prompt sentence with major grammatical errors which distort meaning; or missing substantial information which changes the main idea of the prompt sentence; or a combination of both.	*a. Purdue students print on campus.* *b. Students have free printers on campus.* *c. Purdue students have free access on campus.* *d. Students print free on campus.*
0 Omission	Complete silence, or only one or few words repeated which do not make sense by itself.	*a. Purdue* *b. Students* *c. Free printing* *d. Etc*

In terms of rating, raters score the EI task solely based on the accuracy of the response. Whether examinees successfully select the correct word mentioned in a sentence stimulus does not affect the score that they receive. Table 3.1 presents the five-point scale currently in use. As shown in the table, the current EI task employs a rating scale with the highest score 4, indicating an exact repetition, for which all words in a sentence stimulus are repeated accurately. The lowest score, 0, indicates no repetition, for which no words or a wrong stimulus has been repeated. As shown in the scale descriptions, item scores are only associated with the accuracy of responses; the more accurate and complete the repetition is, the higher the score the student will receive. The total score of the EI task that examinees receive is the sum of the twelve item scores.

3.2 Participants and procedure

The present dataset consists of pre-test item scores and post-test item scores collected in Fall 2016. The pre- and post-test scores of 100 examinees were drawn from the ACE-In Fall 2016 pool which contains the test scores of 460 students. The sampling was based on the following considerations: first, all sample subjects should complete EI pre- and post-test and all test recordings should be available; second, all sample subjects should have TOEFL iBT scores; third, the makeup of sample subjects by native country should simulate the student makeup in ENGL 100 Fall 2016 cohort (China 52%, India 10%, and South Korea 9%); random sampling was applied with in the Chinese, Indian and South Korean subgroups after the first two conditions were made (Cheng, personal communication, March 17, 2020). All examinees were enrolling in ENGL 100 of the PLaCE program at the time taking the pre-/post-test. The pre-test took place at the beginning of the semester (late August or early September) and the post-tests were administered at the end of the same semester (late November or early December).

Except for the seven examinees who did not identify their sex, 28 females and 65 males completed the test. The examinees come from 27 countries; the three largest groups are from China (44), India (13) and South Korea (10). Furthermore, the examinees were enrolled in a wide range of programs with the largest group coming from the College of Engineering (29). As for the English language proficiency of the examinees in the dataset, Table 3.2 displays the distribution of TOEFL iBT total scores and sub-section scores. As shown in the table, TOEFL iBT total scores range from 78 to 100 with an average score of 93.8. The average scores of reading, listening and writing subsections are above 23, while the average scores for the speaking subsection is slightly lower than the other three sections. Their language proficiency range represents the proficiency level of the majority of the English as a second language (ESL) students who are currently studying at Purdue University and have been identified as benefitting from additional language support.

Table 3.2 TOEFL iBT score distribution

Subsection	Mean	Standard Dev.	Min. Score	Max. Score
Reading	23.82	2.85	14	30
Listening	23.96	2.75	17	30
Speaking	22.71	2.12	19	28
Writing	23.41	2.28	17	29
Total	93.80	4.79	78	100

3.3 Scoring and rater performance

Examinees' responses were recorded during the test and saved in an internet platform for rating. Thirteen trained raters who were then the PLaCE program instructors or testing staff members participated in rating. All the raters attended several rounds of training and workshops held by the PLaCE program before rating. The reported EI test item scores used for data analysis were the mean scores of all given scores from the trained raters. While for the pre-test rating, 72 exams were rated by all 13 raters and 28 exams were rated by two of the 13 raters, all 100 post-test exams were randomly assigned to two of the 13 raters. Ten out of 14,316 of score entries were recorded as missing in the dataset because the raters were unable to rate those responses due to technical difficulties. All ten missing values were item scores, and the missing item scores were replaced by using the average item score of that exam.

Prior to data analysis, rater performance was examined to ensure the quality of the test data. Examining rater performance is a crucial step to detect any ill-performed raters and to ensure that rating is not a major source of error. Krippendorff's Alpha coefficient, which measures the agreement among raters was selected as the reliability coefficient. Krippendorff's Alpha is a reliability coefficient developed by Klaus Krippendorff to measure the agreement of observers, coder, judges and raters (Krippendorff, 2011). Although the method was primarily developed and used for content analysis, it has been widely applied in other areas where two or more methods are utilized to generate data (Krippendorff, 2011). Krippendorff's Alpha allows researchers to judge a variety types of data including datasets that are small in size and with missing values. This coefficient can also be applied to different types of scales including nominal, ordinal, interval and ratio. As the EI task rating process involved in 13 raters and most raters did not rate all exams, Krippendorff's Alpha is a suitable inter-rater reliability coefficient for the present study. It is also a robust statistic with datasets containing missing values and is able to provide inter-rater reliability coefficient for the dataset with more than one pair of raters. According to Krippendorff (2011), the general form for α can be expressed in the following mathematical equation:

$$a = 1 - \frac{D_O}{D_e}$$

- D_o = the observed disagreement among values assigned to units of analysis
- D_e = the disagreement one would expect when the coding of units is attribute to chance rather than to the properties of these units

α falls between 0 to 1. When raters agree perfectly, the observed disagreement D_o equals to 0 and α equals to 1. A higher value of α indicates a higher rater reliability. The present study computed the rater reliability α for both pre- and post-test data, and the analysis was performed on the IBM Statistical Package for Social Science (SPSS) Version 24.0.0. Based on the analysis results, the Krippendorff's Alpha for the raters who rated the pre-test exams is 0.8955 and the raters for the post-test is 0.9034. The high α values indicate that the raters of the EI task are likely to agree to the scores given to the same responses. In other words, the rater performance in both pre-test and post-test is reliable and satisfactory.

3.4 Data Analysis

The main data analysis procedure for this study comprises three stages: 1) examining the measurement quality of the EI items, which includes item difficulty, item discrimination, and item reliability, 2) investigating the effectiveness of the task that captures the difference between pre-test and post-test, which performs the analysis of a matched pairs t-test of the total score and the instructional sensitivity of the items, and lastly, 3) exploring the correlation between EI scores and TOEFL iBT subsection scores and total scores. Each stage addresses one research question respectively.

The data analysis procedure started with examining descriptive statistics including mean, standard deviation of item scores and total scores for both pre-test and post-test data, and any possible outliers. This preliminary data analysis procedure provided a general overview of the score distribution and uncovers any possible data recording errors.

To address the first research question, I first examined item difficulty of all EI items. Item difficulty is argued as one of the most important characteristics of an item, as the misfit of item difficulty to person ability leads to low reliability of the test even if the items are carefully written (Henning, 1987). For the present study, the item difficulty level of pre- and post-test exams were analyzed separately. Although the items used in the pre- and post-test are the same, the item difficulty of the same item may change when the item is used for a different test setting and purpose. For example, an item with an appropriate difficulty level for a pre-test may become too easy for a post-test. In addition, four forms of the EI task were also analyzed respectively, so the study examined item difficulty of 48 items (four forms with twelve items per form). The item difficulty index values suggested by literature (i.e. Crocker & Algina, 1986, Fulcher, 2013) are often used for item scored dichotomously such as item scores generated from multiple choices items. When items sored using a binary format (the item score is either 0 or 1), item difficulty index is the proportion of the examinees who answered an item correctly (Fulcher, 2013), and the value ranges between 0 to 1. The greater the item difficulty index value the easier the item, as it shows that a larger proportion of examinees can answer the question correctly. However, the current EI task employs an ordinal scale with five possible item score values (0-4). The traditional computation cannot be applied to EI item scores. Thus, for the present study, the item difficulty index was expressed as the mean score of each item. The possible value of item difficulty ranges from 0 to 4. The present study did not refer to a specific standard value to determine whether the item was too easy or too difficult. Instead, I focused on the distribution of item difficulty of each form, the variability of the item difficulty levels and the appropriateness of the item difficulty level for our particular testing purposes. I also identified comparatively easy items and difficult items to provide insight into the influence of the item features over item difficulty. Based on the analysis results, test developers could gain further understanding of the item features of difficult or easy items as well as the strategies to balance the overall difficulty level.

Second, I examined item discrimination of all EI items by forms. Item discrimination indicates whether an item is capable of discriminating weak and strong examinees in the ability being measured. The two statistics, item difficulty index and item discrimination index are often utilized together to offer suggestions about whether an item should be included, modified or deleted. In terms the index, I selected the item-total score correlation to represent item discrimination. Item-total score correlation is an alternative way to investigate the discriminating power of an item, which does not require dichotomous item scores (Thorndike & Thorndike, 2009). The polyserial correlation coefficient can effectively demonstrate the correlation between the observed item and the total score. It is especially suitable for the present study because the dataset contains one variable (total score) in ratio scale and the other variable (item score) in ordinal scale.

Lastly, I examined the internal consistency of each form for both pre- and post-test exams. As each form consists of twelve items, whether all the items are functioning appropriately together to measure the same construct is the foundation to a reliable instrument. To examine the item reliability, Cronbach's coefficient alpha was selected as the item reliability index. Besides providing reliability evidence, the Cronbach's coefficient alpha also allows test developers to gain a basic understanding of whether all items are functioning alike. For a set of items with lower Cronbach's coefficient alpha, the analysis can also help to identify any items that could negatively influence the overall reliability of the set.

