2015

An argument-based validation study of the English Placement Test (EPT) – Focusing on the inferences of extrapolation and ramification

Zhi Li
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Educational Assessment, Evaluation, and Research Commons, English Language and Literature Commons, Higher Education Administration Commons, and the Higher Education and Teaching Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/14538

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
An argument-based validation study of the English Placement Test (EPT)

– Focusing on the inferences of extrapolation and ramification

by

Zhi Li

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirement for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics & Technology

Program of Study Committee:

Carol Chapelle, Major Professor
Volker Hegelheimer
Tammy Slater
Barbara Blakely
Frederick Lorenz

Iowa State University
Ames, Iowa
2015

Copyright © Zhi Li, 2015. All rights reserved.
Dedication

To my beloved and inspiring wife, Ling Luo, lovely daughters Ziqian Li and Sijin Li
# TABLE OF CONTENTS

LIST OF TABLES ......................................................................................................................... viii

LIST OF FIGURES .................................................................................................................... x

ACKNOWLEDGEMENTS ........................................................................................................... xii

ABSTRACT ................................................................................................................................. xiv

CHAPTER 1. INTRODUCTION ..................................................................................................... 1

1.1. ESL students and English placement tests ................................................................. 2

1.2. Key concepts in argument-based approach to validation ............................................. 3

1.3. Context of this study and research questions ............................................................... 7

1.4. Significance of this study .............................................................................................. 9

1.5. Overview of the dissertation ........................................................................................ 11

CHAPTER 2. LITERATURE REVIEW ......................................................................................... 13

2.1. Argument-based approach to validating English placement tests in general .......... 13

2.1.1. Brief introduction to argument-based approach to validation ................................. 14

2.1.2. An interpretation and use argument for the English Placement Test (EPT)......... 16

2.2. Relationship between the EPT and other measures of English proficiency .......... 27

2.2.1. Standardized English proficiency tests as external criteria .................................. 27

2.2.2. Self-assessment as an external criterion ................................................................. 30

2.2.3 Summary and research question 1 ............................................................................ 36

2.3. Impact of language tests .............................................................................................. 37

2.3.1. Test stakeholder’s perceptions of impact of English placement tests ................. 37

2.3.2. The relationship between of English proficiency and academic achievement ....... 40
2.3.3. The roles of motivational factors in predicting academic achievement .................43
2.3.4 Summary and research questions 2 and 3 .................................................................46

CHAPTER 3. METHODOLOGY .................................................................................................48

3.1. Overall research design .................................................................................................48

3.2. Participants .....................................................................................................................50

3.2.1. ESL students ..............................................................................................................50
3.2.2. ESL instructors ..........................................................................................................55
3.2.3. Academic advisors .....................................................................................................56
3.2.4. Content course instructors .........................................................................................57

3.3. Instruments ....................................................................................................................58

3.3.1. The English Placement Test (EPT) ...........................................................................58
3.3.2. In-class tests and final exams in ESL courses ...............................................................60
3.3.3. Questionnaire ............................................................................................................61
3.3.4. Interview questions for test stakeholders ..................................................................66

3.3.5. Lexical Complexity Analyzer (LCA) and L2 Syntactic Complexity Analyzer (L2SCA) .........................................................................................................................67
3.3.6. Criterion .....................................................................................................................69

3.4. Data collection procedures .............................................................................................70

3.5. Data analysis methods .....................................................................................................73

3.5.1. Descriptive statistics and reliability analysis .................................................................73
3.5.2. Rasch model analysis of the self-assessment (RQ1) ......................................................74
3.5.3. Confirmatory factor analysis (RQ1) ............................................................................78
3.5.4. Multitrait-multimethod analysis of the EPT, the TOEFL iBT and the self-assessment (RQ1) .........................................................................................................................79
3.5.5. Inductive analysis of the interview transcripts (RQ2) ........................................82
3.5.6. Paired-samples t-tests of ESL course performance data (RQ3) ..........................83
3.5.7. Confirmatory factor analysis (RQ4) ..................................................................85
3.5.8. Structural equation modeling (RQ4) .................................................................86
3.6. Chapter summary .................................................................................................88

CHAPTER 4. RESULTS ...............................................................................................90

4.1 The relationship among the EPT and self-assessment and the TOEFL iBT ........90

4.1.1. Descriptive statistics and reliability analysis of the EPT, the TOEFL iBT and
the self-assessment ........................................................................................................91
4.1.2. Rasch model analysis and confirmatory factor analysis of the self-assessment.....95
4.1.3. Factor structure of self-assessment and factor scores ........................................102
4.1.4. Multitrait-multimethod correlation coefficient matrix .....................................107
4.1.5. Correlated trait-correlated methods (CTCM) model and correlated trait-
correlated uniqueness (CTCU) model ........................................................................114
4.1.6. Summary ..........................................................................................................118

4.2 Test stakeholders’ perceptions ..............................................................................119

4.2.1. Perceptions of the EPT decisions ......................................................................119
4.2.2. Summary ..........................................................................................................127

4.3. Impact of the EPT placement decisions on ESL learners’ English learning ........127

4.3.1. ESL students’ progress in Engl99L ..................................................................128
4.3.2. ESL students’ progress in ESL writing classes .................................................130
4.3.3. Qualitative Findings regarding the Impact of the EPT decisions on ESL learners’
English learning ..............................................................................................................139
4.3.4. Summary ..........................................................................................................143
4.4. Impact of the EPT decisions on ESL learners’ academic achievement ................. 144

4.4.1. Descriptive statistics and reliability analysis ............................................. 144

4.4.2. Confirmatory factor analysis of the subscales .......................................... 150

4.4.3. Structural equation modeling analysis ...................................................... 153

4.4.4. Qualitative Findings regarding the Impact of the EPT decisions on ESL learners’ academic achievement ................................................................. 159

4.4.5. Summary .................................................................................................... 165

4.5. Chapter summary .......................................................................................... 166

CHAPTER 5. CONCLUSIONS AND DISCUSSION ................................................. 168

5.1. Summary of primary findings to the research questions ................................. 168

5.1.1. Research question 1: The relationship between the EPT and two external criteria of English proficiency ................................................................. 168

5.1.2. Research question 2: Test stakeholders’ perceptions of the EPT placement decisions and their impact ............................................................... 169

5.1.3. Research question 3: The impact of the EPT on English learning .............. 171

5.1.4. Research question 4: The impact of the EPT on academic achievement ......... 172

5.2. The validity argument for the EPT ............................................................... 173

5.2.1. The extrapolation inference ...................................................................... 174

5.2.2. The ramification inference ....................................................................... 176

5.3. Implications for English placement tests in general and the EPT in particular . 180

5.3.1. Theoretical and methodological implications ............................................ 180

5.3.2. Practical implications for future development of the EPT ......................... 181

5.4. Limitations and future studies ..................................................................... 185

5.5. Concluding comments .................................................................................. 188
REFERENCES .......................................................................................................................... 189

APPENDIX A. SURVEY ABOUT SELF-ASSESSMENT OF ENGLISH USE AND DLEARNING MOTIVATION .................................................................................................................. 203

APPENDIX B. SAMPLE INFOMRED CONSENT FORM .................................................................. 211

APPENDIX C. INTERVIEW QUESTIONS FOR ESL STUDENTS .................................................. 213

APPENDIX D. INTERVIEW QUESTIONS FOR ESL COURSE INSTRUCTORS .......................... 215

APPENDIX E. INTERVIEW QUESTIONS FOR CONTENT COURSE INSTRUCTORS ........... 217

APPENDIX F. INTERVIEW QUESTIONS FOR ACADEMIC ADVISORS .................................. 219

APPENDIX G. COURSE SYLLABUS FOR ENGL99L ............................................................... 221

APPENDIX H. COURSE SYLLABUS FOR ENGL101b ............................................................. 223

APPENDIX I. COURSE SYLLABUS FOR ENGL101C ............................................................. 226

APPENDIX J. ESSAY PROMPTS USED IN THE IN-CLASS TESTS IN ENGL101B/C ............ 230

APPENDIX K. RATING RUBRIC FOR THE EPT WRITING SECTION .................................. 231

APPENDIX L. SUMMARY OF THE LEXICAL COMPLEXITY MEASURES IN THE LEXICAL COMPLEXITY ANALYZER (LCA) .......................................................... 233

APPENDIX M. SUMMARY OF THE SYNTACTIC COMPLEXITY MEASURES IN L2 SYNTACTIC COMPLEXITY ANALYZER (L2SCA) ......................................................... 235

APPENDIX N. TRAIT FEEDBACK CATEGORIES AND EXAMPLES ON Criterion ........... 236
LIST OF TABLES

Table 3.1 Characteristics of the Participants of Self-assessment and Motivation Questionnaire ........................................52
Table 3.2 Demographic Information of the ESL Student Interviewees .................................................................54
Table 3.3 Demographic Information of the ESL Instructor Interviewees ..............................................................56
Table 3.4 Demographic Information of the Academic Advisors ........................................................................57
Table 3.5 Demographic Information of the Content Course Instructors .............................................................57
Table 3.6 Summary of Data Collection and Analysis ........................................................................................88
Table 4.1 Descriptive Statistics for the EPT \( (n = 202) \) ....................................................................................91
Table 4.2 Descriptive Statistics for the TOEFL iBT \( (n = 202) \) ...........................................................................93
Table 4.3 Descriptive Statistics for the Self-assessment \( (N = 347) \) .................................................................94
Table 4.4 Results of Rasch Measurement Analysis: Reliability and Unidimensionality Check for the Self-assessment \( (N = 347) \) ...................................................................................95
Table 4.5 Results of Rasch Measurement Analysis: Item measure and Fit Statistics in the Self-assessment \( (N = 347) \) ..................................................................................................................97
Table 4.6 Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Reading Section \( (N = 347) \) .................................................................99
Table 4.7 Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Listening Section \( (N = 347) \) .................................................................101
Table 4.8 Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Speaking Section \( (N = 347) \) .................................................................101
Table 4.9 Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Writing Section \( (N = 347) \) .................................................................102
Table 4.10 Model fit indices of the competing models for the self-assessment \( (n = 339) \) ..................105
Table 4.11 Descriptive Statistics for the Factor Scores of the Self-assessment \( (n = 202) \) ..................106
Table 4.12 Multitrait-multimethod Correlation Matrix for the Fall 2014 Data \( (n = 202) \) ........111
LIST OF FIGURES

Figure 1.1 Toulmin’s model of argument........................................................................................................6
Figure 2.1 An interpretation and use argument for the English Placement Test (EPT) ..................17
Figure 2.2 The domain description inference for the EPT.................................................................................19
Figure 2.3 The evaluation inference for the EPT..............................................................................................21
Figure 2.4 The generalization inference for the EPT .....................................................................................22
Figure 2.5 The explanation inference for the EPT............................................................................................23
Figure 2.6 The extrapolation inference for the EPT .........................................................................................25
Figure 2.7 The ramification inference for the EPT...........................................................................................26
Figure 3.1 Mixed-methods sequential explanatory design ..............................................................................49
Figure 3.2. Structure of the English Placement Test (EPT) and corresponding ESL courses ..........60
Figure 3.3 Procedures and phases of data collection .....................................................................................70
Figure 3.4 Hypothesized correlated trait-correlated method (CTCM) model (left) and correlated trait-correlated uniqueness (CTCU) model (right). .................................................................82
Figure 3.5 Steps for interview data analysis ..................................................................................................83
Figure 3.6 The pre- and posttest design for research question 3 .................................................................84
Figure 3.7 Hypothesized model of the relationship among the latent variables with GPA........87
Figure 4.1 Category probability curves for the subscales of self-assessment..............................................100
Figure 4.2 Models tested as potential explanatory models for response patterns on the self-assessment survey (n = 339) of English proficiency as measured by the self-assessment ..........................................................103
Figure 4.3 Correlated trait-correlated uniqueness (CTCU) model for the MTMM data .....................116
Figure 4.4 Category probability curves for the subscales of academic self-efficacy (ASE), motivation (MO), and self-regulated learning strategy (SRL) ........................................................................150
Figure 4.5 Hypothetical model for the academic self-efficacy, motivation, and self-regulated learning ........................................................................................................................................151
Figure 4.6 A simplified fully recursive model of the relationship among the EPT, motivational factors, and cumulative GPA, with standardized parameters ..........................157

Figure 4.7 A simplified final model of the relationship among the EPT, motivational factors, and cumulative GPA, with standardized parameters ..........................................................158
ACKNOWLEDGEMENTS

Five years ago, I made the decision of joining the Applied Linguistics and Technology (ALT) program at Iowa State University. At the end of this dissertation project, I found this decision brought me extremely rewarding and pleasant experiences in the States, academically and personally. First and foremost, I would like to express my sincere and deep gratitude to my committee chair, Dr. Carol Chapelle, who supported me with great mentorship and guidance throughout the course of this project. Like many of my colleagues, Dr. Chapelle was the main reason why I, as a passionate language tester, chose Iowa State University. In the past five years, I grew professionally through taking her courses, working with her on projects, and learning more from her in the dissertation project. I was so blessed to have had her as my major professor.

I am indebted to other committee professors. Dr. Volker Hegelheimer has been a great mentor, boss, and role model to me as he demonstrated top-notch scholarship, integrity in research, and more importantly, an impressive balance of family and academic work. Dr. Slater is such a knowledgeable and inspiring professor. She introduced Systemic Functional Linguistics (SFL) to me and more importantly promising applications of SFL to language testing research. Dr. Blakely was my mentor in Engl500 Proseminar (teaching first-year composition) about four years ago. It is a great honor to have her on the committee as she always offers timely and insightful comments regarding pedagogy and learning. Dr. Lorenz played an important role in my study at Iowa State and I gained so much new knowledge and understanding in statistics and advanced research methodology from his courses. My committee professors have helped me
shape the dissertation with patient guidance and constructive comments. Of course, I am solely responsible for all remaining errors.

I am sincerely thankful to other faculty members in the ALT program, especially, Dr. Gary Ockey for his teaching of advanced measurement theories and sharing of research experiences. I am appreciative of the help of my colleagues in the ALT program, especially the large cohort of 2010 PhD students, with whom I shared so many good memories in various classes, conferences, and parties.

My appreciation goes out to the students, ESL instructors, academic instructors, and academic advisors who willingly participated in my study. Without their valuable time and help, this dissertation would not have been possible. I am also grateful for the financial support from the Educational Testing Service (ETS) through the Small Grant for Doctoral Research in Second or Foreign Language Assessment. Without this generous support, the data collection and analysis procedures would not have gone smoothly.

I would like to thank my parents and parents-in-law for their love and encouragement when I was away from them. I am also indebted to Professor Yunnan Xiao, my former advisor at Hunan University, China. Professor Xiao introduced me to the field of language testing and encouraged me to pursue further development in academia.

My special thanks go to my caring and supportive wife, Ling, who sacrificed her own career as a university teacher and became a “domestic engineer”, taking care of me and two daughters in the States. Her constant support made my study more enjoyable. I owe too much to my amazing daughters, Ziqian and Sijin. I wish I had spent more time playing with them and reading bedtime stories to them.
ABSTRACT

English placement tests have been widely used in higher education as post-admission assessment instruments to measure admitted English as a second language (ESL) students’ English proficiency or readiness in academic English, usually upon their arrival at universities in English-speaking countries. Unlike commercial standardized English proficiency tests, many English placement tests are locally developed with comparatively limited resources and are relatively under-investigated in the field of language testing. Even less attention has been directed to the score interpretation and the impact of placement decisions on ESL students’ English learning and academic achievement. Undoubtedly, this scarcity of research on English placement tests is inappropriate in view of their status as one of the most frequently used language testing instruments, which may exert immediate and strong impact on ESL students’ learning in general.

By employing a mixed-methods approach, this dissertation project investigates the validity of test score interpretation and use of the English Placement Test (EPT) used at Iowa State University (ISU) under an argument-based validity framework. More specifically, this study started with an interpretation and use argument for the EPT, which states the score meaning and intended impact of the EPT explicitly, and focused on the last two inferences in the interpretation and use argument, namely extrapolation and ramification. The extrapolation inference links expected scores of the EPT (scores that exhibit adequate test reliability) to target scores or actual performance in the target domain. In this study, the extrapolation inference requires investigation of the relationship between ESL students’ English placement test performance and two external criteria of English performance, including the TOEFL iBT and a
self-assessment. The ramification inference links the use of the EPT results to its actual impact and in this study the ramification inference requires investigation of the impact of the placement decisions in a specific educational context.