To answer the second research question, how well does the EI task capture the difference in performance before and after instruction, I performed a matched pairs t-test for the pre-post total test scores and examined instructional sensitivity for each test item. Although a matched pairs t-test does not directly yield an effect size measure, it tells the test developers whether there is a significant difference in average item difficulty between pre- and post-test occasions and the magnitude of the difference between pre-test and post-test scores. Prior to performing the matched pairs t-test, I checked the assumption of normality by computing descriptive statistics, examining histograms and QQ plot of the total test scores. In addition to the matched pairs t-test, to further understand the technical quality of the EI items, I also examined instructional sensitivity of the test items. Instructional sensitivity or instructional validity refers to the ability of the test or the item to detect differences in student performance before and after instruction. In other words, an instructionally sensitive test or item should be capable of measuring the differences in instruction received by students (Polikoff, 2010). Instructional sensitivity is often included as a part of instrument evaluation because this measure is argued to provide empirical support for inferences on test scores expected to be influenced by instruction (Naumann et al., 2017). This study utilized the standardized mean difference to represent instructional sensitivity. Based on the literature, Hedge's *g* was selected as the instructional sensitivity index.

The third stage of data analysis focused on the relationships between the EI task and the TOEFL iBT test. Person *r* correlation was computed to estimate the relationship between EI scores and TOEFL total scores as well as the relationship between EI scores and TOEFL subsection scores. EI total scores and TOEFL iBT total scores were expected to have a significant positive correlation because both tests are argued to measure language proficiency. However, I did not expect the correlations between EI scores and all TOEFL iBT subsection scores to be strong. This is because these two tests aimed at measuring different aspects of language proficiency. For example, EI aims to provide information regarding language learners' real-time processing ability, which is not a major targeted skill assessed by the TOEFL iBT reading subsection. Thus, I expected that the correlation between EI scores and TOEFL iBT reading subsection scores would be lower. The correlation patterns between EI total scores and TOEFL iBT total and subsection scores may provide validity evidence for EI as a measure of language proficiency; meanwhile, the correlation patterns may also demonstrate that EI elicits performance data that standardized tests may fail to offer. The additional information provided by EI can be beneficial for both the institutions and the admitted students.

Person *r* correlation was computed to estimate the relationship between TOEFL iBT scores and pre- and post- EI test scores, including both observed correlations (r) and correlations corrected for range restriction. Correlations corrected for range restriction was computed because the current sample represents a restricted sample of the entire TOEFL iBT test taker population as only students admitted to the PLaCE program were included in the analyses. The restricted range is a common problem in predictive validity studies in educational and psychological research. In this case, as the observed correlations tend to underestimate the relationship between the examined variables, corrected correlation is a more accurate representation (Ginther & Yan, 2018). For the present study, the corrected correlation was computed using the Thorndike Case II formula (Thorndike, 1949), which can be expressed in the following mathematical equation:

$$r_{XY} = S_X r_{xy} / ({S_X}^2 {r_{xy}}^2 + {s_x}^2 - {s_x}^2 {r_{xy}}^2)^{1/2}$$

where

- r_{XY} = the observed correlation between X and Y in the restricted sample;
- S_X = the estimated standard deviation of X in the restricted sample;
- s_x = the estimated standard deviation of X in the unrestricted sample; and
- r_{xy} = the estimated corrected correlation between X and Y in the unrestricted sample when only the restricted sample has been used.

The mean and standard deviation of TOEFL iBT scores for undergraduate-level students reported in TOEFL iBT test and data summary (ETS, 2016) were used as the unrestricted population parameters. The statistical analyses were performed on the IBM SPSS Version 24.0.0 and Statistical Analysis System Version 9.4.

CHAPTER 4. RESULTS AND DISCUSSION

This chapter consists of four major sections. Except for the first section which reports the descriptive statistics of the test scores, the other three sections address each research question respectively. The second section presents the results from item analyses, which focus on the measurement qualities, including item difficulty, item discrimination, and total score reliability. The third section reports the relationship between the pre/post-test total scores and item scores. While a pairwise t-test examines if examinees had a significant gain after a semester of instruction, the instructional sensitivity index (Hedge's *g*) is used to evaluate the strength of the effect. Lastly, the fourth section examines the relationships among TOEFL iBT test scores and EI test scores.

4.1 Descriptive Statistics of EI Test Score

The data analysis started with a preliminary examination of the descriptive statistics for pre-test and post-test total scores. The possible total score of the current EI test ranged from 0 to 48 because each item was rated based on a five-point scale (0 to 4), and there were 12 items of each test form. Table 4.1 and Figure 4.1 display the descriptive statistics of total test scores for the pre-test by form.

Table 4.1 Descriptive Statistics of Total Test Scores for Pre-test

Form	N	M	SD	Min	Max	95% CI
F1	26	26.93	5.65	19.00	36.77	[24.65, 29.22]
F2	31	26.81	6.86	16.69	40.15	[24.29, 29.32]
F3	17	28.92	8.33	17.54	42.50	[24.63, 33.20]
F4	26	26.16	7.25	14.23	42.62	[23.23, 29.09]
Overall	100	27.03	6.90	14.23	42.62	

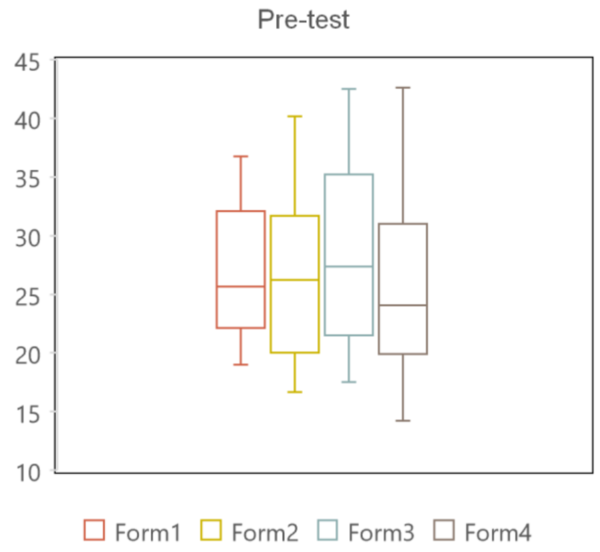

Figure 4.1 Total Test Scores for Pre-test in Boxplot

When comparing pre-test total scores across test forms, I found that the descriptive statistics of Form 1, 2, and 4 were comparable; however, Form 3 had a slightly different profile. First of all, Form 3 had a higher mean total score (M = 28.92) indicating that Form 3 was easier. While the other three forms had a mean total score between 26 to 27, the mean total score of Form 3 was nearly 2 points higher than the other forms. In addition, the wider bandwidth of the 95% confidence interval [24.63, 33.20] and the larger standard deviation (SD = 8.33) of Form 3 indicate that the performance of examinees assigned to Form 3 were more variable. The differences observed from the descriptive statistics of Form 3 may result from the smaller number of examinees (N = 17) assigned to Form 3 during the pre-test. Although no outlier or influential cases were identified in Form 3, due to the small sample size, the impact of the performance of each individual could be more pronounced.

However, despite the small differences between Form 3 and the other three forms in descriptive statistics, the mean total scores of all four test forms fell within the 95% confidence interval of the other three forms. Moreover, as shown in Figure 4.1, the distribution of test scores demonstrates similar patterns across forms in the pre-test. Therefore, for this study, the four forms of the present EI test were considered comparable forms and were included in subsequent analyses.

Table 4.2 Descriptive Statistics of Total Test Scores for Post-test

Form	N	M	SD	Min	Max	95% CI
F1	23	29.13	5.88	18.50	42.00	[26.59, 31.67]
F2	26	29.35	7.20	18.00	43.00	[26.44, 32.25]
F3	28	28.54	6.93	20.50	43.00	[25.85, 31.22]
F4	23	28.93	6.16	20.00	40.00	[26.27, 31.59]
Overall	100	28.98	6.51	18.00	43.00	

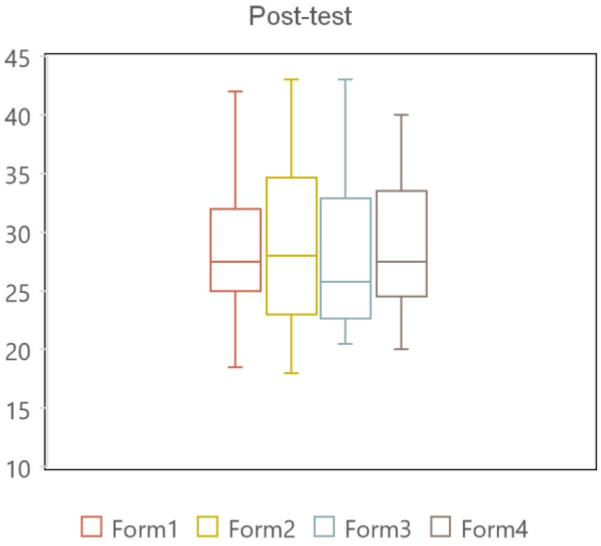

Figure 4.2 Total Test Scores for Post-test in Boxplot

As for the post-test, the number of examinees of each form was more balanced compared to the pre-test. As shown in Table 4.2, the descriptive statistics demonstrates greater similarities across forms in the post-test. The mean total score of each form fell into the 95% confidence interval of the other three forms in the post-test, and the largest difference among the mean total scores across forms was less than 1. Similarly, Figure 4.2 shows larger overlapping areas of the score distribution across forms. The descriptive statistics of the post-test shown in Table 4.2 and Figure 4.2 provide additional evidence that the four EI test forms were adequately comparable and could all be included in subsequent analyses.

Next, the descriptive statistics of the pre- and post-test were compared. The mean total scores increased from 27.03 to 28.98. The increase was anticipated as it was reasonable to find a gain in the students’ language proficiency after instruction. However, regardless of the fact that the overall mean total score improved, the mean total score of Form 3 surprisingly decreased by 0.38 from the pre-test to the post-test. While the mean total score of Form 3 appeared to be in the normal range when compared with the other three forms, the unexpected drop of the post-test mean total score was more likely to be attributed to the unexpectedly higher mean total score of Form 3 in the pre-test.

It is interesting to note that the range of the mean total score among the four test forms narrowed in the post-test from 2.76 to 0.81, which indicates that the examinees’ performance becomes more homogeneous on the post-test. One possible explanation for this observation is that the students scored lower in the pre-test were able to achieve greater gains after a semester of instruction, while the performance of more proficient students remained relatively stable. This observation echoed the fact that the lowest score among all the exams improved 4 points on the post-test whereas the highest score only increased by less than half-point on the post-test. An alternative explanation is that there is a possible ceiling effect in the scores.

4.2 Measurement Quality: Item analysis

To answer the first research question: What are the measurement qualities of the EI items of the ACE-In, this section reported the results from the item analysis. The three measurement qualities evaluated in the item analysis include item difficulty, item discrimination, and total score reliability. Although the four test forms were believed to be comparable based on the descriptive statistics of the total scores, each item was examined individually in the item analysis. Thus, the measurement qualities of all 48 items were reported in this section.

4.2.1 Item difficulty

Item difficulty was conceptualized as the mean observed item score. For the present study, EI items were categorized into three difficulty levels: hard (item difficulty index < 1.5), average (1.5 ≤ item difficulty index ≤ 2.5), and easy (item difficulty index > 2.5).

First of all, the overall item difficulty of each form in the pre- and post-test was computed. As shown in Table 4.3, the item difficulty indices of all forms in both the pre-test and post-test range between 2.18 to 2.45. Based on the above criteria, the overall item difficulty of the present

EI task was found to fall between the average difficulty range. As Fulcher (2013) suggested, for dichotomous items, test developers' ideal was that the overall item difficulty should be around .50. In other words, 50% of the examinees are given full credit. Having an overall difficulty index above 2 on a scale of 0 to 4, the item difficulty of the current EI test was acceptable but slightly easy for the target testing population especially during the post-test.

Table 4.3 Overall item difficulty by Form

Test	F1	F2	F3	F4	Overall
Pre-test	2.25	2.23	2.42	2.18	2.27
Post-test	2.43	2.45	2.38	2.41	2.42

As for the comparison between pre-test and post-test exams, a higher item difficulty index for the post-test was anticipated. After a semester of instruction, the language proficiency of the examinees was expected to improve, so examinees on average should find the items less difficult in the post-test. The results supported the above assumption with one exception. Form 3 had a smaller overall item difficulty index in the post-test, which indicates that Form 3 becomes more difficult in post-test exams. According to Table 4.3, Form 3 was the easiest test form for pre-test exams and became the hardest test form for post-test exams. As discussed previously in the descriptive statistics, a smaller number of examinees assigned to Form 3 in the pre-test might be a possible explanation for this unexpected result.

Table 4.4 The Distribution of Item Difficulty by Form

Item Difficulty	Pre-test				Post-test			
	F1	F2	F3	F4	F1	F2	F3	F4
Hard (0-1.5)	2	0	0	0	1	0	0	0
Average (1.5-2.5)	6	8	9	12	6	6	8	7
Easy (2.5-4)	4	4	3	0	5	6	4	5

Next, I examined item difficulty of each form in the pre- and post-test respectively. As shown in Table 4.4, the result indicates that the majority of the items are either easy or at an average difficulty level in both pre- and post-test exams. What is more noteworthy is the absence of difficult item across forms. In both pre- and post-test exams, only three items were found to be difficult, and all three items were from Form 1 (I10 and I11 in the pre-test, I10 in the post-test). Figure 4.2, 4.3, 4.4 and 4.5 provide visual representations of item difficulty across items of each form by sorting the item difficulty level by a descending order with item difficulty for the post-test next to it. The figures also indicate the lack of difficult items across forms as the majority of the items pass the cut-off score, 1.5, between average and difficult items. Having too many easy or difficult items will lead to the floor and ceiling effect that contaminates the speech data and will reduce the score variance (Fulcher, 2013; Naiman, 1974). The absence of difficult items may be problematic because the test can provide limited information concerning examinees with higher language proficiency. However, one of the uses of the information provided by the ACE-In is to exempt proficient students from the language support course. Without the help of difficult items, the test may not be able to efficiently identify proficient students. To achieve the test purpose, more difficult items should be developed.

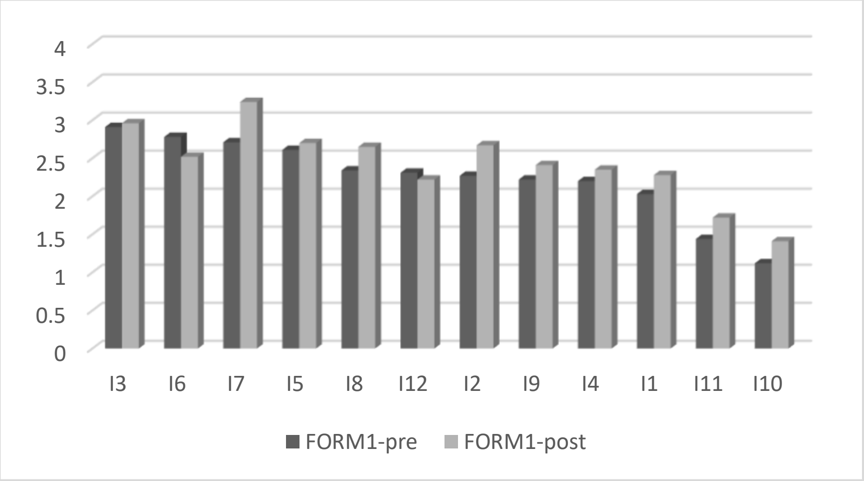

Figure 4.3 Item difficulty of Form 1

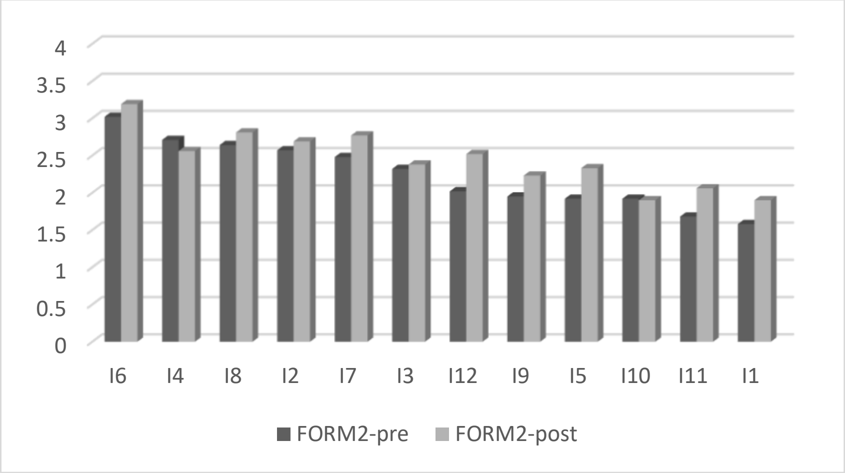

Figure 4.4 Item difficulty of Form 2

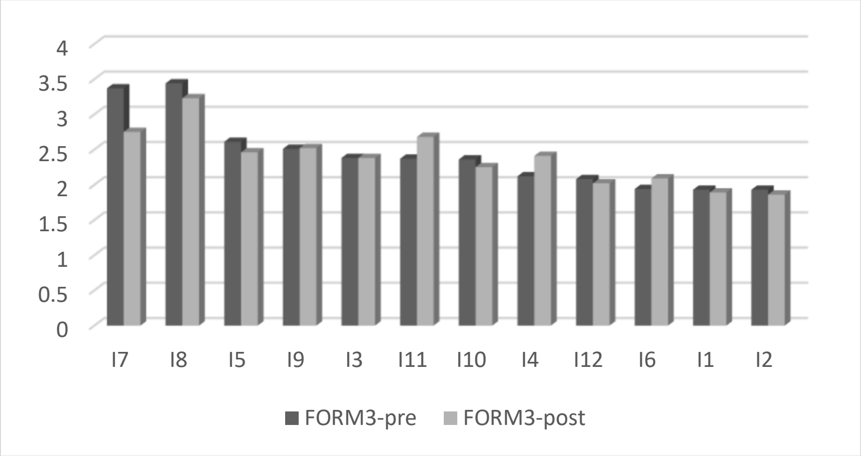

Figure 4.5 Item difficulty of Form 3

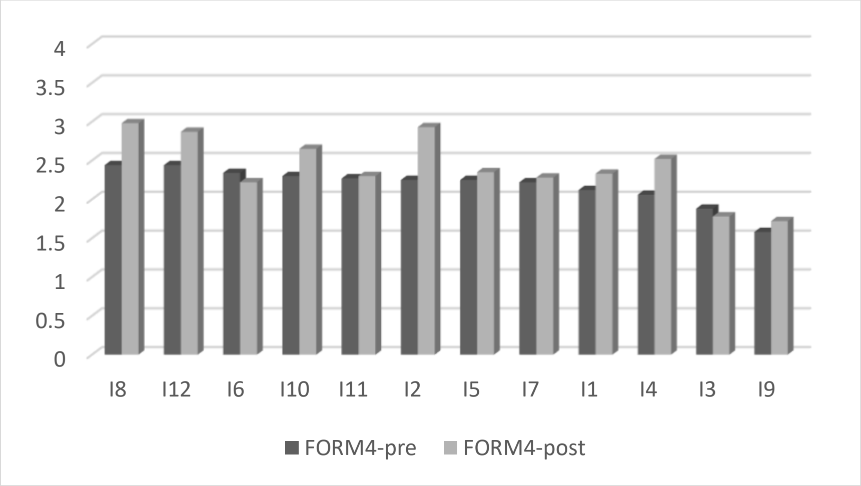

Figure 4.6 Item difficulty of Form 4

Moreover, the results suggest that the item difficulty level of the current test is slightly restricted within each form. Specifically, the cluster of item difficulty appeared to be a problem for Form 4 where all twelve items in the pre-test and more than half of the items in the post-test were found to be of average in difficulty. Although Fulcher (2013) noted that the ideal item difficulty across items was .50, the item difficulty of each item should not be uniformly .50. A set of items that demonstrate variability in item difficulty is more desirable as a wider range of item difficulty can allow the test to discriminate examinees at different proficiency levels. In addition, to maintain the comparability of test forms, it might be also beneficial to improve the diversity of the item difficulty of Form 4. To improve the variability of the item difficulty of Form 4, test developers may consider replacing current items that are of similar item difficulty with either easy or preferably more difficult items.