For the extrapolation inference, quantitative data such as test performance data on the EPT, the TOEFL iBT, and the self-assessment were collected and analyzed using multitrait-multimethod (MTMM) analysis techniques. The findings indicated that the EPT was found to have moderate relationships with the TOEFL iBT and weak to moderate relationships with the self-assessment. The EPT showed some of the expected convergent correlations as well as discriminant correlations based on the MTMM correlation coefficient matrix as well as the factor loading parameters in a correlated trait-correlated uniqueness (CTCU) model. For the ramification inference, three types of analyses were conducted to seek support with regard to 1) test stakeholders’ perceptions of the EPT placement decisions, 2) the impact of the EPT placement on ESL students’ English learning, and 3) the relationship between ESL students’ EPT performance and their first-semester academic achievement. The interviews with test stakeholders were coded and analyzed to identify statements indicating their perceptions of the impact of the placement decisions. The qualitative findings are also utilized to help interpret the quantitative findings. Multiple paired-samples t-tests are used to investigate ESL students’ progress in the ESL courses that they were placed into. In addition, a structural equation modeling (SEM) approach was used to model the relationship among students’ performance on the EPT, ESL courses, and their first-semester GPA, mediated by individual difference constructs, such as learning motivation, academic self-efficacy, and self-regulated learning strategies.
The qualitative analyses of the interviews with four groups of test stakeholders show that the interviewed ESL students in general experienced initial frustration regarding the placement decisions, in retrospect, they understood why they were placed into ESL courses and appreciated the benefits of taking the required courses, especially ESL writing courses. The ESL course instructors were satisfied with the placement accuracy, even though occasionally they identified a few cases of potentially misplaced students in the ESL courses. The interviewed undergraduate academic advisors showed positive perceptions of the EPT and the placement decisions. They also reported observing that the majority of the ESL advisees were receptive to the EPT placement decisions.

The analyses of ESL course performance data collected at the beginning and the end of the course indicate that ESL students in Engl99L, an ESL listening course focusing on listening strategies, made statistically significant progress in terms of score gains on the same listening test administered at two time points. However, only nine out of 38 ESL students made satisfactory progress with reference to the course standard. Students in Engl101B (a lower-level ESL academic English writing course) and Engl101C (a higher-level ESL academic English writing course) did not show much progress in terms of lexical complexity, syntactic complexity, and grammatical accuracy. However, the Engl101C students on average wrote longer essays at the end of the course. Based on the ratings of the essays written in the final exams using the EPT scoring rubric, 14 out 18 Engl101B students (77.8%) and eight out of 16 Engl101C students (50%) showed satisfactory progress in these classes and were deemed ready for the next level of English study. The SEM analysis results indicate that ESL students’ EPT performance had significant and direct impact on their academic achievement. What’s more, students’ EPT
performance predicted their academic self-efficacy and affected extrinsic goal orientation. However, these motivational factors did not have direct impact on academic achievement.

The findings in this study contribute to building the validity argument for the EPT with two of the assumptions underlying the warrant for the extrapolation inference and ramification inference found supported and the other three partially supported. This findings in this study contributed to a better understanding of the score interpretation and use of the EPT at Iowa State University through constructing a validity argument. These findings shed light on the future development of the EPT and other similar English placement tests. The findings in this study as well as the research methodology can be informative for other institutions where English placement tests are used.
CHAPTER 1. INTRODUCTION

With a steady increase in the enrollment of international students at the universities in English-speaking countries, such as the United States, the United Kingdom, and Australia, the English proficiency of non-native English-speaking students has been recognized as one of the key factors contributing to their academic success (Andrade, 2006; Feast, 2002). At the university level in these English-speaking countries, English proficiency is usually measured with standardized English proficiency tests such as the Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS). As a complement to the standardized tests, English placement tests are also widely used as a post-admission testing tool to assess English proficiency or readiness in academic English of the admitted non-native English-speaking students and at the same time to help English as a second language (ESL) programs make course placement decisions accordingly (Green, 2012; Luoma, 2013).

In this chapter, I start with a brief description of the use of university-level English placement tests for ESL students in English-speaking countries and point out that English placement tests in general are under-researched especially for their impact in various educational contexts. I, then, propose the argument-based approach to validation as an appropriate way to address the need for more studies on English placement tests and introduce key concepts in this approach. Following this, I narrow down the focus to the context of my dissertation study to explain my research questions and discuss the significance of this study. I end this chapter with an overview of the structure of this dissertation.
1.1. ESL students and English placement tests

The wide use of placement tests results from a practical need for a test to measure the English of non-native English-speaking students locally (Fox, 2009). English placement tests can play an important role in facilitating English teaching and learning through grouping ESL students who share similar needs in English learning. Given their close connection with ESL courses, English placement tests could exert a direct impact on English learning and teaching in the courses (Brown, 1989; Fulcher, 1997; Green & Weir, 2004). In this scenario, ESL students as the major test stakeholders are under the immediate impact of the decisions based on the test results of English placement tests. The users of the placement tests assume that ESL students with limited English proficiency are likely to benefit from the placement decisions through taking the recommended or required ESL courses. On the other hand, ESL students with adequate English proficiency, if they are misplaced, could be affected adversely and may show low motivation in the ESL courses (Fox, 2009). Students’ perception of their English proficiency can be affected by the test results, and this in turn may influence their confidence and motivation in learning new content in English. In this sense, English placement tests in general function as gatekeepers and door openers at the same time for ESL students in local contexts.

Unlike standardized proficiency tests, most English placement tests in use are locally developed tests with comparatively limited resources, despite their significant impact on various test stakeholders, especially ESL students. Due to the limited investment of resources, there are relatively few studies on English placement tests in the field of language testing and the majority of the studies have examined the psychometric properties of the test in question (Blais & Laurier, 1995; Clark, 2004; Isbell, Nelly, & Arvizu, 2014; Kunnan, 1990, 1992; Nakamura, 2007). Apart from the psychometric focus in language testing, even fewer studies have focused on the impact
of English placement tests (Bradshaw, 1990; Green & Weir, 2004; Plakans & Burke, 2013). Overall, the scarcity of studies on English placement testing is inappropriate in view of their status as one of most frequently used language testing events, the outcomes of which have immediate impact on ESL students’ English learning and potential influence on their academic achievement. More systematic validation efforts are needed to facilitate a better understanding of the test scores and the consequences of decision-making based on the test scores.

1.2. Key concepts in argument-based approach to validation

To address the need for more studies on English placement tests, a useful framework of validity is needed to guide the research design and structure the validity argument. Earlier studies on English placement tests mainly approached validity with a single question (to what extent does the test measure what is purported to measure?) supported by three types of validity, namely content validity, concurrent or criterion-referenced validity, and construct validity (Brown, 1989; Fulcher, 1997; Kunnan, 1992; Wall, Clapham, & Alderson, 1994). More recent studies followed Messick’s unitary concept of validity (Fujita, 2005; Gere, Aull, Green, & Porter, 2010; Lee, 2005; Schmitz & del Mas, 1991) or Bachman and Palmer’s test usefulness framework (Banegas, 2013). These approaches to validation have contributed greatly to the field of language testing. However, the piecemeal validity evidence collected through these approaches tends to appear disconnected with each other. By comparison, an argument-based approach could be appropriate for validating test score interpretation and use in English placement tests because it provides a systematic way to investigate multiple linking inferences that underlie score interpretation and use so that a coherent validity argument can be constructed (Chapelle, Jamieson, & Enright, 2010; Kane, 2006, 2012). There are a small number of studies
applying the argument-based approach to validating score interpretation and use in English placement tests (Johnson, 2011; Le, 2010). This dissertation study will add more empirical evidence to help build a comprehensive and coherent validity argument for the EPT.

In this section, I will briefly introduce the following key concepts used in argument-based approach to validation: interpretive and use argument, validity argument, claims, inference, assumption, warrants, rebuttal, and backing.

Kane (2006) conceptualizes validation as a process involving “an evaluation of the extent to which the proposed (score) interpretations and uses are plausible and appropriate” (p. 17). In this process, claims about the score interpretations and uses are made explicitly in what Kane referred to as an interpretive argument, or more recently an interpretation and use argument in Kane (2013). This argument about the claims concerning the score interpretations and uses is then evaluated with empirical evidence and/or theoretical rationales, which constitute the validity argument. In this sense, the interpretation and use argument is the ‘specification’ of the proposed score interpretations and uses, whereas validity argument is the ‘evaluation’ of these proposed score interpretations and uses.

Using Toulmin’s (2003) argument model, an interpretive argument consists of claims, grounds or data, at least one inference, assumptions, rebuttals, and backing (see Figure 1.1). In argument-based validation studies, claims are the general statements or assertions about test score interpretations and uses. They are the conclusions about test-takers’ performance or the consequences of using scores based on various observations and data, which are called grounds. The linking process from grounds to a claim can be labeled as a specific inference depending on the nature of the claim. For instance, Kane (2006) lists four major inferences in an example of interpretive argument for placement systems: scoring (using a scoring rubric to assign scores to
test-takers’ performance), generalization (extending the interpretation of the observed scores to expected performance over multiple parallel test forms), extrapolation (linking test-takers’ expected performance to real-life performance in the target domain or target score), and decisions (making decisions based on the interpretation of target scores). Chapelle, Enright, and Jamieson (2008) explored six inferences in the validation project of the TOEFL iBT: Domain description (connecting performance in the target domain to the observations of performance in the test domain), evaluation (Awarding scores to the observations of test performance based on scoring rubrics), generalization (linking the scores of the observed performance to expected scores from parallel tasks and/or raters), explanation (linking expected scores to theoretical constructs which can explain test performance), extrapolation (linking the constructs to target scores or real-life performance in the target domain), and utilization (linking target scores to the uses of test score). This project will mainly follow the examples in Chapelle et al. (2008). In addition, in this study, the last in the chain of inferences in this study ends with the ramification inference, instead of the utilization inference because the ramification inference connects the use of target scores to the consequences of such use (Chapelle, Cotos, & Lee, 2015), which is critical in validating the use of a test for placement purposes. I will provide more detailed descriptions of each inference in Chapter 2.
Each of the inferences is supported by at least one warrant and each warrant entails certain assumptions. A warrant for an inference can be generally accepted rules, standard procedures, or established guidelines (Chapelle et al., 2008). For example, a warrant for the extrapolation inference in the interpretation and use argument for the English Placement Test (EPT) is that “the construct of academic language proficiency as assessed by the EPT accounts for students’ self-assessment of English proficiency in academic contexts.” The underlying assumption is that “performance on the EPT is related to students’ self-assessment of English proficiency in academic contexts” (EPT office, 2014). Assumptions underlying the warrant should be satisfied with adequate evidence or backing in order to provide support to the warrant. For example, one backing for the extrapolation inference would be a strong correlation between the EPT and the self-assessment of English proficiency. At the same time, rebuttals to the inferences should be used to indicate the conditions under which the inference cannot be made and thus rebut the claims about score interpretations and uses. Rebuttals are supported with backings consisting of evidence indicating that the proposed rebuttal is actually supported, and
the inference should not be made. One example rebuttal backing for the extrapolation inference would be the poor psychometric properties of the self-assessment instrument if there were a rebuttal specifying this as a threat to the extrapolation inference.

1.3. Context of this study and research questions

The test investigated in this study is the English Placement Test (EPT) used at Iowa State University (ISU) to determine whether the incoming non-native English-speaking students or English as a second language (ESL) students would need additional ESL instruction in English reading, listening, and writing. This EPT is administered to more than 1000 ESL students in every academic year and the test results have a direct impact on ESL students’ academic journey at Iowa State University. More information about the EPT can be found in the instrument section of Chapter 3.

This dissertation project utilized the argument-based approach to validation to address the need for a systematic effort for validating the test score interpretation and use of the EPT. Specifically, this study built on findings from other EPT-related research projects (Le, 2010; Manganello, 2011; Yang & Li, 2014) and focused on the last two inferences in the interpretation and use argument for the EPT, i.e., the extrapolation inference and the ramification inference. The extrapolation inference advances the argument from the expected scores (the scores showing adequate test reliability) to the target score or actual performance in a target domain. The ramification inference links the use of target scores to its intended impact on test stakeholders.

Four topics are listed below to link evidence to the mentioned inferences: 1) the relationship between the EPT and two external measures of English proficiency: the TOEFL iBT and ESL students’ self-assessment (extrapolation); 2) test stakeholders’ perception of the use of
test scores for placement purposes and the impact of such a test score use (ramification), 3) the impact of the placement decisions on ESL students’ English learning (ramification), and 4) the relationship between ESL students’ EPT performance and their academic achievement in light of individual difference constructs in academic self-efficacy and learning motivation (ramification). The first topic represents the key concern in extrapolation inference and the last three topics deal with the ramification inference. More detailed explanation of these two inferences as well as the warrants and underlying assumptions will be provided in the literature review section.

Multiple data were examined quantitatively and qualitatively using a variety of analytic methods, including multitrait-multimethod (MTMM) analyses, Rasch model analysis, confirmatory factor analysis, qualitative analysis of interview data, paired-samples $t$-tests, and structural equation modeling (SEM). Specifically, MTMM analyses were used to seek backing for the extrapolation inference because the results revealed the relationship between the EPT and other two measures of English, whereas qualitative analysis of interview data, paired-samples $t$-tests, and structural equation modeling were used to provide backing for the ramification inference as these methods addressed three assumptions for its warrant.

In this study, an interpretation and use argument for the EPT used at Iowa State University is developed to help lay out a research plan and to guide the choice of research methods that were conducted one step at a time. This dissertation study focused on the extrapolation and ramification inferences of the EPT for two reasons. Firstly, considering the direct impact of the EPT decisions on the ESL students, it is imperative to investigate the relationship between students’ performance on the EPT and some external criteria of English proficiency in the target domain (extrapolation), and the impact of the placement decisions based on the EPT scores on students’ English learning and academic achievement (ramification). These
two inferences go beyond the psychometric properties of the test and address the concerns over the use of test scores. Secondly, in addition to the documentation of the EPT development and regular item analysis of the EPT, there have been several validation studies conducted for the EPT, but very limited efforts have been devoted to the extrapolation and ramification inferences of the EPT validity argument. For example, Le (2010) proposed an interpretive argument for the listening section of the EPT and conducted empirical studies on the four main inferences, including domain analysis, evaluation, generalization, and explanation. Yang and Li (2013) examined the explanation inference of the EPT through an investigation of the factor structure and the factorial invariance of the reading and listening sections of the EPT. This dissertation study intends to fill in the gap with an investigation of the degree to which support can be found for the extrapolation and ramification inferences of the EPT.

1.4. Significance of this study

Considering the growing enrollment of non-native English-speaking students in the U.S. higher education as well as a strong need for English placement tests to help make placement decisions, this validation study makes a timely contribution to the field of language testing through offering insights into the score interpretation and impact of score use of a particular English placement test. Specifically, this dissertation study is significant to the field in three ways. Firstly, following the argument-based approach to validation, this study shows how the score interpretation and use of a particular English placement test can be investigated through seeking support for the assumptions that warrant the claims about the inferences of extrapolation and ramification. The external criteria of English proficiency used in this study, namely the TOEFL iBT and the self-assessment, can be highly relevant in many other institutions as the
TOEFL iBT is very likely already used for admission purposes and the self-assessment instrument developed in this study can be modified to accommodate various academic contexts. Therefore, as shown in this study, these two instruments can be used for validation purpose elsewhere. As for the inference of ramification, the assumptions underlying the warrant helped determine multiple groups of test stakeholders as well as two aspects of the impact of the EPT for a finer investigation. In addition, the warrants and assumptions listed in the interpretation and use argument for other inferences for the EPT could be useful in other validation studies if other English placement tests have similar score interpretations and uses.