Additionally, another interesting finding is that there is no apparent pattern between the length of the sentence stimuli and its item difficulty. In other words, a longer sentence stimulus does not necessarily indicate a more difficult item. It is well-acknowledged that the length of sentence stimuli is a central consideration for EI task design, as it may determine whether EI items elicit the use of implicit linguistic knowledge or elicit simple rote repetition. When designing the present EI task, the test developers intentionally included sentence stimuli of different length bands. For all test forms, there were four medium-length sentences (I5, I6, I7, I8) which contained 15 or 16 syllables, and eight long sentences (I1, I2, I3, I4, I9, I10, I11, I12) which contained 20 or 21 syllables. Intuitively, I expected that the longer the sentence stimulus the harder the item would be. However, as the results shown, some long sentence stimuli (e.g. F1I3, F2I2, F2I4, F3I3, F3I9, F4I10, F4I12) turned out to be easier than the medium-length sentence stimuli. Although the only two difficult items (F1I10 and F1I11) were long sentences, the rest of the long sentences were identified as either easy or average.

Sentence length may have an effect on item difficulty under some circumstances; however, what was considered a long sentence by test developers in this case did not guarantee a difficult item. This finding supports Earlm (2006)'s argument that reconstructive EI task design will show no correlation between the length of sentence stimuli and successful performance. One possible explanation is that for a reconstructive EI task, a long sentence can be converted into a "short" sentence if examinees have sufficient implicit knowledge to "chunk" the information. In other words, examinees can shorten the sentence stimuli by reformulating the original sentence into several lexicalized sentence stems by using implicit knowledge. For example, F2I2, *"Before you arrive on campus, you need to make sure that you have a place to live"*, was an easy long sentence stimulus. The lexicalized stem, *to make sure that*, can be processed as one unit instead of four independent syllables. Hence, a long sentence stimulus is shortened and becomes easier to reproduce. In addition, the content of the sentence stimuli may also influence item difficulty. For example, F4I10, "*Meeting people and making friends should be an important part of your college life*", was an easy long sentence stimulus, while F3I6, "*Purdue has the third largest Greek system among all public schools*", was a difficulty medium length sentence. The topic of F4I10 is a more general and commonly discussed topic, while F3 I6, which describes Greek system is more culture specific and may be new to many international students. The unfamiliarity of the topic may account for the high item difficulty index of F3I6, in spite of its shorter sentence length. As the number of syllables of both medium length sentences and long sentences exceeds the limit of short-term memory, the length of the sentence did not appear to be a crucial factor of item difficulty. Therefore, to increase the overall item difficulty of the test, in addition to the sentence length, test developers also need to consider the number of embedded lexicalized sentence stems and the content of the sentence stimuli.

Lastly, it is also worth noting that the first item appears to be a difficult item across forms, especially for pre-test exams (F1I1 = 2.03, F2I1 = 1.58, F3I1 = 1.93, F4I1 = 2.12). The low item difficulty index of I1 may result from the fact that the examinees are not well-prepared at the beginning of the test or not familiar with the test format. As they gained more familiarity with the test, their performance gradually improved. For the current EI test, the first and last four items are long sentences, and the four sentences in the middle are medium length sentences. In this case, test developers may consider switching a medium length sentence with low item difficulty to the beginning of the test. A shorter and easier item may help examinees quickly familiarize themselves with the test format. Alternatively, starting with a practice item that is unscored might be useful.

In sum, the overall item difficulty of the present EI task is at an average level. Although the current test is a slightly easy test, the difficulty is acceptable for the target test population and test purposes. However, there are a few future steps that test developers can take to improve the test. First and foremost, a further investigation into Form 3 is necessary. The item difficulty level for Form 3 has decreased in post-test exams, which means that the items in Form 3 became more difficult for the examinees after they had a semester of instruction. To investigate whether the unexpected statistics resulted from either the imbalance of the sample or the test content (i.e. the design of the sentence stimuli) of Form 3, another round of analysis with a larger or more balanced sampling should be conducted. Secondly, although the overall difficulty level is acceptable, the combination of items with different levels of item difficulty could be improved. More importantly, including more difficult items into each form can be beneficial. Test developers may also consider replacing some average difficulty items from Form 4 as Form 4 shows a restricted range of item difficulty. Improving the variability of item difficulty of Form 4 can also maintain the comparability of all test forms.

4.2.2 Item discrimination

Table 4.5 reports the item discrimination of all items by form. The discrimination index of the pre-test was sorted in descending order. For the pre-test, the item discrimination index ranged from .90 to -.04. All but four items had an item discrimination index above .40, indicating that the majority of the items were functioning satisfactorily in terms of distinguishing proficient examinees from examinees with lower language proficiency.

Table 4.5 Item discrimination index by form

Test		Item											
		1	2	3	4	5	6	7	8	9	10	11	12
Pre-Test	F1	.58	**.35**	.56	.52	.67	.70	**.39**	.83	.67	.48	***-.04***	.66
	F2	.86	.84	.73	.74	.58	.46	.71	.73	.71	.65	**.39**	.72
	F3	.82	.87	.53	.71	.85	.73	.81	.90	.79	.75	.80	.84
	F4	.72	.77	.75	.53	.62	.75	.77	.73	.80	.74	.82	.58
Post-Test	F1	.45	.54	.54	.66	.68	.49	.74	.67	.78	.76	**.22**	.68
	F2	.68	.51	.56	.73	.70	.50	.64	.90	.66	.46	.60	.75
	F3	.66	.72	.73	.59	.84	.66	.77	.83	.84	.61	.75	.66
	F4	.59	.75	.67	.62	.85	.61	.61	.84	.68	66	.79	**.31**

Among the four items that had low discrimination indices, three of them (F1I11, F1I7, F2I11) fell between .30 and .39, suggesting that very minor revision was needed. The only problematic item that required further investigation or deletion was F1I11 whose item discrimination was -.04. A negative index indicates that examinees with low proficiency tend to score high on this item while more proficient examinees tend to score low. Although F1I11 is a long sentence with a complex syntactic structure that may increase item difficulty (pre-test item difficulty = 1.44, post-test item difficulty = 1.72), a difficult item does not necessarily yield a low discrimination index. A possible explanation for its low discrimination may be related to the uncommon syntactic structure of the sentence stimuli. In fact, EI has been criticized by using inauthentic grammatical structures that often strain examinees' energy and affect the validity of the test (Lightbown, 1978; Slobin, 1973). The problematic item, "*Taking a part-time job on campus has been shown to help students succeed in college*", contains a raising verb, "shown", which usually appears with a syntactic argument that is the semantic argument of an embedded predicate. Raising verbs such as *appear*, *seem*, and *show* often take part in an *it*-extraposition structure. However, the verb "shown" was not used as a raising-to-subject verb in F1I11, which may interrupt examinees' language processing regardless of language proficiency. Another possible explanation is associated with the topic of the sentence. The finding about the relationship between working while studying may be contrary to examinee's expectations as students may believe that taking part-time jobs could negatively influence academic performance. Both high and low proficiency examinees failed to reproduce this sentence stimuli. As a consequence, this item may be unable to discriminate examinees with high proficiency from the examines with low proficiency. A possible solution is to rewrite the sentence and change the sentence structure to *it*-extraposition. For example, the sentence can be written as "*It is shown that taking a part-time job on campus helps students succeed in college*". It is worthwhile to test this revised sentence and to investigate if changing the syntactic structure would improve the item's discrimination.

As for the post-test, the item discrimination index of two-thirds of the items slightly decreased, but the majority of the items still functioned satisfactorily in discriminating examinees. In fact, forty-six out of 48 items maintain an item discrimination index above .40. In other words, most items functioned satisfactorily in both the pre-test and the post-test. It is also noteworthy that the discrimination index of the three items that need minor revisions improved in the post-test; the item discrimination indices of all three items increased above .40 in the post-test. As for the problematic item, F1I11, although the item discrimination index increased from -.04 to .22, major revision was still necessary if this item were to be used in the post-test.

While the majority of items maintain their discrimination indices above .40 in the post-test, the discrimination index of a few items experienced a dramatic decline. For example, the item discrimination index of F2I2 dropped from .85 to .50. However, since its item discrimination index maintains above .40, deletion or revision is unnecessary. While the majority of the item still functioned well in the post-test, the decrease of the item discrimination index of F4I12 appears to be problematic. The item discrimination index of F4I12 dropped from .58 to .31, suggesting that this item needs a minor revision when used for post-test exams. F4I12, "*You should talk to your advisor if you are not sure what courses to take next semester*", is an easy long sentence stimulus. The topic of this sentence may account for the decrease; the examinees became more familiar with the topic in relation to selecting courses after a semester of studying in college regardless of their proficiency levels. The familiarity of the topic may facilitate the information processing and allow better performance. As an easy item in the post-test, F4I12 is not as effective as it is used in the pre-test.

Overall, the analysis of item discrimination results report that the majority of the items discriminate satisfactorily for the current testing population in both pre- and post-test. There were no obvious patterns between the sentence length and the item discrimination index. Three items, F1I7, F2I11, and F4I12, need minor revisions. The most problematic item, F1I11 should be re-written or removed from the current test as its discrimination indices for both the pre-test and the post-test were very low. Additionally, the results suggest that the item discrimination index may change when the same item is used in different testing stages. Some items' discrimination may improve while others become less effective in discriminating examinees when used in post-test exams. Therefore, item analyses should be conducted for each testing condition, even if the same set of items functioned well previously.