Secondly, the mixed-methods design used in this study can inform language testing researchers on the use of multiple sources of data as well as multiple analytic techniques to investigate the inferences of extrapolation and ramification. In this study, the quantitative analyses included Rasch model-based analysis of the self-assessment data and the motivation questionnaire responses, correlation analyses of the multitrait-multimethod (MTMM) data, confirmatory factor analysis of the MTMM data, confirmatory factor analysis of self-assessment data as well as motivating questionnaire responses. Meanwhile, the interview data from the test stakeholders provided major evidence for the ramification inference as they pertain to the perceptions of the EPT placement decisions. These qualitative data also functioned as supplementary information to help understand the quantitative findings in this study with regard to the relationship between the EPT and the TOEFL iBT as well as the self-assessment, and the impact of the placement decisions on ESL students’ English learning and academic achievement.

Lastly, the findings of this study contribute backings needed to build the validity argument for the EPT with the evidence collected for the extrapolation and ramification
inferences at Iowa State University. More specifically, identifying the relationship between the EPT and two external measures of English (the TOEFL iBT and the self-assessment) will set a frame of reference for test score interpretations. Comparing students’ performances on these three measures will also help test developers better understand to what extent the EPT measured something similar to what is tapped by the TOEFL iBT and the self-assessment. In addition, studying test stakeholder’s perceptions of the placement decisions based on the EPT scores will inform the future development of the EPT, especially with regard to the test structure and the score reporting practice. Furthermore, a closer look at the impact of the EPT placement on students’ English learning and academic achievement will shed light on the effectiveness of the ESL program at the university as well as the relationship between English proficiency and academic achievement for ESL students. The findings in this study can be useful for other institutions which have similar English placement instruments for ESL students in English-speaking countries.

1.5. Overview of the dissertation

This dissertation consists of five chapters, including this introductory chapter. The second chapter reviews previous studies related to the two inferences, namely extrapolation and ramification in the field of language testing, as well as the studies of ESL students’ motivational constructs and academic achievement. Chapter three details the research design, participant information, instruments, data collection procedures, and data analytic procedures used to answer each of the research questions. Chapter four presents the data analysis results for each research question and discusses these results in light of findings from previous studies. The last chapter, Chapter five, summarizes the primary findings of this study and discusses the implications for
validation study design and future test development. The dissertation ends with a short recommendation for future studies.
CHAPTER 2. LITERATURE REVIEW

The goal of this study is to validate the score interpretation and use in the English Placement Test (EPT) used at Iowa State University following the argument-based approach and the focus of this study is on two of the inferences in an interpretation and use argument for the EPT, namely the extrapolation inference and the ramification inference. To lay out the theoretical background of this study, this chapter starts with a brief review of the background of the argument-based approach to validation and highlights two major inferences for English placement tests that are of interest in this dissertation study: the extrapolation inference and the ramification inference. The rest of the chapter focuses on a review of the studies on 1) the relationship between English placement tests and the scores from other English language tests that supposedly measured similar constructs (extrapolation), and 2) the impact of English language tests on students’ English learning and academic achievement (ramification). The former contains two major subtopics, namely standardized English proficiency tests and self-assessment as external criteria for evaluating English proficiency measured by English placement tests and other English language tests. The latter includes discussion of stakeholder’s perceptions of test impact of English placement tests, the relationship between of English proficiency and academic achievement, and the role of motivational factors in predicting academic achievement.

2.1. Argument-based approach to validating English placement tests in general

The section briefly reviews the validation approaches in language testing research and introduces main concepts in an argument-based approach to validation. An interpretation and use
argument for the English Placement Test used at Iowa State University is then presented with special attention to the inferences of extrapolation and ramification.

2.1.1. Brief introduction to argument-based approach to validation

Validity theory has undergone several stages in the last few decades, as embodied in four validation approaches in the field of language testing according to an analysis by Chapelle and Voss (2013), who name four states: 1) one question and three validities, 2) evidence gathering, 3) test usefulness, and 4) argument-based approach. The first approach centers on the question “to what extent does a test measures what it purports to measure?” and identifies three types of validity to be investigated: content validity, concurrent and criterion-referenced validity, and construct validity. This approach has been criticized for its view of validity as an inherent property of a test (Messick, 1989). However, the traces of this approach are still visible and the counterpart of each validity type remains important in other approaches or frameworks.

The evidence-gathering approach to validation reflects the unitary view of validity as proposed by Messick (1989) and acknowledged in the Standards for Educational and Psychological Test (AERA, APA, &NCME, 1999). In this approach, construct validity is presented as the core of validity and validation efforts aimed to construct “an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1989, p. 13). Another important contribution of this approach is its inclusion of consequences of test score use as a part of validity. One limitation of this approach is its lack of definite guidelines to inform researchers and practitioners how much evidence is adequate. The third approach of test usefulness was introduced specifically to language testing research by Bachman and Palmer (1996) and partially addressed the limitation of the second approach by operationalizing
validation in six aspects of ‘test usefulness’: construct validity, reliability, authenticity, interactiveness, impact, and practicality.

The argument-based approach to validation represents a newer trend in validating test score interpretation and use in language testing research (Chapelle & Voss, 2013). According to Kane (2006), validation is a “process of evaluating the plausibility of proposed interpretations and uses” (p. 17). Therefore, two key steps are involved in an argument-based approach to validation: 1) construction of an interpretation and use argument, which frames the proposed interpretations and uses of the test scores and 2) evaluation of the plausibility of the interpretation and use argument (Kane, 2013). In an interpretation and use argument for a test, a series of inferences linking test performance data to the uses of test scores and consequences are proposed. For example, Kane (2006) discussed four inferences for a placement test: scoring, generalization, extrapolation, and decision making. Bachman and Palmer’s framework of Assessment Use Argument (AUA) consists of four general inferences about assessment records, score interpretations, decisions, and intended consequences and each general claim has more fine-grained claims that require backing and/or rebuttals. In a comprehensive validation study of TOEFL, Chapelle, Enright, and Jamieson (2008) explored six inferences, including domain definition, evaluation, generalization, explanation, extrapolation, and utilization. These inferences will be further explained in next section. These linked inferences are stated in the interpretation and use argument for a particular test and then the validation process consists of investigation that seek backing to support the inferences. Following Toulmin’s argument model, each inference in an interpretation and use argument is elaborated with a detailed claim, whose plausibility depends on the backing for its underlying warrants and their assumptions as well as
the backing for potential rebuttals (see Figures 2.2–2.7). This outcome of the validation is a structured validity argument.

According to Chapelle, Enright, and Jamieson (2010), the advantages of this argument-based approach to validation are in four aspects: “framing the intended score interpretation, outlining the essential research, structuring research results into a validity argument, and challenging the validity argument” (p. 3). Compared to the concepts of validity and validation in Messick (1989) and the 1999 AERA/APA/NCME Standards for Educational and Psychological Testing, an argument-based approach to validation provides a more accessible framework for educational measurement practitioners to design validation studies based on an interpretive argument or interpretation and use argument and thus enables them to structure a validity argument in a more systematic and coherent way. Recent years have witnessed an increase in the application of the argument-based approach to validation in various language testing projects (Aryadoust, 2009; Bachman & Palmer, 2010; Chapelle et al., 2008; Chung, 2014; Jia, 2012; Jun, 2014; Kadir, 2008; Llosa, 2008; Voss, 2012). As observed by Chapelle and Voss (2013), the number of argument-based validation studies is on the rise in the last few years (2006-2011).

2.1.2. An interpretation and use argument for the English Placement Test (EPT)

The scores from English placement tests are usually claimed to be indicative of students’ English proficiency levels in an academic context and thus can be used to make decisions about academic ESL course placement. The placement decisions based on English placement test scores also reflect an underlying belief that adequate English proficiency is necessary for ESL learners to achieve academic success at a university. However, the intended score interpretations and impact of the placement decisions of English placement tests in general are still an under-researched area (Green, 2012) and more validation efforts should be directed to constructing
specific validity arguments. Since test score interpretation and use are test-specific or dependent on the contexts where the tests are used, in this section, I use the English Placement Test (EPT) used at Iowa State University as an example test for English placement purpose and present such an interpretation and use argument for the EPT.

Following the structure of the interpretive argument for the new TOEFL outlined by Chapelle et al. (2008), the overall structure of the interpretation and use argument for the EPT is constructed, as depicted in Figure 2.1. This interpretation and use argument for the EPT specifies the inferences involved in the proposed claims about test score interpretation and use. This interpretation and use argument could be modified for validating score interpretation and use of other similar English placement tests in other contexts, but in this study, I am working with the specific validity argument for test interpretation and use at Iowa State University.

![Figure 2.1. An interpretation and use argument for the English Placement Test (EPT) (adapted from Chapelle et al., 2008).](image)

Similar to the interpretive argument for the TOEFL in Chapelle et al. (2008), the interpretative argument for English placement tests covers six inferences starting from domain description and concluding with ramification. This series of inference, which explicitly contains
explicit statement about test score interpretation and use, also guides a systemic validation to build a validity argument: identifying target domain (domain description), scoring student’s performance on the test with pre-determined scales or answer keys (evaluation), estimating the reliability of the test (generalization), accounting for students’ performance from theoretical perspectives (explanation), matching scores with actual performance in the target domain (extrapolation), and establishing the impact of decisions based on the scores (ramification). Each inference will be briefly introduced in the following paragraphs.

2.1.2.1. Domain description inference

For the English Placement Test, the target domain is usually the university contexts in English-speaking countries. The target domain in the case of the EPT is the academic contexts at Iowa State University. The target domain is a starting point in a validation process and the domain description inference links the knowledge, skills, and abilities required in the target domain to the test-takers’ performance on a test, which is supposed to consist of assessment tasks representative of the target domain.
Figure 2.2. The domain description inference for the EPT (adapted from Chapelle et al., 2008).

As shown in Figure 2.1, the warrants to the Domain description inference are that the assessment tasks that are representative of academic domains are identifiable and the tasks in the academic domains can be modeled. To provide supports to the warrant, domain analysis and task modeling should be carried out. For example, constructing test task frameworks can be a useful approach to characterizing tasks used in a target domain, thus helping test developers create test tasks that model the tasks in the target domain (Jamieson, Eignor, Grabe, & Kunnan, 2008). The rebuttal to the claim could be a lack of task authenticity due to a failure of modeling essential contextual factors in the target domain language use. Making the claim from this Domain
description inference paves the way to other following inferences by providing detailed and actionable definition of the target domain.

2.1.2.2. Evaluation inference

The second inference in the interpretation and use argument for the EPT is the evaluation inference which focuses on the scoring process of the EPT through linking ESL students’ observed performance on the EPT to the EPT scores or grades. As shown in Figure 2.3, the warrants to the claim about observed scores are that 1) the testing conditions are standardized and appropriate for test-takers to perform to their best, 2) the scoring rubric and scoring conditions are appropriate for raters, and 3) the items and the whole test exhibit good psychometric qualities. Accordingly, three types of studies could be conducted to either support or refute the warrants: evaluation of the comparability of test administration conditions, investigation of rater’s use of scoring rubric, and statistical analysis of items and the whole test.
2.1.2.3. Generalization inference

The generalization inference advances the interpretation and use argument by linking the observed scores on the EPT to expected scores which would be obtained if the same ESL student took multiple parallel forms of the same test and were rated by different raters. The claim one wants to be able to make is that the observed scores could be generalized to other parallel test forms or parallel scoring sessions. The warrant is that the observed scores are reliable enough to be treated as expected scores. Two assumptions underlie this warrant are that 1) a sufficient number of items or tasks are used in the test to provide a reliable estimate of test-takers’
performance and 2) different versions of the same test are made parallel based on test specifications.

Figure 2.4. The generalization inference for the EPT (adapted from Chapelle et al., 2008).

The Generalization inference links the observed scores verified in the Evaluation inference to expected scores. As shown in Figure 2.4, the warrants include 1) a sufficient number of tasks are used in English placement tests, and 2) parallel tasks and test forms are created based on well-defined test specifications and scoring rubric. The Generalization inference is similar to the traditional concept of reliability. Therefore, reliability analysis is one of the major methods to
provide backing to the warrant. In addition, quantitative and/or qualitative methods can be used to investigate the comparability of parallel tasks and test forms.

2.1.2.4. Explanation inference

The Explanation inference links the expected scores to theoretical explanations of test-takers’ performance. In this sense, it is closely related to the traditional concept of construct validity, which also highlights theoretical accounts of what is measured in a test.

![Diagram of Explanation Inference](image)

Figure 2.5. The explanation inference for the EPT (adapted from Chapelle et al., 2008).

The warrants to the Explanation inference include 1) the internal structure of English placement tests is consistent with a theoretical view of language proficiency, 2) the test-taker’s
use of knowledge and strategies are in line with those described in test specifications. Two types of analysis can provide backing evidence to the warrants: correlation analysis of test components and confirmatory factor analysis of test structure.

2.1.2.5. Extrapolation inference

In the interpretative argument for English placement tests, the extrapolation inference links the construct of language proficiency as represented by the scores or levels of sub-skills (reading, listening, and writing skills) to the target scores, which represent the quality of performance in the real-world domain of interest. The diagram in Figure 2.6 presents the extrapolation inference with its grounds, claims, assumptions, rebuttal, and backing.

As shown in Figure 2.6, the claim warranted by the extrapolation inference is that the expected scores of the EPT reflect test-takers’ the target scores, in other words, test-takers’ actual English proficiency in academic contexts at the university. The assumptions underlying this inference include that the constructs of academic language proficiency as assessed by the EPT account for the quality of linguistic performance at that university. Typical backing for the assumptions include criterion-related validation studies, in which an external criterion that represents test-takers’ performance in targeted domain is employed for comparison. Therefore, a key step in evaluating the extrapolation inference is to identify an appropriate external criterion and use it as a reference to compare test-takers’ performance on the EPT. Possible external criteria are concurrent measures of English proficiency, such as standardized English proficiency tests, student self-assessment, teacher evaluation, in-class tests, and so on. I will review the studies on standardized tests and self-assessment as potential external criteria in the dissertation study and leave out the in-class test because there are very few, if any, studies on the relationship between in-class tests and English placement tests.
2.1.2.6. Ramification inference

In the interpretative argument for English placement tests, the ramification inference allows the argument to advance to the claim that the decisions based on the placement test scores are conducive to learners’ English learning and academic success at the institution. As shown in Figure 2.7, the assumptions underlying the warrant are 1) the decisions of ESL course placement are justifiable and comprehensible to test stakeholders; 2) the decisions are beneficial for
learners’ improvement of their academic English proficiency; 3) The decisions are beneficial for learners’ achievement of academic success at the university. The backing for the aforementioned assumptions can be collected using qualitative methods, such as interviewing test stakeholders for their perceptions of the placement decisions, as well as quantitative methods, such as analyzing students’ performance in the ESL courses and analyzing the predictive power of the EPT placement decisions on English proficiency improvement and academic achievement with an involvement of motivational factors and self-regulated learning variables.

*Figure 2.7. The ramification inference for the EPT (adapted from Chapelle et al., 2008).*
In this dissertation study, the focus of the dissertation study is on the last two inferences: the extrapolation inference and the ramification inference. The following sections will review the studies that are relevant to these two inferences and these are somewhat related to the traditional notion of criterion-referenced validity: concurrent validity and predictive validity.

2.2. Relationship between the EPT and other measures of English proficiency

The relationship between the English language test of interest and other English language tests that measure similar constructs constitutes a main piece of backing for the extrapolation inference because such a relationship provides a reference for test score interpretation, especially in the case that other tests tap into the constructs in the target domain. Identifying appropriate reference tests and setting up reasonable expectation of the relationship are therefore very important in investigating the extrapolation inference. This section reviews the studies relying on external criteria of English proficiency to validate the interpretation of test scores in English language tests and English placement tests in particular. Two general categories of external criteria are reviewed, namely standardized English proficiency tests and self-assessment as types of non-test indicators.