4.2.3 Item reliability

Cronbach's coefficient alpha (α) was computed to represent item reliability, which provided evidence for the internal consistency of the twelve items used in each test form. The result from item reliability allows test developers to identify whether a set of items proposed to measure the same construct produces similar scores. Table 4.6 reports that almost all test forms except Form 1 have total score reliability above .85 in both the pre- and post-test, indicating very high total score reliability of the present EI test. A large reliability index value shows that all 12 items of each test form are functioning together as a group and producing similar test scores about the examinees' language proficiency. Although the previous analyses reported some unexpected results regarding Form 3, Form 3 had excellent internal consistency with α greater than .90 in both test administrations.

Table 4.6 Item reliability index by form

	Form 1	Form 2	Form 3	Form 4
Pre-test	.77	.89	.93	.91
Post-test	.81	.85	.90	.86

When compared to the other three forms, Form 1 reported slightly lower α in both the pre-test (α_{pre} = .77) and the post-test (α_{post} = .81). Although the values were adequate, considering the high internal consistency of the other forms, I conducted a further investigation of the items in Form 1. To identify possible problematic items, I examined Form 1's α when each item was deleted from the set. As shown in Table 4.7, Form 1's α improved if I11 was removed for both pre-test (α_{pre_del} = .80) and post-test (α_{post_del} = .82). This result indicates that I11, "*Taking a part-time job on campus has been shown to help students succeed in college*", is potentially a problematic item that fails to measure the same construct that all other items measure. It was not surprising to note that I11 was not highly correlated with all other items in the same form, given the fact that I11 had inadequate item discrimination for both the pre- and post-test. The incapability of distinguishing proficient examinees from low proficiency examinees might result from the fact that this item was not measuring the same construct as other items in the same form.

Table 4.7 Cronbach Coefficient Alpha with deleted items for Form 1

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12
Pre-test	.73	.78	.76	.75	.75	.73	.77	.72	.73	.75	**.80**	.73
Post-test	.80	.79	.80	.78	.81	.77	.81	.79	.77	.77	**.82**	.78

In sum, all test forms of the current EI test have high internal consistency, and the performance remains relatively stable when used in different testing settings. If test developers rewrite I11 of Form1 or delete it from the current item bank, the total score reliability of Form1 could be further improved.

4.3 Pre- and Post- Test Analysis

This section focuses on the second research question: Does EI capture a difference between pre- and post-test scores. The analysis was conducted in two stages: the analysis of a matched pairs t-test, which uncovered whether there was a significant difference between pre- and post-test total scores, and the analysis of the instructional sensitivity of test items, which reported how well each test item captured the changes of the performance after the instruction.

4.3.1 The instruction between the pre- and post-test

Between the pre- and post-test, all examinees were enrolled in a 3-credit course, ENGL 100: American Language and Culture for International Students I in the 2016 Fall semester. ENGL 100 is the first course of the two-course sequence offered by Purdue Language Cultural Exchange (PLaCE) program. This course is designed to benefit international students’ development of the academic, linguistic, and cultural competences needed fully engage in available academic opportunities that Purdue University offers. According to the ENGL 110 course syllabus, by taking this course, the students will be able to speak and read English more fluently, communicate in English with increased clarity, and develop and apply effective process for cross-cultural comparison and reflection. The instructors of this course guide students to practice advanced reading, writing, listening and speaking skills while exploring the local culture. Most students who take ENGL 110 during the fall semester will complete the sequence by taking ENGL 111 in the spring semester.

For ENGL 110, each class has a maximum of 16 students, and class attendance is mandatory (Allen, personal communication, February 21, 2020). The development of integrative language skills is strongly emphasized and reflected by the ENGL 110 curriculum. Unlike many traditional language support programs where each language skill is often taught and practiced individually, the course projects and activities of ENGL 110 require the use of all four language skills. The students are encouraged to employ integrative language skills to complete class activities and homework. In addition to the four core language skills, the PLaCE program is designed to promote students' intercultural knowledge and competence as well as thinking and learning skills and strategies. Therefore, the ENGL 110 class content covers a variety of topics tightly associated with college life in the U.S.

4.3.2 Matched pairs t-test

The analysis of pre-post changes in total scores on the EI section started from a matched pairs t-test. The assumptions of independence and normality were examined before performing the t-test. Figure 4.7, the QQ plot of the sample shows that the underlying testing population follows a normal distribution, so the current sample meets the assumption of normality. In addition, each examinee's score was independent of one another, and no potential outliers were observed in the dataset. As the present sample met both assumptions, normality and independence, a matched pairs t-test was conducted to investigate if there was a significant difference between pre- and post-test scores.

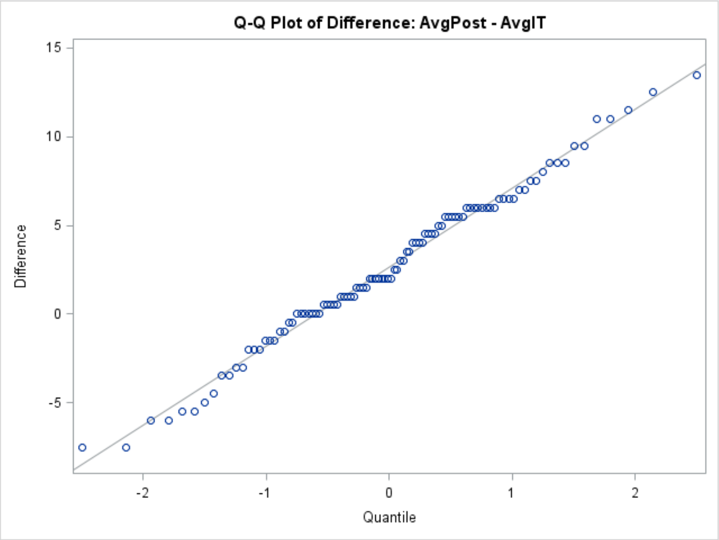

Figure 4.7 QQ Plot of The Difference

The results of the matched pairs t-test reports that there is a significant difference between pre-test scores (M = 27.03, SD = 6.90) and post-test scores (M = 28.98, SD = 6.51); t (99) = 5.95, $p < .0001$, $d = 0.29$. In other words, students' EI test scores significantly improved after a semester of instruction. To clarify, the effect sizes in fields that are attempting to measure latent traits and lack experimental control are expected to be "small", but those effects should not be necessarily interpreted as unimportant (Cohen, 1988). Although the effect size ($d = 0.29$) is not large according to the traditional labeling system, it adequately shows the importance of the pre-post changes in performance on EI.

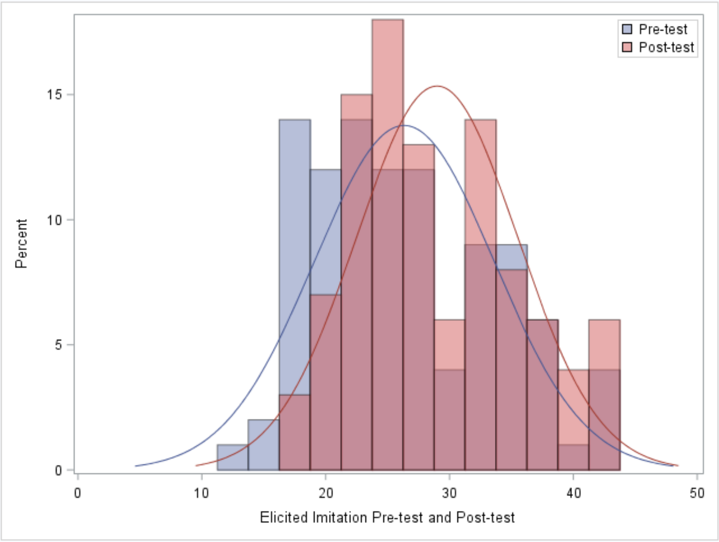

Figure 4.8 Overlay Histogram of the Pre- and Post-test

In addition, Figure 4.8, an overlay histogram of pre- and post-test scores provides a more straightforward representation of the change of the score distribution. The overall score distribution of the post-test shifted to the right, indicating the increase of the overall test scores. The overlay histogram also shows that although the number of students at the high end only has a slight increase, the number of examinees at the low end reduces notably in post-test exams. Especially, the number of examinees whose total test scores were between 10 to 20 declined dramatically during the post-test. This result again indicates that the students who started the program with lower proficiency have more evident gains after a semester of instruction. The students also are likely to have been engaged with peers in English outside of class during the same time frame. These students have made satisfactory progress over time.

4.3.3 Instructional sensitivity

Since a significant difference was found between the pre- and post-test total scores, I further investigated the pre- and post-test item scores by examining the instructional sensitivity index of each item. The index selected to represent instructional sensitivity was Hedge's *g*.
As shown in Figure 4.7, the present EI items have a wide range of instructional sensitivity levels, ranging from -.64 to .78. The items were categorized into two groups based on the absolute value of the instructional sensitivity index: sensitive items (Hedge's g ≥ .30) and insensitive items (.00 < Hedge's g < .30). Sixteen of the 48 items were identified as sensitive items. As shown in Figure 4.9, despite the fact that there were several items with high instructional sensitivity values, more than half of the items obtained a value less than .30. However, finding a relatively small proportion of items with even modest sensitivity is not unusual. Among tests that are distal from any specific classroom curriculum, this pattern appears to be typical (Polikoff, 2016).