2.2.1. Standardized English proficiency tests as external criteria

Criterion-referenced evidence for validity has been a major type of evidence used in language testing research (Bachman & Palmer, 1996). A typical practice of this kind is to correlate the scores on a test under investigation with those on a well-established test, such as TOEFL or IELTS, assuming that both the target test and the reference test measure a set of similar, if not the same, constructs (Riazi, 2013).
In a concurrent validation of the Pearson Test of English Academic (PTE Academic), a relatively new English proficiency test developed by Pearson, Riazi (2013) compared the performances of 60 international students on both PTE Academic and IELTS as a criterion test as well as their GPA in an Australian university. It was found that the correlation between the overall scores from PTE Academic and IELTS was .824 and the correlations for individual subskills ranged from .661 to .723. Ockey, Koyama, Setoguchi, and Sun (2015) studied the relationship between the speaking test in the TOEFL iBT and oral English tasks used in a Japanese university, including group oral discussion, picture and graph description, and prepared oral presentation tasks. Two hundred twenty-two Japanese university students took the TOEFL iBT speaking test and the oral tasks. Their overall scores on these measures were correlated and the disattenuated correlation coefficients ranged from .63 to .76. All of the analytic scores for the university oral tasks were also found to be significantly correlated with the TOEFL iBT speaking scores, ranging from .50 to .74. Based on the moderate to high correlations, Ockey et al. (2015) claimed that the TOEFL iBT speaking scores could be interpreted as “good overall indicator of academic oral ability” (p.40).

This type of correlational analysis for criterion-referenced evidence is not uncommon in validation studies on English placement tests. For example, there are a few studies investigating the relationship between English placement tests and the standardized English proficiency tests. Manganello (2011) found that the TOEFL iBT scores had a moderate correlation with the EPT administered from fall 2009 to spring 2011 at Iowa State University, following Cohen (1988)’s conventions of effect size interpretation (small: $r = .10$, moderate: $r = .30$, large: $r = .50$ or above). He reported that the correlation coefficient (Pearson’s $r$) for the reading section between the two tests was .363 and the correlation coefficient for the listening section between the two
tests was .413 ($n = 338$). The correlation coefficient (Spearman’s rho) for the writing section between the two tests was .317 ($n = 324$). The correlation magnitudes reported in Manganello (2011) indicate that the constructs measured in these two tests may be different to some extent. It is noteworthy that several factors could affect the correlations between two tests even though they may have measured something very similar. For example, the reliability of each test has direct impact on the magnitude of the correlation coefficients as the theoretical upper limit of the correlation is determined by the square root of the reliability of either measure (Bachman, 2004, p. 94).

These findings in Manganello (2011) echoed the questioning of the common practice of using TOEFL and other standardized English proficiency tests for placement purposes, even they are not explicitly claimed to be used for those purposes (Fox, 2009; Kokhan, 2012; 2013). For example, Kokhan (2012) studied the possibility of using TOEFL scores for university ESL course placement decisions at the University of Illinois at Urbana-Champaign. She found that the correlation coefficients between the TOEFL scores and the scores on the local English Placement Test varied when the lag between the TOEFL test and the English Placement Test was taken into consideration. Overall, the highest correlation coefficient was below .4 in the case where TOEFL was taken most recently by the students. However, with a wider interval gap between the TOEFL and the English Placement Test in time, the correlation coefficients became even weaker. Considering the potential impact of misplacement using the TOEFL scores, Kokhan (2013) made an explicit argument against using standardized test results from SAT, ACT, and the TOEFL iBT for placement purposes at the University of Illinois at Urbana-Champaign. Kokhan’s argument is also supported by Fox (2009)’s study on the impact of a top-down policy about using TOEFL and IELTS scores for placement purpose in an English for Academic Purpose (EAP) program at
a Canadian university. Since the standardized English proficiency tests are not designed for placement purposes, one of the consequences of this policy was a noticeable number of misplaced and at-risk ESL students identified by the EAP program teachers. In addition, Fox (2009) reported that the concordance relationship among scores on different standardized English proficiency tests and an in-house English placement test used in the EAP program was not stable enough to ensure an accurate placement using the standardized proficiency tests alone for that purpose. One of the possible reasons for this consequence is that the scores on the standardized English proficiency tests do not “test the incremental developmental changes that define learning within the EAP program” (p. 38).

In view of the existing studies on the relationship between standardized English proficiency tests and local English placement tests, I acknowledge that the standardized proficiency tests can provide key English proficiency information, but are not necessarily appropriate for placement purposes. For this reason, I expect significant, but relatively weak to moderate correlations between the TOEFL iBT and the English Placement Test (EPT) in this study. Apart from the TOEFL iBT, I intend to include a self-assessment as another external criterion of English proficiency in seeking the backings for the assumptions in the extrapolation inference. The next subsection will review self-assessment-related studies in the field of language testing.

2.2.2. Self-assessment as an external criterion

Self-assessment of second or foreign language abilities is defined as “language learners’ evaluation of his or her own language skills, usually in connection with a language course or part of other forms of language assessment” by Luoma (2013). Self-assessments are usually made with explicit reference to clear learning goals in a target domain. Therefore, self-assessments
could be another potential external criterion used in the study of the extrapolation inference. In addition, self-assessments can be a reliable learner-directed measure of foreign language or second language proficiency that brings test-takers’ voices to the validation process (LeBlanc & Painchaud, 1985). As for the efficacy of self-assessment by naive or untrained learners, research shows that self-assessment is feasible even for young learners. For example, the correlation between self-assessment of reading skills made by 5271 third graders in Sweden was found to be .58 with a standardized test of reading skills and .59 with teacher judgment (Johansson, 2013). The self-assessment reported in that study contained four statements on a 4-point Likert scale. Johansson (2013) also found that gender and socioeconomic status did not affect third-graders’ self-assessment, which, together with the correlation coefficients, leads to his claim that for these young learners, self-assessment of reading literacy skills was a reliable assessment tool.

In terms of data collection and accessibility, self-assessment is easier to use and could reach out to a larger number of targeted participants with a set of uniform or standardized tasks/items and/or rubrics, compared to teacher evaluation and peer assessment. Another advantage of using self-assessment is that it is convenient and less time-consuming to ask students to directly reflect on or evaluate their abilities in a given context (Coronado-Aliegro, 2006). Self-assessments are usually administered in a low or non-threatening environment with minimal consequences. As mentioned in Powers et al. (2009), self-assessment tends to be more valid in the cases where participants have “no incentive to intentionally distort their reports” (p. 15).

Self-assessment has been recognized as an important instructional and learning instrument in the field of education, where the major interests are the facilitative effects of self-assessment on learning outcomes and the relationship between self-assessment and other
individual difference constructs, such as self-regulated learning strategies (Kostons, van Gog, & Pass, 2012; Lew, Alwis, & Schmidt, 2010; Ross, 2006). The importance of self-assessment for learning has been widely acknowledged for its positive effect on learner awareness and learning outcomes (Luoma, 2013). Self-assessment can be made in multiple formats, such as open-ended questions, checklists, and can-do statements with a Likert-scale. Self-assessment also has its value as an alternative to some exiting tests or as a tool to validate a test (Alderson, 2005; Engelhardt & Pfingsthorn, 2013; Hellekjær, 2009; Lee & Greene, 2007).

The utility of self-assessment of English skills has been explored mainly via correlational analyses with other measures, for example, scores on standardized English proficiency tests and teacher ratings. Overall, the findings about the correlation between self-assessment tools and other measures are promising, although the magnitude of correlation coefficients varies from study to study depending on the item format and specificity of item content (Brantmeier, 2006; LeBlanc & Painchaud, 1985; Luoma, 2013; Ross, 1998). LeBlanc and Painchaud (1985) used a planned self-assessment questionnaire as a placement tool, which contained 60 “can-do” statements with reference to specific situations. They found that the self-assessment tool produced high quality results and placed students in a similar way as the standardized tests did. Malabonga, Kenyon, and Carpenter (2005) investigated the relationship between university students’ performances on a self-assessment and a computerized oral proficiency test of foreign language. It was found that 98% of the students in that study could successfully use the self-assessment to select the test tasks that were appropriate to their foreign language proficiency levels. In addition, the correlation between the self-assessment and teacher ratings of oral proficiency ranged from .74 to .81. Strong-Krause (2000) proposed a use of self-assessment for
ESL placement purposes. Other studies also suggested or examined the use self-assessment for similar purposes, especially in writing assessment (Gere et al., 2010).

Considering self-assessment as an instrument to elicit individuals’ responses with reference to relevant real-life tasks, a series of self-assessment-based validation efforts has been made in recent years, especially for the Test of English for International Communication (TOEIC) (Powers, Roever, Huff, & Trapani, 2003; Powers, Kim, & Weng, 2008; Enright, Bridgeman, Eigorn, Lee, & Powers, 2008; Powers, Kim, Yu, Weng, & VanWinkle, 2009; Ito, Kawaguchi, & Ohta, 2005). As Powers et al. (2008) pointed out, “one kind of evidence that has proven useful in elucidating the meaning, or validity, of language test scores has come from examinees themselves, in the form of self-assessments of their own language skills” (p. 8). In Powers et al. (2008), 49 can-do statements about English reading and listening in everyday life activities were administered to 7,292 test-takers in Japan and 3,636 test-takers in Korea immediately after the TOEIC test. It was found that the self-assessment had a relatively strong correlation with the TOEIC scores, ranging from .47-.53. In another research report, Powers et al. (2009) attempted to link test-takers’ performance on the speaking and writing sections of the TOEIC test with test-takers’ self-evaluation of real-life activities. Forty can-do statement about speaking tasks and 29 can-do statements about writing tasks were administered to 2,947 test-takers in Japan and 867 test-takers in Korea. The correlation coefficients between the self-assessment and the TOEIC scores ranged from .52 to .54. Powers et al. (2008) concluded that the use of self-assessment provided evidence for the linkage from test performance to test-takers’ performance in a variety of English-related activities in real life and that the can-do statements in the self-assessment can be “reasonably trustworthy validity criteria” (p. 12). Ito et al. (2005) studied 8,386 Japanese company employees’ self-assessment of functional job performance and
compared the self-assessment with their TOEIC scores. Their self-assessment consisted of 65 five-point Likert scale can-do statements related with job activities in seven situational domains. It was found that the self-assessment was highly correlated with the TOEIC scores, with the correlation coefficient ranging from .62 to .71.

Self-assessment was also used to represent test-takers’ real life experiences in an academic context in a recent criterion-referenced validation study of the General English Proficiency Test (GEPT) in Taiwan (Weir, Chan, & Nakatsuhara, 2013). However, the relationship between the GEPT test and test-takers’ self-assessment was not reported due to some technical issues in their study. Instead, Weir et al. (2013) focused on the relationship between the GEPT test and the IELTS test as well as that between the GEPT test and some real life academic performance on course assignments and examinations.

In a validation study of the TOEFL iBT, a self-assessment, along with academic placement and instructor’s ratings, was used as a piece of evidence for the extrapolation inference (Enright, Bridgeman, Eignor, Lee, & Powers, 2008). Using confirmatory factor analysis, Enright et al. (2008) identified four factors corresponding to the four sub-skills (reading, listening, speaking, and writing) in the self-assessment. The four factors in the self-assessment were found to have a moderate and positive correlation with test-taker’s performance on both TOEFL PBT and the prototype measures of the TOEFL iBT, with the correlation coefficients ranging from .30 to .62. Enright et al. (2008) regarded the magnitude of correlation to be “high” and “similar in magnitude to other test-criterion relationships” (p. 178).

However, like standardized English proficiency tests, self-assessment is not without problems when used as an external criterion for validation purposes. There are several factors that can influence individual’s self-assessment, including the characteristics of self-assessment
items and personal traits. Zell and Krizan (2014) in their meta-synthesis of 22 meta-analytic studies on self-assessment in various fields identified five moderators between self-assessment and other measures: performance domain, academic discipline, task objectivity, task familiarity, and task complexity. In terms of task characteristics of self-assessment, Butler and Lee (2006) compared off-task and on-task self-assessment tools in measuring the oral English proficiency of the fourth grade and sixth grade Korean learners of English. The two types of self-assessment were similar content, but differed with regard to the time of administration and reference to learning tasks. It was found that on-task self-assessment had a higher correlation with general proficiency test scores and teachers’ assessment. In addition, on-task self-assessment appeared to be less prone to the influence of attitudinal and personality factors. Their findings about the differences between the two types of self-assessment echo with the research on the effect of task specificity in self-assessment (Oscarson, 1997; Pajares, 1996; Zell & Krizan, 2014; Strong-Krause, 2000). In a meta-analysis of self-assessment studies in the field of language testing, Ross (1998) identified that domains could be a factor contributing to the variability of the correlation between self-assessment and other measures. For example, self-assessment of receptive skills (reading and listening) tended to have a higher correlation with other outcome measures in the meta-analyzed studies than that of productive skills (speaking and writing).

Of course, self-assessment instruments are not without problems. The accuracy in self-assessment in general has been questioned as self-perceptions are sometimes found to be “off the mark”, conflicting with external judgment in certain domains (Zell & Krizan, 2014, p.112). This could be accounted for by research participants’ experience in evaluating themselves and the influence of participants’ social desirability, especially in sociological and psychological research.
Overall, comparing the scores from English placement tests with those from standardized English proficiency tests and self-assessments should shed light on the plausibility of the extrapolation inference; even they might not be optimal external criteria of the English proficiency in the target domain.

**2.2.3 Summary and research question 1**

The studies on the relationship between English placement tests and standardized English proficiency tests indicate that only weak to moderate correlation coefficients existed and thus suggested a limited utility of the standardized English proficiency tests in validating the interpretation of scores from English placement tests. On the other hand, studies on self-assessment reveal that tailored self-assessment appears to be a promising instrument for validation purposes. Considering the inherent limitations of self-assessment as well as possible discrepancy in the construct measured by self-assessment and English placement tests in general in this dissertation, I did not intend to use either one as the sole proxy of English proficiency. Instead, I will have multiple indicators of English proficiency in my study, including standardized English language tests. Therefore, the first research question is raised as below for the extrapolation inference.

1) To what extent are ESL students’ EPT performances related to two external criteria: students’ self-assessment of English use and their TOEFL iBT scores?

(Extrapolation)

The first research question addresses the extrapolation inference of English placement tests by focusing on the relationship between the English Placement Test and relevant criteria of English proficiency in the target domain. The criteria in this study will be operationalized as
2.3. Impact of language tests

Corresponding to the ramification inference in the interpretive argument for English placement tests, this section reviews the studies on three aspects of the impact of the placement decisions in English placement tests in general: test stakeholder’s perceptions of the placement decisions, the impact of the placement decisions on ESL students’ English learning, and the impact of the placement decisions on ESL students’ academic performance.

2.3.1. Test stakeholder’s perceptions of impact of English placement tests

Unquestionably, the placement decisions based on students’ performance on English placement tests are important for test stakeholders: test-takers, ESL instructors, content course instructors as well as academic advisors or other university staff. Test stakeholders’ voices should be heard in a validation of English placement tests (Bradshaw, 1990; Rea-Dickins, 1997). Studying test stakeholder’s perceptions of a test could yield critical information for a validity argument (Winke, 2011). However, there are very few studies involving test stakeholders in their investigation of the impact of the placement decisions in the EPT. Stressing the consumer validity of an English placement test in the UK, Bradshaw (1990) used a questionnaire to examine the reactions of test-takers and teachers to the test in nine dimensions: time availability, clarity of instructions, familiarity, nervousness, perceived test difficulty, test anxiety, fairness, interest, and pleasantness. The participants included 141 Italian students, 27 Spanish students, and 21 teachers. Bradshaw (1990) found that the C-test (an integrative instrument requiring test-takers to restore missing parts of words) of test measuring was negatively perceived by the
students, whereas the multiple-choice questions and the reading were more positively perceived as valid by teachers, compared with students’ perceptions. Moreover, lower scoring students were less positive toward the test than higher scoring students were even though they were similar in viewing the test difficulty, fairness, pleasantness, and interest. The findings in this study were only indirectly related to the consequences of using the test scores, but they contributed to decisions about test development.

Some of findings in Bradshaw (1990) are also noticed in a validation study of an English placement test used at a private Japanese university. Fujita (2005) took six approaches to collecting validity evidence, with the last one being a questionnaire about test consequences. In Fujita’s study, 2200 incoming students and 229 freshman English teachers responded to the surveys in 2002 and 2003. It was found that the students in higher-level courses were more likely to feel satisfied with the placement than the students in lower level courses. In addition, teachers tended to hold a slightly more positive attitudes toward the placement decisions than the students did. This piece of qualitative evidence about test consequence was used along with quantitative evidence to validate the score interpretation and use in the Japanese university. It is noteworthy that the qualitative studies mentioned above focused on test stakeholders’ general perception of the test, but did not address the impact of the placement decisions on students’ English learning and academic performance.