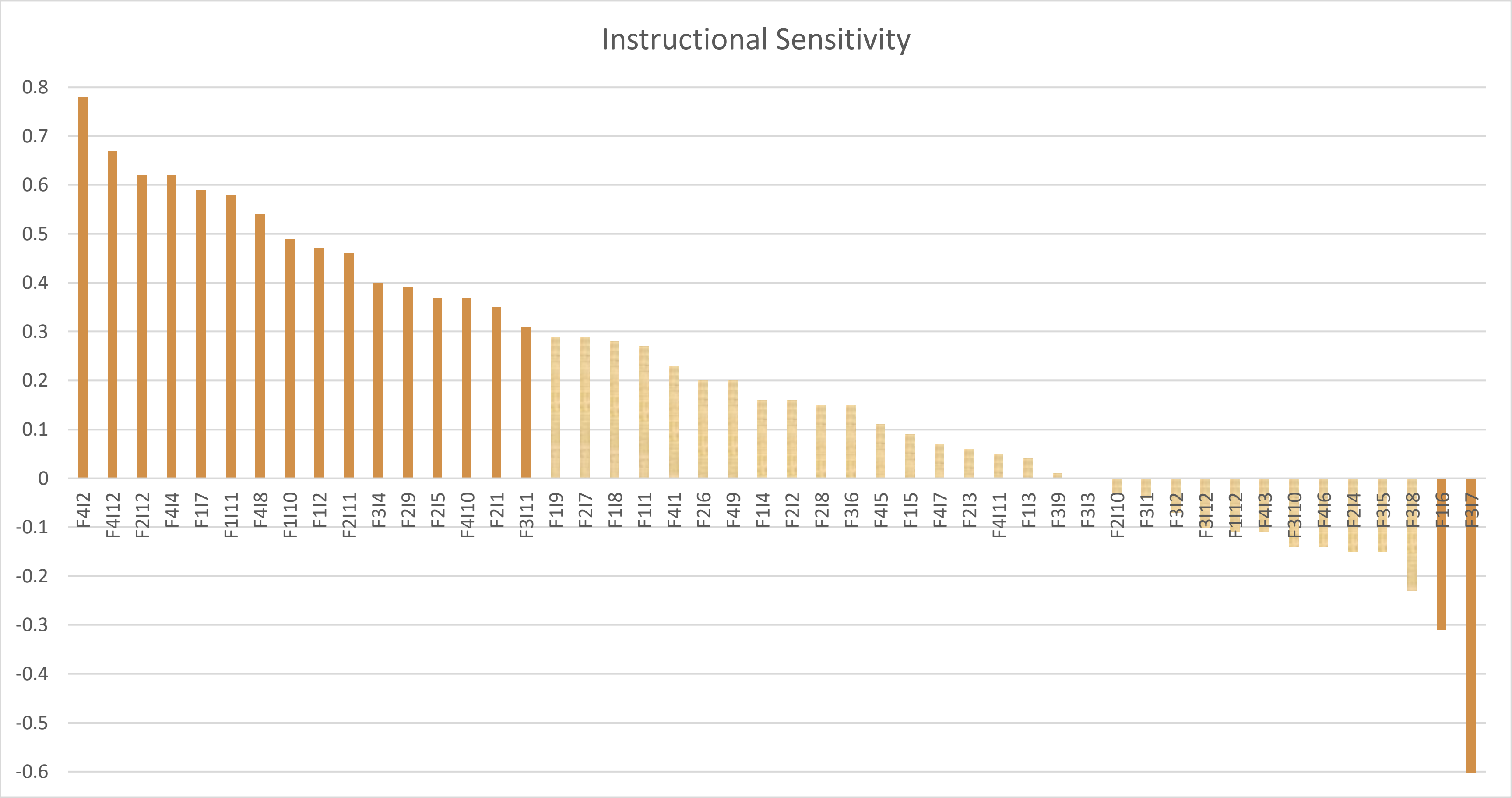

Figure 4.9 Instructional Sensitivity Index of All Items

A small instructional sensitivity index value suggests that either the item is not sensitive enough to capture the gain after instruction or the instructions do not fully match the assessment. For the present EI test, both the quality of the items and the instruction may contribute to the small values, but the restricted range of item difficulty is argued to be the main reason. The absence of difficult items weakens the ability of the test to discriminate at the high end of the test population. As a consequence, the test items were less likely to detect the effect of instruction. In addition, the examinees only took one course of the two-course sequence of the PLaCE program. The small instruction sensitivity may also attribute to the short period of instruction. The instructional sensitivity may improve if the post-test is administered by the end of the second semester.

It is interesting to note that among the sixteen sensitive items, thirteen of them are long sentence stimuli. Although the length of the sentence stimuli did not appear to be correlated with item difficulty or item discrimination based on the reported results, the length of the sentence stimuli may play a role in instructional sensitivity. Long sentences appear to be more sensitive to instruction than medium-length sentences. If the test aims at offering insights into students' gains after instruction, test developers may consider employing more long sentence stimuli.

Furthermore, as shown in Figure 4.9, thirteen items obtained negative Hedge's *g* values, but the absolute values of the majority of the items were smaller than .30. A negative value may reflect true changes in language proficiency as a result of instruction, and it may also capture the failure of random assignment, the problems with item contents, or any systematic measurement errors by the raters. The majority of the items with negative instructional sensitivity index were from Form 3, especially F3I7 (Hedge's *g* = -.63), which gained the largest negative value. Based on previous analyses, most unexpected results that reported by Form 3 were likely to be associated with the imbalance of the sample for which a smaller number of examinees were assigned to Form 3 during the pre-test. The imbalanced sample may also contribute to the negative instructional sensitivity values of Form 3 items. Besides F3I7, another item that had a larger negative instruction sensitivity value was F1I6 (Hedge's *g* = -.31). A possible explanation for the negative value might be the content of the sentence stimuli. The contents of all sentence stimuli were related to university life such as course enrollment, social activities, and on-campus jobs. F1I6, "*Purdue has the third largest Greek system among all public schools*", introduces a topic about the Greek system. The Greek system is most likely an unfamiliar concept for many international students. Item contents that are highly culture-specific may create extra cognitive load for the examinees and influence their performance. As a consequence, the item may be less sensitive to instruction. To disentangle whether there were problems with the item content, the rating or the imbalance of the sample, it might be helpful to draw and examine a more balanced pre-test sample across forms or collect background information that could be used as covariates to count for the imbalances.

Table 4.8 Instructional Sensitivity Levels

Level of Instructional Sensitivity	Item
Sensitive (N=18)	F1: I2, *I6, I7, I10, I11 F2: I1, I5, I9, I11, I12 F3: I4, *I7, I11 F4: I2, I4, I8, I10, I12
Insensitive (N=30)	F1: I1, I3, I4, I5, I8, I9, *I12 F2: I2, I3, *I4, I6, I7, I8, *I10 F3: *I1, *I2, I3, *I5, I6, *I8, I9, *I10, *I12 F4: I1, *I3, I5, *I6, I7, I9, I11

Note. *items with negative values

Besides having the largest number of items with negative instructional sensitivity, Form 3 also has fewer sensitive items. Table 4.8 shows that except for Form 3, all three forms had five highly sensitive items.

Finally, it is surprising to find that F1I11, a problematic item identified by the item discrimination analysis gained a high instructional sensitivity index value (Hedge's g = .57). In other words, although this item failed to separate those of higher proficiency from those of lower proficiency, the item appeared to be sensitive to the changes in language proficiency as a result of instruction. Interestingly, this finding supported the argument made by Polikoff (2010) that the instructional sensitivity index might offer a qualitatively different estimation of the items that were identified as poor items by traditional item statistics such as item difficulty and item discrimination.

4.4 EI and TOEFL iBT

The last stage of data analysis answered the third research question: What are the relationships between EI test scores and TOEFL iBT test scores? Table 4.9 reports both the observed and adjusted Spearman correlation coefficients of the EI test scores with TOEFL iBT subsection scores and total test scores. The adjusted Spearman correlation was reported because the current sample represented a restricted sample of the entire TOEFL iBT test taker population as only students admitted to the PLaCE program were included in the analyses.

Table 4.9 Observed and adjusted Spearman correlations between EI and TOEFL iBT scores (N = 100)

	Pre-test		Post-test	
	r	Adjusted *r*	*r*	Adjusted *r*
Reading	-.21*	-.46	-.11	-.46
Listening	.22*	.49	.14	.49
Speaking	.35***	.62	.40***	.68
Writing	.19	.38	.12	.25
Total	.26**	.76	.23*	.72

Note. **p* < .05, ***p* < .01, ****p* < .001.

Overall, as shown in Table 4.9 the correlation patterns with both the pre- and post-test were similar. There are moderate and positive correlations observed between EI total scores and TOEFL iBT total scores ($r_{pre\text{-}total}$ = .26, $r_{post\text{-}total}$ = .23). However, when broken down into subsection scores, EI scores show different correlation patterns with TOEFL iBT subsections. There are strong and positive correlations observed between speaking subsection scores and EI scores in both the pre-test and post-test ($r_{pre\text{-}speaking}$ = .35***, $r_{post\text{-}speaking}$ = .40***). After correction for range restriction, the correlation coefficients increased ($r_{A\_pre\text{-}speaking}$ = .62, $r_{A\_post\text{-}speaking}$ = .68). However, the correlations between TOEFL reading and EI scores are negative in both the pre- and post-test ($r_{pre\text{-}reading}$ = -. 21*, $r_{post\text{-}reading}$ = -.11). After correction for range restriction, the correlation coefficients increase ($r_{A\_pre\text{-}speaking}$ = -.46, $r_{A\_post\text{-}speaking}$ = -.46).

The significant moderately positive correlation between EI and the speaking subsection of the TOEFL iBT test was anticipated because both tests assessed examinees' productive skills and were rated based on examinees' oral production. Additionally, both tests require examinees to process the given audio information to complete the task. Although I argue that EI is designed to measure examinees' implicit linguistic knowledge, and EI is more than a speaking test, speaking skills are one of the key aspects of language competence assessed by EI. A moderate positive correlation between EI and the speaking subsection of the TOEFL iBT test also provides external validity evidence for the current EI test. While EI has been frequently criticized by its lack of authenticity, free-response speaking items used by TOEFL speaking subsection are argued to be authentic and have high face validity. The significant moderate correlations between EI and TOEFL speaking and total scores indicate that EI is likely to be an effective measure of English language proficiency, especially speaking skills. Regardless of the different task design, EI provides useful information regarding examinees' speaking skills similar to free-response tasks. Compared to free-response items, the simplicity of the test format and the ease of rating place EI in an advantageous position to be widely implemented in language assessment. EI test might be an ideal alternative to free-response items if the institution cannot afford a long period of testing and rater training.