Recently, there have been some sponsored qualitative studies on test stakeholders’ perceptions of IELTS. Using surveys and individual interviews, Coleman, Starfield, and Hagan (2003) investigated how students and university staff perceived the IELTS test used at the universities in Australia, the UK, and China. Through analyzing responses from 624 respondents to a survey and 37 interviews, Coleman et al. (2003) found that students in general had a positive
view of the IELTS test and they were satisfied with the predictive nature of the cut-off score requirement. On the other hand, the university staff wished to raise the standards of the IELTS scores for admission purposes. In addition, Coleman et al. (2003) noticed some “unprincipled” uses of the IELTS scores, such as admitting students whose IELTS scores were below the minimum requirement (p.163). Coleman et al. (2003)’s findings provided important evidence about the validity of score interpretation from the test stakeholder’s perspectives. Hyatt (2013) conducted a qualitative study examining test stakeholders’ perceptions of using IELTS for admission purposes in the UK. Hyatt (2013) collected qualitative data through 100 surveys delivered via email and 12 follow-up telephone interviews with academic and administrative staff in 15 higher education institutions in the UK. He reported that most of the interviewees believed that the IELTS scores were a “very useful indicator” of the level of academic English required at higher education institutions in the UK (p. 857). Meanwhile, the interviewees acknowledged a significant tension between the minimum scores for admission and the pressure to recruit international students, which entails a stronger need for post-entry English language support in these institutions. Hyatt also noticed that the interviewees had limited knowledge about the content and process of the IELTS test and called for a better communication between the test developers and the university staff.

The qualitative evidence collected in these studies is useful in validating the score interpretation. However, they did not delve into the impact or consequences of the test score use. This dissertation study will cover not only the test stakeholders’ perception of the placement decision, but also their perception of the impact of the decisions on students’ English learning and academic achievement.
2.3.2. The relationship between of English proficiency and academic achievement

The ramification inference could be evaluated with evidence of a test score’s prediction of students’ English learning outcome and/or academic achievement. However, there are very few studies, if any, on the impact of English tests on English learning, and the most predictive validation studies focused on standardized English proficiency tests such as TOEFL and IELTS.

Academic achievement of the English language learners in the U.S. educational system has attracted much attention in the secondary schools partly due to the No Child Left Behind Act (Ardasheva, Tretter, & Kinny, 2012; Duran, 2008; Genesee, Lindholm-Leary, Saunders, & Christian, 2005; Guglielmi, 2008). By comparison, there is much less attention directed to ESL learners at tertiary institutions (Bretag, 2007; Phakiti, Hirsh, Woodrow, 2013; Vinke & Jochems, 1993).

Usually academic achievement is seen as “evidence of learning, which may be measured by successful completion of course requirements, grade point average (GPA), satisfactory academic standing or retention” (Andrade, 2006, p.134). GPA is arguably the most frequently used indicator of academic achievement even if it may not be the best criterion measure, psychometrically. Sadler (2009) cautioned that there are four threats to grade integrity: random error, bias, contamination of the object graded, and confusion in the grading principles (p. 812). Admittedly, GPA is one of the most accessible and comparable indicators of academic achievement.

Research on the relationship between standardized English proficiency tests such as TOEFL and IELTS, and ESL students’ academic achievement, typically operationalized as GPA, suggests that these tests usually have a low predictive power and there are many other personal factors that may contribute to students’ academic achievements. This type of predictive study is
usually motivated by a concern over the relationship between ESL students’ English proficiency and their academic achievement. For example, Hill, Storch, and Lynch (1999) compared the predictive validity of TOEFL and IELTS at Australian universities. Hill et al. (1999) found that IELTS was comparatively a better predictor of the GPA and TOEFL only had a weak relationship with GPA. In a meta-analysis of empirical studies on relationship between the TOEFL and GPA as well as course completion published in 1987-2009, Wongtrirat (2010) confirmed a weak predictive power of TOEFL and thus suggested admission offices make cautious use of TOEFL scores.

The weak relationship between English proficiency and academic achievement is also captured by Cho and Bridgeman (2012), who studied 2594 undergraduate and graduate students in four general disciplines at ten American universities for the relationship between the TOEFL iBT scores and GPA. It was found that the disattenuated correlation coefficients for graduate students ranged from .22 to .41 in four disciplines (Business, Humanities and Arts, Science and Engineering, and Social Science), while the disattenuated correlation coefficients for undergraduate students ranged from .19 to .41. Cho and Bridgeman (2012) also reported a general pattern shown in expectancy graphs, that is, students with higher TOEFL iBT scores tended to obtained higher GPAs.

Using self-reported data, Berman and Cheng (2010) investigated the relationship between perceived difficulty in academic English and ESL students’ academic achievement in a Canadian university. They developed a questionnaire to elicit perceived difficulty in academic English skills from both native English-speaking (NES) students and non-native English-speaking (NNES) students, and then associated their responses to self-reported TOEFL scores, self-reported GPA, and demographic traits such as first language, gender, age, and field of study. A
total of 186 students responded to the questionnaire. Different difficulty levels of the required academic English skills were reported by NES students and NNES students. In other words, NES and NNES students faced somewhat different challenges in using academic English skills. Furthermore, Berman and Cheng (2010) found that the perceived difficulties in academic English skills seemed to have negative impact on NNES graduate students, but not on NNES undergraduate students, who reported similar academic performance with their NES undergraduate peers. Their main explanation for these different impact of academic English skills on academic achievement for NNES graduate and undergraduate students is that graduate level studies had higher English language demands.

The aforementioned studies on the relationship between English proficiency and academic achievement are revealing, but at the same time may have suffered from some typical limitations in predictive validation studies (Cho & Bridgeman, 2012). For example, Graham (1987) identified four challenges in predictive validation studies, including the operationalization of academic achievement, the quality of the English proficiency measure, interpretation of correlation coefficients, and other confounding variables. Other possible factors that account for the weak relationship between English proficiency measures, such as TOEFL and IELTS, and GPAs, include “cultural capital,” defined as “attitudes, dispositions, skills to navigate the (educational) system, academic preparation, disciplinary knowledge, and educational qualifications” (Chang & Kanno, 2010, p. 674). The “cultural capital” may be more influential for graduate students who have already mastered the fundamental knowledge in their own fields of study, as exemplified by four non-native English-speaking doctoral students at an American university in Chang and Kanno (2010). In addition, Daller and Phelan (2013) pointed out that the
issue of “truncated samples” in predictive validation studies may have affected the relationship between English proficiency and academic achievement.

In the following sub-sections, I will turn to the individual difference constructs as major confounding factors affecting the relationship between English proficiency and academic achievement and review the studies on the roles of motivational factors in academic achievement.

2.3.3. The roles of motivational factors in predicting academic achievement

Existing predictive studies show that standardized English proficiency tests, such as IELTS and TOEFL, have a weak to moderate correlation with academic achievement, which is usually operationalized as students’ grade point average (GPA) (Ardasheva, Tretter, & Kinny, 2012; Cho & Bridgeman, 2012; Lee & Greene, 2007). As pointed out by Phakiti et al. (2013), the correlational analyses only provide limited information about the predictive validity when the effect of non-linguistic factors is overlooked. Admittedly, many factors can contribute to ESL learners’ academic achievement (Cho & Bridgeman, 2012). For example, Phakiti et al. (2013) reported that motivation, self-regulation, self-efficacy, and English proficiency all played important roles in students’ academic success in an Australian university. The findings in Phakiti et al. (2013) are in line with other studies on the relationship between motivational factors and academic success. From a socio-cognitive perspective, the individual difference constructs of interest include the following interrelated aspects: self-regulated learning strategies, academic self-efficacy, learning motivation, and anxiety for English use (Phakiti et al., 2013; Zimmerman & Schunk, 2008).

The constructs of motivation in this dissertation study are closely connected with the concept of self-regulated learning (SRL), which is defined as “the degree to which students are
metacognitively, motivationally, behaviorally active participants in their own learning process” (Zimmerman, 2008, p.167). Dörnyei and Skehan (2003) posited that self-regulation and motivation are “inextricably bound together because they both concern the antecedents of increased academic achievement” (p. 612). In this sense, SRL, as a multidimensional construct for learning in general, became an umbrella term, covering metacognitive, motivational, and behavioral constructs or aspects in learning. In a discussion of the relationship between motivation and SRL, Zimmerman (2008) claimed that motivation could serve one or more of the four functions: a precursor to SRL, a mediator of SRL, a concomitant of SRL outcomes, and a primary outcome of SRL.

There is a large body of research investigating the effects of individual difference constructs, such as motivational factors, self-efficacy, and self-regulated learning, on academic achievement (Liao, Ferdenzi, & Edlin, 2012; Pintrich, 1999; Pintrich & De Groot, 1990; Richardson, Abraham, and Bond, 2012; Winne & Hadwin, 2008; Zimmerman, 1990). Most of the studies focused on native English-speaking students in secondary education and tertiary education in the United States and only a handful of studies paid attention to English language learners (Phakiti et al., 2013). For example, in a meta-analysis of the psychological correlates of university students’ academic achievement, Richardson et al. (2012) synthesized 217 research papers on the relationship between non-intellective constructs and academic achievement through a systematic and cyclical search for relevant papers published from 1997 to 2010. The contribution of five categories of non-intellective constructs to university students’ academic achievement was analyzed, including personality traits, motivational factors, self-regulated learning strategies, students’ approaches to learning, and psychological contextual influences. The meta-analysis results indicated that effort regulation as a self-regulated learning strategies,
academic self-efficacy, and grade goals, both as a motivational factor, are significant predictors of university students’ academic achievement.

According to Bandura (2006), self-efficacy refers to the beliefs that people have about their capability to accomplish an anticipated task or take a course of action. Previous empirical studies on motivation reveal that self-efficacy could be a strong predictor of academic achievement (Mills, Pajares, & Herron, 2007). In a review of motivation as an academic enabler, Linnenbrink and Pintrich (2002) contended that motivation, in the social cognitive framework, should be considered as a multifaceted construct, comprising at least four components: self-efficacy, attribution, intrinsic motivation, and goal orientation. Self-efficacy has been found to have positive relationship with a number of learning behaviors, including task choice, persistence, cognitive engagement, use of self-regulatory strategies, and ultimately academic achievement (Linnenbrink & Pintrich, 2002). Gore (2006) reported two empirical studies on the predictive power of academic self-efficacy and other variables on academic achievement of college students in the U.S. Using hierarchical multiple regression and logistic regression, Gore argued that academic self-efficacy could predict students’ success in college.

Self-regulation is defined as “the control of one’s present conduct based on motives related to a subsequent goal or ideal that an individual has set for him or herself” (Zimmerman & Schunk, 2008, p. 1). Self-regulation has been found to play a critical role in learners’ academic achievement. For example, McKenzie, Gow, and Schweitzer (2004) conducted a large-scale research project to investigate the factors contributing to Australian college students’ first-year achievement. Using a structural equation modeling approach, McKenzie et al. (2004) found that self-regulated learning strategies, along with previous achievement and certain personality traits, had a direct effect on first-year achievement, while achievement motivation had an indirect effect.

As the studies revealed, the mentioned individual difference constructs should be able to contribute to ESL learners’ academic achievement. There are limited studies on these constructs in ESL contexts. Furthermore, the relationships among these constructs and the placement of ESL learners based on the scores from English placement tests have barely been investigated. Therefore, this dissertation study will fill in the gap through an empirical investigation of the extrapolation and ramification inferences for English placement tests.

2.3.4 Summary and research questions 2 and 3

The review of the studies mentioned in this section highlights the importance of investigating test stakeholders’ perceptions as backings to assumptions in the ramification inference. However, most of the studies did not include test stakeholders’ views on the impact of the score use. This study aims to fill in the gap and frames this line of inquiry within the argument-based approach to validation. Another aspect of the ramification inference is the actual impact of the placement decisions on ESL students’ English learning and academic achievement. However, there is a scarcity of studies on the impact of placement decisions on ESL students’ English learning. On the other hand, previous predictive validation studies indicated a weak to moderate relationship between English proficiency and academic achievement while acknowledging a lack of control over confounding factors that have been proved to contribute to
students’ academic achievement. To present a more complete picture of the impact of placement decisions, this dissertation attempts to include individual constructs in the collection of validity evidence for the ramification inference for the English Placement Test used at Iowa State University.

Based on the studies discussed earlier, the following research questions are proposed with regard to the ramification inferences of the EPT validity argument:

1) How did the EPT test stakeholders, including ESL students, content course instructors, and ESL course instructors, perceive the placement decisions and their impact on students’ English learning and academic achievement? (Ramification)

2) To what extent did the ESL courses help ESL students improve English proficiency? (Ramification)

3) To what extent was ESL students’ English proficiency related to their academic achievement in light of individual differences in motivational constructs and selfregulated learning? (Ramification)

The three research questions correspond to the ramification inference of the EPT validity argument by focusing on the impact of the course placement decisions on students’ English learning and academic achievement. Exploring these research questions will contribute directly to the validity argument of the EPT and inform future development of the EPT. Furthermore, a closer look at the impact of the EPT and the relationship between the EPT and the ESL courses will shed light on the effectiveness of the ESL program at the university.
CHAPTER 3. METHODOLOGY

This chapter starts with an explanation of the overall research design with regard to the research questions. I describe the participants in this study with a focus on their characteristics, relevance to research questions as well as sampling techniques used. Then, I elaborate on the instruments used in this study and the data collection procedures. This chapter ends with a detailed description of the analytical techniques for each research question.

3.1. Overall research design

The overall research design in this study is mixed-methods sequential explanatory design (see Figure 3.1), in which the first step is to collect and analyze quantitative data, and the second step is to collect and analyze qualitative data. The results of the qualitative data analyses are used to explain and interpret the findings in the quantitative data analyses (Creswell & Plano-Clark, 2007). A mixed-methods approach is particularly useful for studying the extrapolation and ramification inferences in this study because it combines the strengths of both qualitative and qualitative methods and thus provides more comprehensive insights for the validity of test score interpretation and use (Creswell & Plano-Clark, 2007).
As shown in Figure 3.1, both quantitative and qualitative data are used to address the research questions. The quantitative data constitute a relatively large portion of the total data collected in this study. Each individual research question has its own research orientation. For example, ESL students’ test performance data on the EPT, the TOEFL iBT, and the self-assessment are used to answer research question 1 regarding the relationship among the three measures (extrapolation inference). ESL course performance data in a pre- and posttest design are used to answer research question 3 regarding the impact of the EPT placement decisions on ESL student’s English learning. ESL students’ EPT performance, first-semester Grade Point Average (GPA), and their questionnaire responses on individual constructs are used to answer research question 4 regarding the relationship between English proficiency as measured with the EPT and academic achievement. The answer to research question 2, which focuses on test stakeholders’ perception of the EPT decisions, is revealed from the use of semi-structured interview data only. The interview data offer useful information for the quantitative inquiries.
For example, ESL students’ interview data about the relationship between English proficiency and academic achievement adds insight to the findings for research question 4, which focused on the quantitative aspect of the relationship.

### 3.2. Participants

To answer the research questions, this study identified and included four types of test stakeholders as participants: ESL students, ESL instructors, academic advisors, and content course instructors. The demographic information and representativeness of these participants, together with the sampling techniques used, are described in the following subsections.

#### 3.2.1. ESL students

The largest body of participants in this study was the newly admitted ESL students at Iowa State University. In this study, the non-native English-speaking students at Iowa State University are the target population and the 618 ESL students who took the EPT in the 2014 fall semester were treated as a subpopulation, representing the majority of the non-native English-speaking students in their first semester at Iowa State University.

Three groups of ESL students were involved in this study to address the research questions: 1) 347 ESL students who responded to a self-assessment and a motivation questionnaire (RQ1 and RQ4); 2) eight questionnaire respondents who were purposefully sampled to participate in a face-to-face semi-structured interview (RQ2); 3) the ESL students three ESL courses (38 in four ESL listening classes, 18 in one lower-level ESL writing class, and 16 in one higher-level ESL writing class) in the 2014 fall semester who were sampled for their ESL course performance data (RQ3).
The first group of the EPT test-takers were asked to finish an online self-assessment as well as a comprehensive questionnaire about individual difference constructs on motivation. A total of 436 ESL students responded to the invitation email and completed the self-assessment and the comprehensive questionnaire. However, only 347 of these were determined to be valid cases and were retained for further analysis after initial screening and data cleaning. This sample size is adequate for estimating robust parameters in Rasch model analysis (Chen et al., 2014; Wright, 1977), which was used for checking the psychometric quality of the self-assessment items as well as the questionnaire items on motivational factors. The rationale of using Rasch model in this study is presented in 3.5.2.