More interestingly, a negative correlation was reported between EI scores and the TOEFL iBT reading subsection ($r_{pre\text{-}reading}$ = -. 21*, $r_{post\text{-}reading}$ = -.11). The negative correlation suggests that students who scored higher in TOEFL reading tend to score lower in the EI test. One possible explanation for this result is that the TOEFL reading subsection measures a different aspect of language proficiency which does not require the use of implicit knowledge which is crucial to EI. EI is designed to assess examinees' ability to process English in real-time whereas the TOEFL reading section which only includes multiple-choice items is less likely to be assessing real-time processing skills, nor it is intended to. In addition, another possible explanation to the negative correlations is that TOEFL reading subsection scores are easier to improve with intensive test preparations than other subsections. Ginther and Yan (2018) noted that an unexpected advantage on the multiple-choice items might result from excessive test preparation. For example, examinees who scored very high in the TOEFL reading section might be experienced test takers who had crammed for the test and were well-trained in test-taking strategies. On the contrary, the possibility of excessive test preparation for an EI test remains low as it measures the real-time language processing skill, and test-taking strategies for common item types can be rarely applied to EI. Therefore, the possible mismatch between TOEFL reading scores and examinee's language proficiency could lead to the negative correlation between the TOEFL iBT reading subsection and the current EI test.

In sum, the correlation patterns between EI scores and TOEFL iBT scores provide external validity evidence for the current EI test. The results also suggest that the present EI test can provide additional information in relation to examinees' language proficiency that the TOEFL iBT test is unable to provide, especially examinees' ability to process language in real-time. Although EI is seemingly less authentic than free-response speaking items, the correlation between EI and the TOEFL iBT speaking subsection is positive and strong providing evidence in favor of the use of EI as a viable alternative to free-response items. In addition, as multiple-choice reading items and standardized writing tasks may lend themselves to a practice effect, EI is a promising alternative that can reduce the impact from excessive test preparation. As higher test scores from standardized tests do not always guarantee more capable language users that can comfortably meet all the language demands in an academic setting, an additional language test such as EI can complement the information provided by standardized tests. The current EI test can be used not only as a post-entry language test that helps institutions to identify students who need additional language support but also as a supporting tool for language instructors to track the language development of the students enrolled in language support classes.

CHAPTER 5. CONCLUSIONS AND IMPLICATIONS

5.1 Summary of the study findings

The present study examined the technical qualities of the EI items designed for the ACE-In, a locally developed English language proficiency test used in the undergraduate English Academic Purpose Program at Purdue University. To offer insights into the quality of the EI subsection of the ACE-In and to provide guidance for continued test development and revision, this study investigated item difficulty, item discrimination, total score reliability, and pre-post changes in performance on the EI. Additionally, the study further examined the relationship between EI test scores and TOEFL total and subsection scores.

The majority of the present EI items were found to be of easy or average in difficulty. The item difficulty is slightly restricted as there are few difficult items. Although the overall test difficulty is acceptable for the intended test population, the variability of item difficulty could be improved within each form. In addition, the majority of the test items have high discrimination in both the pre and post-test showing that the present EI items are effective in terms of separating those of higher language proficiency from those of lower language proficiency. Furthermore, all four test forms were found to have satisfactory internal consistency. This result provided evidence that within each form, all 12 items functioned effectively and measure the same underlying construct as a group.

The study also examined pre-post changes in performance on the EI section and found a significant pair-wise difference between the pre- and post-test performance after a semester of instruction. However, the analysis of instructional sensitivity showed that over half of the items were not particularly sensitive given their Hedge's *g* values. Several factors may contribute to small instructional sensitivity values. The absence of difficult items was believed to be one of the main reasons. In addition, the content of the sentence that biases the test results may also account for small or negative instructional sensitivity values.

Lastly, the significant positive correlation between EI scores and TOEFL iBT speaking subsection scores provided validity evidence for EI items if and when TOEFL iBT scores are considered the appropriate criterion. EI tasks can be reasonably used as an alternative to free-response speaking items especially when the test and rating has to be completed within a short period of time. Furthermore, the negative correlation between EI scores and TOEFL reading subsection scores suggested that the EI task can provide additional information related to examinees' language proficiency such as their real-time language processing ability that multiple-choice reading tests were not intended to measure. Due to the unique test format, EI, to a large extend, also avoids the practice effect that often observed in multiple-choice items and traditional writing tasks.

5.2 Limitations of the study

One main limitation of this study is that the random assignment of examinees to forms probably failed to fully balance the form samples on student characteristics. The effects of the imbalanced sampling were more pronounced as this study had a relatively small sample size. The number of examinees assigned to Form 3 was much smaller than the other three forms during the pre-test. Most of the unexpected results reported in item analysis were associated with Form 3. For example, Form 3 obtained an unexpected high mean total score in the pre-test, and many items became more difficult when used in the post-test. As the imbalance of the sampling could impact the analyses, the reported results might not accurately reflect the technical qualities of Form 3 items. Thus, I encountered some difficulties interpreting the results.

Although several unexpected results related to Form 3 were reported, the results of the other three forms demonstrated many similarities and provided references for overall result interpretations. As the sampling process was based on several considerations including the completion of the pre-/post-test, the availability of TOEFL iBT scores, and the makeup of the native countries of subjects, it is likely to produce an imbalanced sample with 100 subjects. To reduce the possibility of imbalanced sampling, one solution is to increase the overall sample size. If the sample is balanced across forms and the problems with Form 3 persists, it might be helpful to examine the TOEFL iBT scores of the examinees who took Form 3 during the pre-test.

Another limitation of this study is that one semester as the interval between the pre- and post-test might not be long enough for the instructional sensitivity study to provide an accurate estimation. The result reported that over half of the items were relatively insensitive to the instructions. As examinees only completed one language support course of the two-course sequence offered by the PLaCE program when taking the post-test, the effect of the instruction may be underestimated. The instructional sensitivity of the items may improve if the post-test were administered at the end of the second semester of instruction.

5.3 Recommendations for Future Research and Test Development

I argue that the EI subsection of the ACE-In can be responsibly used as a measure of L2 language proficiency. EI requires careful task design and regular examinations of the technical qualities of the items. Therefore, future research and further discussions on EI are beneficial. Two recommendations for future research are drawn based on the findings and limitations of the present study. First, although item analysis is fundamental to establish validity evidence for EI, examining technical qualities alone is insufficient. As the ACE-In is a recently developed language test, the test can benefit from test validation studies that exceed item analysis. Future studies that focus on providing construct-related and content-related validity evidence can complement the present study and provide more evidence for using EI as a measure of L2 proficiency. For example, results from generalizability studies that demonstrate EI scores would remain consistent under different measurement conditions might be useful.

Furthermore, it is worth further exploring the instructional sensitivity of EI items by examining post-test scores collected after the two-course sequence. Although the ACE-In was originally developed as a post-entry language test to screen the language proficiency of incoming international students, the major use of this test has gradually shifted to examine students' language development over time. The information of the EI test has also been used to exempt students from the program. Therefore, developing items that are sensitive to instruction has become increasingly important for the program. Investigating the instructional sensitivity by collecting post-test scores by the end of the second semester can offer insights into the effectiveness of EI items and provide valuable information for the program.

To continue improving the current EI test, test developers may consider the following aspects. First of all, more difficult items should be added to each form. With the help of difficult items, the test can assess the language proficiency at the high end of the current test population. Increasing the overall difficulty may also improve the instructional sensitivity of the items. Test developers may consider replacing medium length sentences that were identified as easy with longer sentences, especially long sentences with lexicalized stems embedded. Secondly, the variability of the item difficulty of Form 4 needs improving. Besides adding difficult items, involving a few easy items could also be beneficial as the item difficulty of Form 4 were the most restricted and clustered. Thirdly, two items of the current test should be removed or rewritten: F1I11, *“Taking a part-time job on campus has been shown to help students succeed in college”*, and F3I6, *“Purdue has the third largest Greek system among all public schools”*. These two items were identified as problematic and functioned unsatisfactorily in several aspects.

Finally, the advantages of EI tasks are evident: the assessment of real-time language processing ability, the ease of rating, and high internal consistency. It is promising that EI tasks are to be widely implemented to measure English language proficiency of L2 speakers. Nevertheless, EI, a seemingly simple task, requires a thoughtful design because the design of the task directly impacts its effectiveness. Based on the present study, I have four general recommendations for future EI test development. First, the EI test design should possess reconstructive features: controlling the length of the sentence stimuli, avoiding inauthentic grammatical structure, including a repetition delay task, and guiding examinees to focus on meaning. Second, the contents of sentence stimuli should be authentic and idiomatic. To reduce construct-irrelevant variance, topics that are culture-specific and context-specific should be avoided. For example, one of the sentence stimuli of the current EI task is related to the Greek fraternity system in the U.S. This topic may be relatively foreign to some groups of international students and may bias the test scores. Considering the diverse educational and cultural background of international students, I recommend selecting general topics that do not require any prior knowledge; topics that advantage any subgroup of the test population are not appropriate. Finally, I recommend using easy items at the beginning of the EI test. Compared to multiple-choice or free-response items, EI may be a less widely implemented language task. As many examinees may not be very familiar with the test format, starting with an easy item can facilitate examinees to quickly familiarize the test format and reduce possible construct-irrelevant variance.

APPENDIX A SAMPLE ITEM

Directions and sample elicited imitation tasks

Introduction. In this section, you will hear 12 sentences. Each sentence will be played once. After each sentence, the screen will change, and two words will appear. One of the two words was mentioned in the sentence.

Task. your task is to (1) identify the word that was mentioned in the sentence, then (2) repeat the sentence that you heard. Try to repeat the sentence exactly as it was stated.

Preparing for your response. Listen to each sentence carefully. You will have 5 seconds to choose the word and 20 seconds to repeat each sentence.

Sample Item:

You will hear the following sentence:

Parking on campus is free on Sunday. (AUDIO ONLY)

Click on the word below that you heard the sentence? *(CLICK ON WORD)*

Parking　　　　Swimming

The word mentioned in the example sentence was *Parking*, so you should have clicked on *Parking.*

OK, now **repeat the sentence you heard after you hear a voice that states, "recording now":**

Parking on campus is free on Sunday.

APPENDIX B TEST ITEMS

Form 1

1. Most students declare their major at the end of their sophomore year in college.
2. College students can ride the bus for free as long as they have a valid student ID.
3. Last month we traveled to Chicago, which is the third largest city in the country.
4. By the way, you can always borrow textbooks from the library or buy them online.
5. First of all, you must attend all the classes to pass this course.
6. It doesn’t matter if you work alone or in a group on your homework.
7. Earning money is the main reason for students to get a job.
8. If you record your lectures, you can revise your class notes later.
9. The way that English classes are taught here might be quite different from your country.
10. Purdue ranks second in foreign student enrollment among all public schools.
11. Taking a part-time job on campus has been shown to help students succeed in college.
12. Although he did not review for the final exam, he scored very high on that test.

Form 2

1. The amount of work involved in studying for final exams can overwhelm you.
2. Before you arrive on campus, you need to make sure that you have a place to live.
3. Joining a student club on campus is a great way to improve your social skills.
4. The wonderful thing about English teachers is that they know their students quite well.
5. You should come up with a topic for your project by midterm.
6. It looks like I only have morning classes this semester.
7. Working part-time will help you develop time management skills.
8. If you have a morning class, you should go to bed early the night before.
9. You can look at the course schedule to see the dates for midterm and final exams.
10. You can tell me what questions you have on the final project during my office hours.
11. Borrowed books form the library must be returned or renewed by the posted due dates.
12. After he worked on the project all evening, the student went directly to bed.

Form 3

1. Sometimes it’s helpful to ask questions in class as opposed to keeping them to yourself.
2. Foreign students are only permitted to work part-time for employers on campus.
3. Students can keep the books that they borrow from the library for a semester.
4. When you look at the course schedule, you will see the dates for midterm and final exams.
5. Regular workouts benefit the body as well as the mind.
6. Purdue has the third largest Greek system among all public schools.
7. You can also talk to senior students about selecting courses.
8. Many students live off campus because the rent is much lower.
9. The senior student was talking about his own story of finding an apartment.
10. Students can take courses that have nothing to do with their major areas of study.
11. It can be very tough for foreign students to speak English on a daily basis.
12. You can make some extra cash if you work part-time at at dining court on campus

Form 4

1. As you can see on the course schedule, we will not have a final exam for this course.
2. In the event of a car accident, you should first stay calm and then call the police.
3. Regular exercise is extremely important for long-term health and well-being.
4. In other words, you must submit all your homework assignments on the course website.
5. The process of applying for graduate school is quite long.
6. Students who enjoy working in groups are more likely to succeed.
7. Most college students move out of the dorms after their sophomore year.
8. It’s hard to express your ideas if your language skills are low.
9. When you take courses here, attendance often counts as a part of the final grades.
10. Meeting people and making friends should be an important part of your college life.
11. Students should know that they can also borrow books from libraries of other schools.
12. You should talk to your advisor if you are not sure what courses to take next semester.

REFERENCES

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment.* Cambridge, UK: Cambridge University Press.

Baker, E. L., & Herman, G. L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement, 20*(2), 149-164.

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language learning*, *29*(1), 81-103.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In Earone, E., Gass, S., & Cohen, A. (Ed.), *Research methodology in second-language acquisition*, (pp. 245-261). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: CBS College Publishing.

DeKeyser, R. (2008). Implicit and Explicit Learning. In C. Doughty & M. Long (Eds). *The handbook of second language acquisition*. Malden, MA: Blackwell Pub.

Dörnyei, Z. (2009). *The psychology of second language acquisition.* Oxford, UK: Oxford University Press.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, N.J: Prentice-Hall.

Educational Testing Service. (2016). *Test and score data summary for TOEFL iBT tests*. Princeton, NJ: Educational Testing Service.

Ellis, N. C. (1994). Introduction: Implicit and explicit language learning – an overview. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages.* London, UK: Academic Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(2), 141-172.

Ellis, R. (2009a). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol, UK: Multilingual Matters.

Ellis, R. (2009b). Measuring implicit knowledge and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching.* Bristol, UK: Multilingual Matters.

Ellis, R. (1993). The structural syllabus and second language acquisition. *TESOL Quarterly, 27*(1), 91-113.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464-491.

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of verbal learning and verbal behavior*, *2*(2), 121-135.

Fulcher, G. (2013). *Practical language testing* (2nd ed.). London, UK: Hodder Education.

Gallimore, R., & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, *31*(2), 369-392.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271-295.

Haan, J. E. (2009). *ESL and internationalization at Purdue University: A history and analysis* (Publication No. 3378759) [Doctoral dissertation, Purdue University]. ProQuest Dissertations Publishing.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement*, *18*(1), 39-53.

Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech, 20*, 86-97.

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, *2*(1), 1-23.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics*, *6*(2), 107-128.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Boston, MA: Heinle & Heinle Publisher.

Hood, L., & Lightbown, P. (1978). What children do when asked to "say what I say" – Does elicited imitation measure linguistic knowledge? *Allied Health and Behavioral Sciences. 2*(1), 195-219.

Hood, L. & Schieffelin, B. B. (1978). Elicited imitation in two cultural contexts. *Quarterly Newsletter of the institute for Comparative Human Development*, *2*(2), 4-12.

Hsieh, A. F. Y., & Lee, M. K. (2014). The Evolution of Elicited Imitation: Syntactic Priming Comprehension and Production Task. *Applied Linguistics*, *35*(5), 595-600.

Huitt, W. (2003). The information processing approach to cognition. Educational Psychology Interactive. Valdosta, GA: Valdosta State University. Retrieved from http://www.edpsycinteractive.org/topics/cognition/infoproc.html

Hulstijn, J. H., & Graaff, R. D. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, *11*, 97-112.

Hulstijn, J. H., & Hulstijn, W. (1984). Grammatical errors as a function of processing constraints and explicit knowledge. *Language learning*, *34*(1), 23-43.

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*(1), 215-238.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

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

Krashen, S. (1982). *Principles and practice in second language acquisition.* New York: Pergamon Press.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/43

Lado, R. (1965). Memory span as a factor in second language learning. *IRAL-International Review of Applied Linguistics in Language Teaching*, *3*(2), 123-130.

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research*. London, UK: Longman.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, Massachusetts: The MIT Press.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement*, *20*(2), 179-189.

McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders*, *47*(1), 19-24.

McLaughlin, B. (1978). The monitor model: Some methodological considerations. *Language learning, 28*(2), 309-332.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (Eds.), *Advancing Human Assessment* (pp. 19-46). Cham, Switzerland: Springer.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, *7*(2), 191-205.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research methodology in second language acquisition*. New York, NY: Psychology Press.

Naiman, N. (1974). The Use of Elicited Imitation in Second Language Acquisition Research. Working Papers on Bilingualism, No. 2.

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and Relative Measures of Instructional Sensitivity. *Journal of Educational and Behavioral Statistics*, *42*(6), 678-705.

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment*, *21*(2), 89-101.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*(3), 337-347.

Ortega, L., Iwashita, N., Norris, J.M. & Rabie, S. (1999). *A Multilanguage Comparison of Measure of Syntactic Complexity*. [Funded Project]. Honolulu, HI: University of Hawaii, National Foreign Language Resource Center.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication, 191*, 225.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3-14.

Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment, 21*(2), 102-119.

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-155.

Purdue University (2017). *International students and scholars enrollment & statistical report*. https://www.purdue.edu/IPPU/ISS/_Documents/EnrollmentReport/ISS_StatisticalReportFall17.pdf

Read, J. (2015). Issues in post-entry language assessment in English-medium universities. *Language Testing*, *48*(2), 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.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595-626.

Sarandi, H. (2015). Reexamining elicited imitation as a measure of implicit grammatical knowledge and beyond…? *Language Testing*, *32*(4), 485-501.

Savignon, S. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied linguistics*, *11*(2), 129-158.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, *11*, 11-26.

Seliger, H. W. (1979). On the nature and function of language rules in language teaching. *TESOL Quarterly*, 359-369.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. In *Studies of child language development* (pp. 175-208). Holt, Rinehart, & Winston.

Smith, C. (1973). An experimental approach to children. In C. Ferguson & D. Slobin (Eds.). *Studies of child language development*. New York, NY: Holt, Rinehart & Winston.

Smith, M. S. (1981). Consciousness-raising and the second language Learner. *Applied linguistics*, *2*(2), 159-168.

Swain, M., Dumas, G., & Naiman, N. (1974). Alternatives to Spontaneous Speech: Elicited Translation and Imitation as Indicators of Second Language Competence. Working Papers on Bilingualism, No. 3.

Thirakunkovit, S. (2016). *An evaluation of post-entry test: An item analysis using classical test theory (CTT)* (Publication No. 10179945) [Doctoral dissertation, Purdue University]. ProQuest Dissertations Publishing.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Thorndike, R., & Thorndike, T. (2009). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325-344.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54-73.

Yan, X. (2015). *The processing of formulaic language on elicited imitation tasks by second language speakers* (Publication No. 3721116) [Doctoral dissertation, Purdue University]. ProQuest Dissertations Publishing.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497-528.