The characteristics of the 347 participants are described in the third column in Table 3.1, along with the characteristics of two sub-samples for specific analyses in the fourth and fifth columns. Among the 347 respondents, there were 129 female participants and 194 male participants; 24 participants did not specify their gender. The majority of the participants were either in their early twenties ($n=197$) or late twenties ($n=106$), which largely corresponded to the number of undergraduate students ($n=178$) and graduate students ($n=135$). About half of the participants spoke Chinese as their native language ($n=173$) and the next biggest native language groups were Arabic ($n=29$) and Hindi ($n=16$). In terms of field of study, 144 participants were from the College of Engineering, followed by students from the College of Liberal Arts and Science ($n=58$), and the Business College ($n=52$). Two hundred and sixteen participants were in their first semester in the U.S. and 97 participants had studied in the U.S. for two semesters or longer.
<table>
<thead>
<tr>
<th>Demographic variables</th>
<th>Demographic information</th>
<th>Study sample (N = 347)</th>
<th>Sub-sample for MTMM (n = 202)</th>
<th>Sub-sample for CFA/SEM (n = 239)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-22</td>
<td>197</td>
<td>110</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>23-30</td>
<td>106</td>
<td>71</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>31-35</td>
<td>12</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Over 35</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>29</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>129</td>
<td>76</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>194</td>
<td>116</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>24</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Native language</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>29</td>
<td>5</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>173</td>
<td>112</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>16</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>14</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Malay</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Thai</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Other or Unknown</td>
<td>83</td>
<td>42</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Educational status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate</td>
<td>178</td>
<td>95</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>Graduate</td>
<td>135</td>
<td>96</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>34</td>
<td>11</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>College</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>144</td>
<td>87</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>52</td>
<td>25</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Liberal Arts and Science</td>
<td>58</td>
<td>36</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Agriculture and Life Sciences</td>
<td>19</td>
<td>13</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Human Sciences</td>
<td>20</td>
<td>15</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Veterinary Medicine</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>33</td>
<td>10</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.1 continued

<table>
<thead>
<tr>
<th>Demographic variables</th>
<th>Demographic information</th>
<th>Study sample (N = 347)</th>
<th>Sub-sample for MTMM (n = 202)</th>
<th>Sub-sample for CFA/SEM (n = 239)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of stay in the U.S.</td>
<td>Less than 3 months</td>
<td>216</td>
<td>152</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>4-12 month</td>
<td>21</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>1-2 years</td>
<td>36</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2-3 years</td>
<td>20</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>More than 3 years</td>
<td>20</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Other or Unknown</td>
<td>34</td>
<td>18</td>
<td>11</td>
</tr>
</tbody>
</table>

Note: MTMM = multitrait-multimethod, CFA = confirmatory factor analysis, SEM = structural equation modeling.

As shown in Table 3.1, there are 202 valid cases for a multitrait-multimethod (MTMM) analysis (fourth column) and 239 cases for a confirmatory factor analysis (CFA) and a structural equation modeling (SEM) analysis (fifth column). They are the sub-samples from the 347 respondents after multiple sources of data are combined for specific analyses. The sample sizes in these two columns are smaller than 347 in the third column due to missing data in one of the three sources: the EPT test, the TOEFL iBT test, the self-assessment and questionnaire on motivational factors. These sample sizes are adequate for the confirmatory factor analysis, but are small for structural equation modeling to meet the requirement of participant to item ratio of 8:1 (Brown, 2006; Worthington & Whittaker, 2006).

Some ESL student respondents who finished the self-assessment and the motivation questionnaire were invited to participate in a face-to-face semi-structured interview and they constitute the second group of participants in this study. Purposeful sampling technique was used for identifying ESL student interviewees from the first group of ESL students who took the self-assessment and the motivation questionnaire. Two to three students from each type of ESL course (reading, listening, and academic English writing) as well as two to three students who passed the EPT were randomly selected from the pool of questionnaire respondents and invited...
to participate in the interview via email. Eight of the 27 ESL students invited agreed to participate in the interview. The demographic information of the participants is listed in Table 3.2. There were four male and four female interviewees. Their native languages included Chinese \((n = 3)\), Korean \((n = 1)\), Malay \((n = 2)\), Arabic \((n = 1)\), and Persian \((n = 1)\). Two participants were from the College of Engineering, two from the Business College, two from the College of Design, one from the College of Human Science, and one from the College of Liberal Arts and Science.

Table 3.2

<table>
<thead>
<tr>
<th>Interviewee (Pseudonym)</th>
<th>Gender</th>
<th>L1</th>
<th>Education</th>
<th>College</th>
<th>ESL courses placed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yuan-Feng</td>
<td>Female</td>
<td>Chinese</td>
<td>Graduate</td>
<td>Business</td>
<td>Engl101D</td>
</tr>
<tr>
<td>Felicity</td>
<td>Female</td>
<td>Malay</td>
<td>Undergraduate</td>
<td>Design</td>
<td>Engl99R, 101C</td>
</tr>
<tr>
<td>Hua-Chen</td>
<td>Female</td>
<td>Chinese</td>
<td>Graduate</td>
<td>Business</td>
<td>Engl99R, 101B, 101D</td>
</tr>
<tr>
<td>Hyun-Ki</td>
<td>Male</td>
<td>Korean</td>
<td>Graduate</td>
<td>Human Science</td>
<td>Engl101D</td>
</tr>
<tr>
<td>Rasha</td>
<td>Female</td>
<td>Arabic</td>
<td>Undergraduate</td>
<td>Liberal Arts &amp; Science</td>
<td>Engl99R, 101B, 101C</td>
</tr>
<tr>
<td>Sanjar</td>
<td>Male</td>
<td>Persian</td>
<td>Graduate</td>
<td>Engineering</td>
<td>Engl101D</td>
</tr>
<tr>
<td>Austin</td>
<td>Male</td>
<td>Malay</td>
<td>Undergraduate</td>
<td>Design</td>
<td>Engl99L, 101C</td>
</tr>
</tbody>
</table>

The third group of ESL student participants were the students enrolled in some of the ESL courses, whose course performance data were collected and analyzed to partially address research question 3 about the impact of the EPT placement decisions on their English learning. Convenience sampling technique was used to collect ESL students’ writing performance data in the ESL courses. For the ease of data access, I followed the approved Institutional Review Board (IRB) procedures and contacted instructors of ESL courses in person for de-identified students’ course performance data. This group of ESL students include 18 students from one of five sections of Engl101B classes, a lower-level ESL writing class for both undergraduate and
graduate students (see Appendix H for sample course syllabus), 16 students from one of 15 sections of Engl101C class, a higher-level ESL writing class for undergraduate students only (see Appendix I for sample course syllabus), and 38 students from four of eight sections in Engl99L class, an ESL listening class (see Appendix G for sample course syllabus). Unfortunately, I could not collect course performance data from Engl99R, an ESL reading class, and Engl101D, a higher-level writing class for graduate students due to some logistical issues in data collection. The Engl99R classes and Engl101D classes did not use the same or comparable final exams to the first-day in-class tests, which made a meaningful comparison of students’ performance at the beginning and end of the course less plausible. In addition, since I only used the de-identified essays in Engl101B/C and test scores in Engl99L for analysis, no demographic information of the ESL students was made available to me.

3.2.2. ESL instructors

In this study, five ESL instructors were interviewed for their perceptions of the EPT placement decisions and the impact of these decisions. The ESL instructors were purposefully sampled considering their ESL teaching experience and native languages.

As shown in Table 3.3, the first five ESL instructors participated in this study as ESL instructor interviewees. Among them, four were female ESL instructors and one was male instructor. Their native languages included Turkish (n = 2), Chinese, Korean, and English. Four ESL instructors were PhD students in the Applied Linguistics and Technology program and one was an MA student in TESL/Applied Linguistics at Iowa State University. Their length of English teaching varied from two years to eight years. Four instructors had taught more than one ESL course at Iowa State University.
### Table 3.3
Demographic Information of the ESL Instructor Interviewees

<table>
<thead>
<tr>
<th>ESL instructors (Pseudonym)</th>
<th>Gender</th>
<th>L1</th>
<th>Education</th>
<th>Teaching Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mengqi</td>
<td>Female</td>
<td>Chinese</td>
<td>PhD student</td>
<td>4 years, Engl101B, C, Engl150</td>
</tr>
<tr>
<td>Adalet</td>
<td>Female</td>
<td>Turkish</td>
<td>PhD student</td>
<td>7 years, Engl101B, C</td>
</tr>
<tr>
<td>Young-Soo</td>
<td>Male</td>
<td>Korean</td>
<td>PhD student</td>
<td>2 years, Engl101C</td>
</tr>
<tr>
<td>Rachael</td>
<td>Female</td>
<td>English</td>
<td>MA student</td>
<td>8 years, high school ESL courses, Engl99R/L</td>
</tr>
<tr>
<td>Sabriye</td>
<td>Female</td>
<td>Turkish</td>
<td>PhD student</td>
<td>3.5 years, Engl101D, Engl99R, Engl150</td>
</tr>
<tr>
<td>Hee-Sook</td>
<td>Female</td>
<td>Korean</td>
<td>PhD student</td>
<td>5 years, Engl101B, C, D</td>
</tr>
</tbody>
</table>

Three ESL instructors were invited to participate in this study as ESL essay raters using the EPT rubric (see Appendix K), including Sabriye, Adalet, and Hee-Sook. Hee-Sook, a female PhD student in the Applied Linguistics and Technology program, had taught ESL courses for five years at the time of this study. Her native language was Korean. These three ESL instructors had rated the EPT essays in the past semesters and could be regarded as experienced raters.

### 3.2.3. Academic advisors

The potential academic advisor participants were identified through multiple sources such as the school directory, professional networking, and colleague recommendation. Three academic advisers responded to my invitation email and agreed to participate in a face-to-face semi-structured interview. As shown in Table 3.4, the academic advisor interviewees were from the College of Liberal Arts and Science, the Business College, and the College of Engineering, which attracted the majority of international students at Iowa State University. Two of the interviewed academic advisors held a Master’s degree and one had a PhD degree. Their advising experience ranged from ten months to five years.
Table 3.4  
Demographic Information of the Academic Advisors  

<table>
<thead>
<tr>
<th>Interviewee (Pseudonym)</th>
<th>Gender</th>
<th>L1</th>
<th>Education</th>
<th>College</th>
<th>Advising Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Male</td>
<td>English</td>
<td>Master’s</td>
<td>Liberal Arts &amp; Science</td>
<td>2 years</td>
</tr>
<tr>
<td>Mary</td>
<td>Female</td>
<td>English</td>
<td>PhD</td>
<td>Business</td>
<td>5 years</td>
</tr>
<tr>
<td>Todd</td>
<td>Male</td>
<td>English</td>
<td>Master’s</td>
<td>Engineering</td>
<td>10 months</td>
</tr>
</tbody>
</table>

3.2.4. Content course instructors  

Four content course instructors were invited to participate in a face-to-face semi-structured interview (see Table 3.5). The content course instructors were identified and contacted using snowball sampling or referral sampling method. I asked for referral of content course instructors from the interviewed academic advisors and various undergraduate students at Iowa State University.

Table 3.5  
Demographic Information of the Content Course Instructors  

<table>
<thead>
<tr>
<th>Interviewee (Pseudonym)</th>
<th>Gender</th>
<th>L1</th>
<th>Education/Title</th>
<th>Department</th>
<th>Teaching Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feng-Chun</td>
<td>Female</td>
<td>Chinese</td>
<td>PhD/Post-Doc Associate</td>
<td>Mathematics</td>
<td>4 years</td>
</tr>
<tr>
<td>Ganling</td>
<td>Male</td>
<td>Chinese</td>
<td>PhD/Associate Professor</td>
<td>Computer Science</td>
<td>8 years</td>
</tr>
<tr>
<td>Jake</td>
<td>Male</td>
<td>English</td>
<td>PhD/Post-Doc Associate</td>
<td>Civil Engineering</td>
<td>1 semester</td>
</tr>
<tr>
<td>Kathleen</td>
<td>Female</td>
<td>English</td>
<td>PhD/Senior Lecturer</td>
<td>Psychology</td>
<td>7 years</td>
</tr>
</tbody>
</table>

The content course instructor interviewees were the instructors teaching fundamental courses in Mathematics, Computer Science, Civil Engineering, and Psychology. Two were male instructors and two were female instructors. Their native languages were English \((n = 2)\) and Chinese \((n = 2)\). The content course instructors were either PhD candidates or PhD degree holders. Their length of teaching varied from one semester to eight years. As observed by the content course instructors, there were ESL students in their classes. However, there was no...
definite information about whether these students were concurrently enrolled in ESL courses or not.

3.3. Instruments

The instruments used for data collection in this study included three types of measures of English abilities (the EPT, the TOEFL iBT, and the self-assessment); a comprehensive questionnaire to gather self-report on motivation, academic self-efficacy and self-regulated learning strategies; one syntactic complexity analyzer; one lexical complexity analyzer; and an automated writing evaluation tool. Each will be described in turn below.

3.3.1. The English Placement Test (EPT)

The English Placement Test is a post-entry English language test for the admitted international students whose native language is not English at Iowa State University. As stated on the official webpage of the EPT, “the results of the test are used to determine whether non-native English-speaking ISU students have met the English requirement (i.e., no further ESL instruction is required) or whether students need to take additional ESL classes” (the EPT website, http://engl.iastate.edu/ept/eptfaq.html).

There are three sections in the EPT, namely the reading comprehension section, the listening comprehension section, and the essay writing section. The EPT is administered to the students at the beginning of each semester in large auditoriums and the length of the test is about three hours with a ten-minute break between the reading and listening sections, and the writing section. The EPT test is mainly delivered as a paper-and-pencil test using a multiple-choice question format for both the reading and listening sections, and an essay writing task for the writing section (see Figure 3.2). A computerized version of the EPT is available on Blackboard.
Learn, a course management system used at Iowa State University, but, the computerized EPT is mainly used for the ESL students enrolled in distance education programs.

Corresponding to the three sections in the EPT, three types of ESL courses are offered at this university, namely Engl99R (Strategies for Non-native Speakers of English: Strategies for Reading), Engl99L (Strategies for Non-native Speakers of English: Strategies for Listening), and a suite of ESL writing courses, including Engl101B (English for Native Speakers of Other Languages: Academic English), Engl101C (English for Native Speakers of Other Languages: Academic English II – Undergraduates), and Engl101D (English for Native Speakers of Other Languages: Academic English II – Graduates). There are two levels of academic writing courses offered to match the EPT essay rating results: Engl101B is a lower-level course for intermediate-high level students with a focus on grammar learning and paragraph writing, whereas Engl101C and Engl101D are higher-level courses for advanced-low undergraduate students and graduate students, respectively, with Engl101C focusing on essay structure and organization and Engl101D on research paper/thesis sections. The proficiency levels (intermediate-high and advanced-low) are labeled using the American Council on the Teaching Foreign Languages (ACTFL) terminology because the EPT essay rating rubric is built on the ACTFL framework.

The placement decisions for these courses are made based on students’ performance on the respective section of the EPT. Therefore, the major interpretation and use of the EPT scores are as follows: The EPT scores represent three skills of ESL learners’ academic English proficiency, namely reading, listening, and writing skills. The scores are used as indicators of whether the students need further ESL assistance and as a criterion for the decisions of ESL course placement. In the fall semester in 2014, a total of 618 new ESL students took the EPT test.
and the majority of the students (about 73% of the total test-takers) were placed into one or more ESL classes based on their performance on the EPT.

**Figure 3.2.** Structure of the English Placement and corresponding ESL courses. MC = multiple-choice (items); UG = undergraduate students, G = graduate students.

### 3.3.2. In-class tests and final exams in ESL courses

In each ESL course, there is a 40 to 50-minute in-class test of the target subskill administered to students on the first day of class. The in-class tests are typically developed by ESL course coordinators and some ESL instructors to measure the skills and strategies taught in the classes. The purpose of the in-class tests is twofold: 1) to provide ESL instructors with the performance information of their students and 2) to identify potentially misplaced ESL students in these classes. Therefore, the results of the in-class tests can be interpreted as ESL students’ proficiency level in the targeted subskills (reading, listening, and writing), and the use of the
results is to make course-waiving decisions for those misplaced students and to help ESL instructors understand their students’ needs.

The in-class test for the listening classes consists of multiple choice questions, true-false questions, fill-in-the-blank questions, and short open-ended questions. The in-class listening test is graded by ESL instructors and cut-off scores are set by the course coordinators for the decision of waiving ESL courses. In the fall 2014 semester, the cutoff score for the in-class test for waiving the ESL listening course was 65 out of 100. The same in-class test of English listening skills was used at the beginning and the end of the semester in the fall 2014 semester.

The in-class test for the writing classes uses essay prompts prepared by the course coordinators. Sample prompts and the scoring rubric are provided in Appendices J and K. The essays are rated by ESL course instructors based on the EPT essay scoring rubric to identify high-quality essays. Decisions about waiving ESL writing course are made in a two-step process: initial recommendations are proposed by ESL course instructors and then adjudicated by the course coordinators. In the ESL writing classes, the same prompts for the in-class tests are often used at the end of the semester as a way to gauge students’ progress in the courses. The same practice of repeating the in-class test at the end of the course is employed in the listening class as well.

3.3.3. Questionnaire

A questionnaire was developed for this study based on my pilot study as well as literature on learning motivation. It is used in the research to tap into students’ self-assessment of English proficiency and their motivational constructs. The questionnaire consists of four sections, namely self-assessment of English proficiency (20 items), academic self-efficacy (5 items), learning motivation (8 items), and self-regulated learning strategies (10 items) (see Appendix A for the
questionnaire). A 6-point Likert scale is used in the questionnaire, with 1 being “Strongly Disagree” and 6 being “Strongly Agree”. In this subsection, I introduce the structure and content of each subsection in the questionnaire. The information about questionnaire administration is provided later in the section of data collection procedures.

3.3.3.1. Self-assessment

The self-assessment contains 20 can-do statements about the four subskills, listening, reading, speaking, and writing, with reference to students’ activities in content courses or major courses (see Appendix A for the self-assessment items). The self-assessment of English proficiency in the context of content courses was developed based on existing literature concerning self-assessment and informal interviews with ESL students at the university for typical English use in their content courses. According to the admission policy for international applicants at the university, the minimum requirement for English proficiency is set as 71 for the TOEFL iBT or 6 for the IELTS (Admission Office, 2014). The corresponding proficiency level on the Common European Framework of Reference for Languages (CEFR) is B1 to B2 (ETS, 2006; Taylor & Jones, 2006). The corresponding proficiency level on the ACTFL is Intermediate High (Slagter, Surface, & Watson, 2010). Considering the fact that all of the admitted ESL students had higher TOEFL or IELTS score than required, I mainly reviewed the self-assessment descriptors at higher proficiency levels in the European Language Portfolio (ELP) (B2-C2), Association of Language Testers in Europe (ALTE, Levels 4 and 5), and ACTFL (Intermediate high to Advanced-high) and developed the items for the self-assessment in this study. All the items were piloted with a small number of students from the target population and reviewed by experts in Applied Linguistics at Iowa State University.
The questionnaire items were designed to assess the self-assessment construct in context by including in each item classroom activities such as reading textbooks, understanding lectures, taking notes, making oral presentations, writing reports, engaging in discussions, and so on. An example item for reading is “I can fully understand a lecture, presentation and/or demonstration in English in and outside my classes.” The student’s requested response could range from 1 (Strongly disagree) through 6 (Strongly agree). Therefore, a high score on this section of the questionnaire was interpreted as higher level of the subskills described in each item.

3.3.3.2. Motivation

The learning motivation items were intended to measure both intrinsic and extrinsic learning motivation. Intrinsic motivation or intrinsic goal orientation reflects the idea that the causes of students’ engagement in activities are mainly for personal interest, curiosity, and intention to master certain skills, while extrinsic motivation or extrinsic goal orientation shows that students’ participatory purposes are to achieve good grades or impress others. Like the items about academic self-efficacy, this motivation construct was included in this study to account for the potential influence of the EPT performance on learning motivation as well as the influence of learning motivation on ESL students’ academic achievement. The items were adapted from the Intrinsic and Extrinsic Goal Orientation scales in the Motivated Strategies for Learning Questionnaire (MSLQ), which was originally developed by Pintrich and his colleagues in 1980s and 1990s to measure college students’ motivation and learning strategies (Pintrich, Smith, Garcia, & McKeachie, 1993). The reason for adapting the items from the MLSQ is that it has been recognized as a reliable measure of its targeting constructs, namely motivation and self-regulated learning strategies, and these constructs have been used in a large number of studies as predictors of college students’ academic performance (Credé & Philips, 2011). In addition, the
MSLQ is found to be useful in a variety of educational settings and has been used in different countries.

The MSLQ consists of 81 items in two broad sections: motivation scales and learning strategy scales. The motivation scales in the MSLQ include three components: Value components, which consist of Intrinsic goal orientation (4 items), Extrinsic goal orientation (4 items), Task value (6 items), Expectancy components, which consist of Control of learning beliefs (4 items), Self-efficacy for learning and performance (8 items), and Affective component or Test anxiety (5 items). The learning strategy scales include two general strategies: Cognitive and metacognitive strategies, which contained Rehearsal (four items), Elaboration (six items), Organization (four items), Critical thinking (five items), Metacognitive self-regulation (12 items), and Resources management strategies, which contained Time/study environmental management (eight items), Effort regulation (four items), Peer learning (three items), and Help seeking (four items).

The motivational sections in the MSLQ were useful and appropriate to this study because of their content relevance to university-level learning at U.S. universities and their documented uses with good psychometric quality (Credé & Philips, 2011; Pintrich, et al. 1993). The Intrinsic and Extrinsic goal orientation in the value components of the motivation scales are particularly relevant to this study as goal has been a major component in various motivational theories, and goal-orientation theory is particularly about a dichotomy of achievement goals in educational context, namely mastery (intrinsic) orientation and performance (extrinsic) orientation (Pintrich, 2000). In addition, goal orientation has been found to affect academic achievement (Diseth, 2011). An example of intrinsic goal orientation item is *In my content courses, I prefer course material that really challenges me so I can learn new things* (1 Strongly Disagree to 6 Strongly
Agree). An example of extrinsic goal orientation item is *If I can, I want to get better grades than most of the other students in my major course(s)* (1 Strongly Disagree to 6 Strongly Agree).

### 3.3.3.2. Academic self-efficacy

Academic self-efficacy refers to students’ perceptions or self-appraisal of their ability to accomplish academic work. Research has established that academic self-efficacy can influence students’ learning motivation and affect students’ academic performance (Richardson et al., 2012). The items of academic self-efficacy used in this study were adapted from the subscale of Academic self-efficacy in the Patterns of Adaptive Learning Scales (PALS) (Midgley et al., 2000) and the Self-efficacy for learning and performance scale in the MSLQ (Pintrich et al., 1993). All the statements were rephrased to reflect the academic contexts at the university. Five statements about students’ expectation and self-efficacy were included in the survey. An example item is *I believe I am able to receive excellent grades in my major course(s) at this University if I work hard.* (1 Strongly Disagree to 6 Strongly Agree).

### 3.3.3.4. Self-regulated learning strategies

Self-regulated learning (SRL) is defined as “the degree to which students are metacognitively, motivationally, behaviorally active participants in their own learning process” (Zimmerman, 2008, p.167). Dörnyei and Skehan (2003) posited that self-regulation and motivation are “inextricably bound together because they both concern the antecedents of increased academic achievement” (p. 612). The items assessing self-regulated learning strategies were adapted from the statements on strategies for the regulation of academic behavior in the MSLQ, mainly from the subscales of Time and study environment, Effort regulation, and Help seeking.
In this study, I did not include the cognitive and metacognitive strategies in the MSLQ because they are self-regulation of cognition and are very specific to learning activities. Instead, I am interested in how students organized and managed their study in general, and thus I focused on the aspects of Time and study environment, Effort regulation, and Help seeking. An example item of time and study environment is *I have a regular time set aside for studying for my major course(s).* An example item of effort regulation is *When I study for my major course(s), I often quit before I finish what I planned to do because I feel bored or lazy* (1 Strongly Disagree to 6 Strongly Agree). An example item of help seeking is *When I can’t understand the material in my major course(s), I ask other students or friends for help* (1 Strongly Disagree to 6 Strongly Agree).

### 3.3.4. Interview questions for test stakeholders

To elicit test stakeholders’ perceptions of the EPT placement decisions and their impact, four sets of interview questions and protocols were developed separately for four groups of test stakeholders, namely ESL students, ESL instructors, content course instructors, and academic advisors (see Appendices C to F for interview questions). Different groups of test stakeholders can provide different perspectives on the EPT. For example, in addition to personal view of the EPT placement decisions, ESL students can share their test-taking experiences as well as their learning experiences in both ESL courses and content courses at Iowa State University. The ESL instructors, who usually are involved in the EPT proctoring and essay rating practices, can contribute essential information about placement accuracy through their observation of ESL students’ attitude and performance in ESL classes. The academic advisors can comment on the EPT placement decisions based on their advising experiences with ESL students as well as their observations of ESL students’ needs and challenges regarding academic English. The content
course instructors can provide feedback on ESL students’ performance in their classes and share their concerns related to ESL students’ English proficiency. These four sets of interview questions were designed to answer the second research question and therefore there were some overlapping interview questions across the four sets of interview questions. The development of interview questions and protocols underwent multiple rounds and the questions were finalized after revisions based on the suggestions and comments from peer debriefing and reviewing. The details of the interviewing procedures and data analysis will be described later in this chapter.

The interview questions consist of three sections, with the first section collecting interviewees’ demographic information, the second focusing on their views of placement decisions and the impact of the placement, and the last section inquiring about their views of the relationship between English proficiency and academic achievement at ISU. The total number of interview questions ranged from 8 to 11, with several probing questions attached to some main questions for each type of interviewee. For example, a sample question in the first section for ESL instructors is “Can you tell me something about yourself and the ESL courses that you have taught at ISU?” A sample question in the second section for academic advisor is “To what extent do you think the EPT is a fair English test to our students?” A sample question in the third section for ESL students concerns “Whether taking ESL courses helped you perform better in your major courses or not?”

3.3.5. Lexical Complexity Analyzer (LCA) and L2 Syntactic Complexity Analyzer (L2SCA)

In this study, ESL students’ essays were analyzed from the perspectives on complexity, accuracy, and fluency (Housen, Kuiken, & Vedder, 2012; Skehan, 2009) as well as for their holistic qualify based on the EPT essay scoring rubric. To better understand students’ progress in essay writing quality across a 16-week semester period, two web-based text analyzers, namely
Lexical Complexity Analyzer (LCA) and L2 Syntactic Complexity Analyzer (L2SCA), (http://aihaiyang.com/software/) were used to analyze the complexity aspects in the sampled essays which were written by 34 ESL students at the beginning and the end of the 2014 fall semester. LCA and L2SCA were developed by Xiaofei Lu and Haiyang Ai at Pennsylvania State University and their web-based versions support a batch mode to process multiple texts in a single run. The complexity features identified by LCA and L2SCA are in line with the research on both syntactic complexity and lexical complexity (Housen et al., 2012; Wolfe-Quintero, Inagaki, & Kim, 1998) and have been empirically investigated (Lu, 2010, 2011, 2012).

According to Lu (2012), LCA is capable of analyzing lexical density, lexical sophistication (5 variables) as well as lexical variation (25 variables). Lexical density refers to the ratio of content words to the total number of words. Lexical sophistication or lexical rareness refers to the proportion of unusual words in a text. Lexical variation, also known as lexical diversity, shows the lexical range used by writers. It is noteworthy that several variables in the output of LCA are similar in nature, but standardized in different ways; for example, there are four variants of Type-Token ratio (TTR) in the output for lexical variation.

L2SCA analyzes texts for 14 syntactic complexity features grouped in five categories: Length of production, sentence complexity, Subordination, coordination, and particular structures (Lu, 2010, 2011). The length of production is captured with three variables: mean length of sentence, clause, and T-unit. Sentence complexity is represented with clauses per sentence. Subordination contains four variables: clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, and dependent clauses per T-unit. Coordination contains three variables: coordinate phrases per clause, coordinate phrases per T-unit, and T-units per sentence.
Particular structures analyzed are complex nominal per clause, complex nominal per T-unit, and verb phrases per T-unit.

In a corpus-based study on these syntactic complexity measures as indicators of ESL writer’s language development, Lu (2011) identified seven of the 14 features to be statistically different in ESL students’ writing from proficiency level 1 to level 3, conceptualized as school level. Complex nominal per clause (CN/C), mean length of clause (MLC), complex nominal per T-unit (CN/T), mean length of sentence (MLS), mean length of T-unit (MLT), coordinate phrases per clause (CP/C), coordinate phrase per T-unit (CP/T). In my analysis of essay characteristics for research question 3, this set of seven syntactic complexity measures or features are used.

3.3.6. Criterion

To examine the grammatical accuracy in the essays, I used Criterion, an automated writing evaluation tool developed by the Educational Testing Service, to automatically identify main grammatical errors in the timed essays written by the ESL students at the beginning and the end of the writing classes. Criterion uses e-rater as a scoring engine to generate 40 types of trait feedback in five categories, namely Grammar, Usage, Mechanics, Style, and Organization & Development (Burstein, Tetreault, & Madnani, 2013). The reason for choosing Criterion as an automated error identifier in this study is that 1) error identification is time-consuming for human readers. Criterion is efficient and objective in detecting typical grammatical errors; 2) Criterion has been found to have relatively high accuracy in detecting rule-based errors, such as article use (Chodorow, Gamon, & Tetreault, 2010). Detailed description of Criterion feedback can be found in Appendix N. In this study, I mainly focus on five grammatical errors that are frequently made by ESL learners: Subject-verb agreement, Ill-formed verb, Missing or extra article, Confused
words, and Preposition error (Bitchener & Ferris, 2012). The counts of the errors were normalized against the essay length for comparison purpose.

### 3.4. Data collection procedures

The data collection procedures and timeline are depicted in Figure 3.3. There were three main phases of data collection. At the beginning of the 2014 fall semester, I collected the EPT data and administered the self-assessment to ESL students, along with the motivation questionnaire. In the second half of the 2014 fall semester, I conducted face-to-face semi-structured interviews with ESL students and acquired the TOEFL iBT and IELTS scores of the newly enrolled international students from the Registrar’s office. At the beginning of the 2015 spring semester, I conducted face-to-face semi-structured interviews with ESL instructors, content course instructors, and academic instructors. I also requested and collected the GPA data as well as the ESL course performance data following the approved procedures in the Institutional Review Board (IRB) application. Detailed data collection procedures are described below.

**Figure 3.3.** Procedures and phases of data collection.
The self-assessment was administered in the format of an online questionnaire in weeks 3 and 4. The scheduling of the self-assessment administration was determined with a consideration that students were supposed to become familiar with the English language requirements in their content courses by then so that they would better be able to self-assess English proficiency with reference to their academic activities in their content courses. This decision was also informed by the findings from the pilot study, which showed that the timing of self-assessment administration might be a factor influencing the relationship between the EPT performance and the self-assessment (Li, 2015). A weak correlation between the EPT and the self-assessment (.150 to .373) was reported in the pilot study, in which the self-assessment was administered in weeks six and seven of the spring 2014 semester. It is speculated that ESL students had made various gains in English proficiency after a one-and-half-month period of immersion at Iowa State University and this relatively long-period of immersion may fail to reflect ESL students’ initial English proficiency as measured by the EPT.

An email list of the test-takers of the fall 2014 EPT was requested from the EPT office and an invitation email was sent to all the EPT test-takers in weeks 3-4 in the 2014 fall semester through an emailing function on Qualtrics, a web-based survey tool (http://www.qualtrics.com). An electronic informed consent form was presented on the first page of the online questionnaire stating that student’s participation was voluntary. In the invitation email, the purpose of this study and its significance to the EPT and the ESL program were explained to the ESL students. Students were also notified that the participants in the study had a chance to win one of the three 10-dollar gift cards if they entered their email address in the questionnaire for a lottery. To improve the response rate to the questionnaire, I sent two reminder emails to the ESL students who had not completed the questionnaire at an interval of five days after the initial invitation.
email. In addition, I sought help from the ESL instructors and asked them to introduce the questionnaire both orally in their classes and by email. In most cases, ESL students completed the questionnaire on their own outside classes. Time stamps and the time spent on each section of the questionnaire were automatically recorded by Qualtrics, which were used as a reference for data cleaning.

To ensure a good quality of self-assessment data, I manually cleaned and screened the data by identifying the respondents who spent little time on the self-assessment or showed disingenuous response patterns in the survey. As observed in the piloting of the questionnaire, it was estimated that five to eight minutes were needed to finish the self-assessment. If a student, for example, only used less than 4 minutes or responded to all the items using the same value, say, 1 or 6, I excluded this student from my data analysis. This screening process was intended to result in good quality of the data for analysis.

The semi-structured interviews with ESL students and ESL instructors were conducted in a neutral, unthreatening environment in the second half of the fall 2014 semester. The interviews with academic advisors, content course instructors, and ESL course instructors were conducted in the interviewee’s offices. All the interviews were conducted in English and were audio recorded with the permission of the interviewees (see Appendix B for a sample consent form). The length of the interview varied from 25 to 45 minutes. A member check was carried out later to ensure the accuracy of transcription and interpretation of interviewee’s responses (Glesne, 2011). In the member check, a sample of theme interpretation with interview excerpts was presented to the interviewees, if they were available. Feedback from the interviewees was incorporated in the data analysis and discussion.
The test score data, including ESL students’ TOEFL iBT or IELTS scores and their EPT scores, were requested from the Registrar’s office and the EPT office with approval from the Institutional Review Board (IRB) at the university. All the test performance data were de-identified for analysis after being matched with students’ questionnaire responses.

ESL students’ performance data in the ESL courses were collected as a source of evidence for progress in English proficiency. A pre- and post-test design was used in the ESL courses. In the reading and listening courses, students’ reading and listening scores on the EPT were recorded as a pre-test data set. The same in-class test was used at the end of the semester as part of the final exam and as a post-test data set. In the academic English writing classes, students’ essays written on the first day of the class were used as a pre-test data set and their essays written for the final exam at the end of the semester were used as a post-test data set. Comparisons between the pre-test data and post-test data shed light on students’ progress in English learning.

3.5. Data analysis methods

In this section, the analytical techniques are introduced following the order of research questions. Prior to conducting the question-specific techniques, descriptive statistics and reliability analysis was the first step for all the instruments in this study.

3.5.1. Descriptive statistics and reliability analysis

The descriptive statistics of the data from each instrument (the EPT, the TOEFL iBT, the self-assessment, and motivation questionnaire) were calculated using SPSS 21 (IBM Corp., 2012). Along with the basic information such as mean, median, and standard deviation, univariate normality information (Skewness, Kurtosis, Shapiro-Wilk test) was also reported.
In terms of test score reliability, Cronbach’s alpha was calculated as the reliability index for the reading and listening sections in the EPT. Due to a lack of item-level information of the TOEFL iBT in this study, I used the reported reliability of the TOEFL iBT subsections on the official website of the TOEFL iBT in this study. The reliability and Rasch separation index for the self-assessment and questionnaire sections of motivation, academic self-efficacy, and self-regulated learning strategies were calculated in the Rasch model analysis. The reliability information was then used in the estimation of disattenuated correlation coefficients, which reflects the correlation between the ‘true scores’ on two measures or the scores without any measurement errors. Disattenuated correlation coefficients are calculated using the formula:

$$r_{TxTy} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$$

where $r_{TxTy}$ is the correlation coefficient corrected for attenuation or disattenuated correlation coefficient, $r_{xy}$ is the correlation coefficient between measure X and measure Y, $r_{xx}$ is the reliability of measure X and $r_{yy}$ is the reliability of measure Y. (Bachman, 2004).

3.5.2. Rasch model analysis of the self-assessment (RQ1)

The first step to address research question 1, which is about the relationship between the EPT and two external measures, i.e., the self-assessment and the TOEFL iBT, is to examine whether the self-assessment as a new instrument developed in this study possesses good psychometric properties. The self-assessment in this study is in the format of can-do statements on a 6-point Likert scale. A Rasch rating scale model-based item analysis was conducted to investigate the item reliability, person reliability, item difficulty, item discrimination, and scale functioning using Winsteps (Linacre, 2006).

The Rasch rating scale model, as a member of the one-parameter item response theory (IRT) models, is capable of analyzing polytomous data such as responses to Likert-scale items.
Other polytomous models include the Rasch partial credit model, the generalized partial credit model (a 2-parameter IRT model), and the graded response model (a 2-parameter IRT model) (de Ayala, 2009). The polytomous IRT models enjoy several advantages over the classical test theory (CTT), another popular approach used for analyzing Likert scale-based data and developing survey instruments. CTT assumes that an observed score consists a true score and measurement errors associated with each item in a test. In the framework of CTT, item difficulty is calculated as the percentage of test-takers who answered the item correctly and item discrimination is calculated as item-total correlation to reflect the degree to which an item can distinguish high-proficiency level test-takers and lower-level proficiency level test-takers. One major limitation in the CTT is that these two indices are sample dependent, in other words, test-takers’ ability may be labeled as drastically different if a much more difficult or easier test were administered (Carr, 2011). In a CTT-based analysis of Likert scale data, the raw score on each item is usually summed to form a scale score. From the perspective of polytomous IRT models for Likert data analysis, the item difficulty is deemed as the extent to which a respondent would endorse a certain category on the Likert scale based on his or her level on the target construct, for example, motivation or self-efficacy (Sick, 2009). When the unidimensionality assumption is met, the IRT models position item difficulty on the same logit scale along with test-taker’s level on a target construct and it conceptualizes an item score as a degree to which test-taker’s level of the target construct matches item difficulty or endorsability. In this sense, the difficulty parameter in the IRT models is sample-independent. This is particularly important in scale development because item parameters that were calibrated using pilot data are expected to be invariant across sample groups.
In a CTT-based analysis of Likert scale data, the raw scores on each item are summed to form a scale score, even though the Likert scale data are ordinal in nature (Sick, 2009). Another of the advantages of using the polytomous IRT models is that these models treat Likert scale data as ordinal data, as opposed to interval data in non-IRT model analysis, but transform the “counts of the endorsement of these ordered Likert categories into interval scales based on actual empirical evidence” (Bond & Fox, 2007, p.106). Therefore, a data-based threshold structure of Likert scale items can be detected empirically and a true measurement scale with equal intervals in the unit of logit can be established (Davidson & Henning, 1985).

Within the IRT family, both the Rasch rating scale model and the graded response model have been widely used in analyzing Likert scaled based instruments for their psychometric properties (de Ayala, 2009). Both model are capable of providing more psychometric property information at item-level and test-level, compared to the CTT approach. The major difference between these two polytomous models is that the Rasch rating scale model estimates a single parameter, the item difficulty or endorsability, while the graded response model estimates both item difficulty parameter and item discrimination parameter. The cost of estimating two parameters in the graded response model is a requirement of larger sample size – at least 500 respondents for a stable estimation of parameters (2009). For this practical reason, the Rasch rating scale model was used in this study.

Like other IRT models, the Rasch rating scale model assumes that the items measure the same unidimensional construct; this is known as unidimensionality assumption (Bond & Fox, 2007). This assumption was checked with both exploratory factor analysis (maximum likelihood extraction and promax rotation) and the Rasch principal component analysis of residuals. According to Bond and Fox (2007), the infit and outfit mean squares (MNSQ) as unstandardized
fit statistics are used to assess whether an item functions as the Rasch model expects. Infiz is a weighted fit statistic and is less sensitive to outliers, compared to outfit, an unweighted fit statistic. The expected mean square value of the infit and outfit statistics is 1.0. The value range from 0.5 to 1.5 is generally regarded as an acceptable fit to the Rasch model (Green, 2013).

Another relevant item quality index is the point-biserial correlation coefficient of each item, which is the discrimination index in the classical test theory (CTT) framework.

Similar to the consideration of item fit in Rasch analysis, the categories of the Likert scale should also exhibit a good model fit. According to Bond and Fox (2007), the following four characteristics of a rating scale should be checked: category frequency, monotonicity of category average measures, threshold or step calibrations, and category fit. Category frequency is the total number of each of level or response category on the Likert scale chosen or endorsed by the respondents. Monotonicity of category average measures refers to the phenomenon in which the average ability measured in the unit of logit increases along with the increase of level or category on the Likert scale. Threshold or step calibrations are the estimated difficulty for choosing one category over its adjacent one. Similar to the item fit in Rasch modeling, category fit includes infit and outfit with both indicating the scale quality through the fit information about the extent to which the response categories function as the model expects. It is recommended that the count for each category should be no less than 10, the distance between thresholds should be at least 1.4 logits, but less than 5 logits, and the infit statistics for each category should be acceptable.

The same procedures of item analysis applied to the questionnaire items assessing motivation, academic self-efficacy, and self-regulated learning strategies.
3.5.3. Confirmatory factor analysis (RQ1)

A confirmatory factor analysis was conducted to investigate the factor structure of the self-assessment, using Mplus 7.0 (Muthén & Muthén, 2012). Three theoretically plausible models of English proficiency, the correlated four-subskill/factor model, the uncorrelated four-subskill/factor model, and the unitary factor model, were proposed and tested according (In’nami & Koizumi, 2012).

In confirmatory factor analysis as well as in structural equation modeling analysis to be used for research question 4, multiple model fit indices were reported to help decide which model fits best: Chi-square ($\chi^2$), the ratio of chi-square to degrees of freedom ($\chi^2$/df), comparative fit index (CFI), Tucker-Lewis index (TLI) or non-normed fit index (NNFI), the standardized root mean square residual (SRMR) or the weighted root mean square residual (WRMR), and the root mean square error of approximation (RMSEA) (Brown, 2006; Tabachnick & Fidell, 2013). Chi-square ($\chi^2$) was reported as the classic goodness-of-fit index in this study. A non-significant chi-square ($p > .05$) indicates that we should fail to reject (i.e., accept) the null hypothesis that the proposed model generates the same variances and covariances as those in the sample data. However, chi-square is sensitive to sample size and tends to be significant with adequate sample size. In this study, the ratio of chi-square to degree of freedom ($\chi^2$/df) was reported with a value less than 2.0 being regarded as good model fit (Tabachnick & Fidell, 2013). In addition, comparative fit index (CFI) as a type of relative fit indices compares the chi-square value to a baseline model. A CFI of .90 or .95 is indicative of good model fit (Byrne, 2012). Tucker-Lewis index (TLI) is similar to CFI in that it involves a comparison of the ratios of chi-square statistics to the degree of freedom between the specified model and the baseline model. The main difference is that TLI compensate for model complexity...
and its possible value can exceed 1. Like CFI, a value closer to 1 is indicative of good model fit. Lastly, the standardized root mean square residual (SRMR), the weighted root mean square residual (WRMR), and the root mean square error of approximation (RMSEA) were used as absolute model fit indices, which penalize poor model parsimony. Both SRMR and WRMR are variants of root mean square residual (RMR). SRMR is a standardized version of RMR. A SRMR value of .10 indicates acceptable model fit and .08 suggests good model fit. WRMR is an alternative to SRMR when categorical variables are used. A WRMR value of one or lower indicates a good model fit (Wang & Wang, 2012). The range of RMSEA is from zero to one. A 90% confidence interval of RMSEA is usually reported to gauge the index precision (Brown, 2006). A value less than 0.05 for RMSEA indicates a good model fit and 0.08 indicates an acceptable model fit. The final factor structure was determined based on the model fit indices and theoretical soundness.

Once the factor structure of the best fitting model for the self-assessment was identified, regression-based factor scores of the identified subskills (i.e., reading, listening, speaking, and writing) in self-assessment were calculated for each respondent and these scores were used in the subsequent correlational analyses.

3.5.4. Multitrait-multimethod analysis of the EPT, the TOEFL iBT and the self-assessment (RQ1)

A multitrait-multimethod (MTMM) matrix of the correlation coefficients was constructed for the three measures (the EPT, the TOEFL iBT, and the self-assessment) of four subskills (reading, listening, speaking, and writing). The participants reporting only IELTS scores were excluded from this MTMM matrix analysis due to their small number and its incompatible score scales with the TOEFL iBT (9 point-scale on IELTS versus 120-point scale on the TOEFL
iBT). The MTMM matrix consisted of Pearson’s $r$ and Spearman’s $\rho$. The latter was for the correlation coefficients involving the EPT writing grade, which is on a three-point ordinal scale. Evidence about convergent correlations, discriminant correlations, and test method effects was obtained from an analysis of the MTMM matrix. Both the correlation coefficients and their disattenuated counterparts or the correlation coefficients corrected for attenuation were presented in this study.

In MTMM data analysis, convergent evidence, represented by the monotrait-heteromethod correlation coefficients, indicates the extent to which different methods measuring the same traits are related. Theoretically, the monotrait-heteromethod correlation coefficients are expected to be high relative to the heterotrait-heteromethod correlations. The low values of the latter are interpreted as discriminant evidence in support of the extrapolation inference in the EPT validity argument. The heterotrait-monomethod correlations show the extent to which method effects are exhibited in the MTMM results.

In addition to the correlation matrix-based analysis of MTMM data, there are two confirmatory factor analysis (CFA)-based approaches to MTMM analysis, namely correlated trait-correlated method (CTCM) model and correlated trait-correlated uniqueness (CTCU) model (see Figure 3.4). The former is currently thought of as a traditional CFA approach or general approach CFA to analyzing MTMM data proposed by Widaman (Byrne, 2012). This approach starts with a baseline model in which all the indicators are loaded on both traits and methods simultaneously then compares the baseline model with three nested models (i.e., no trait/correlated method model, perfectly correlated trait/freely correlated methods model, freely correlated traits/uncorrelated methods model) to determine convergent evidence, divergent evidence, and method effects. One common issue with this general CFA approach is that the
baseline model may not converge due to its complexity. An alternative approach is to use a correlated trait-correlated uniqueness (CTCU) model, also known as correlated residual model and corrected error model. Measurement errors are usually assumed to be random and their magnitudes determine measurement reliability. According to Brown (2006), the correlated trait-correlated uniqueness model assumes that two or more indicators from different latent variables covary due to certain method effects, for example, the indicators measured by the same type of method. In this study, both correlated trait-correlated methods (CTCM) model and correlated trait-correlated uniqueness (CTCU) model were used to analyze the MTMM data.

In the analysis of CTCU model, the factor loadings to the latent trait factors on the left side of the model were examined for both convergent and divergent evidence, whereas the correlations among the measurement error or unique variance (uniqueness) on the right side of the model were examined for method effects. Correlating measurement errors or unique variance can be justified with the assumption that individual items or indicators are influenced by shared measurement methods. The same set of model fit indices and criteria described above for the confirmatory factor analysis of the self-assessment were reported in the analysis of CTCU model, including chi-square ($\chi^2$), the ratio of chi-square to degrees of freedom ($\chi^2$/df), comparative fit index (CFI), Tucker-Lewis index (TLI) or non-normed fit index, the standardized root mean square residual (SRMR) or the weighted root mean square residual (WRMR), and the root mean square error of approximation (RMSEA).
3.5.5. Inductive analysis of the interview transcripts (RQ2)

The second research question aims to elicit ESL students’ perception of the EPT itself through individual interviews as well as test stakeholders’ perception of the placement decisions and their impact on ESL students’ English learning and academic achievement. The EPT stakeholders in this study include ESL students, ESL course instructors, content course instructors, and academic advisors.

All the interviews were transcribed verbatim in Microsoft Word and the transcripts were coded in a line-by-line fashion using highlighting and commenting functions in Microsoft Word (Saldana, 2009). Only pseudonyms of the participants were used in this study. The procedures of qualitative data analysis are depicted in Figure 3.5. The transcripts were segmented and the initial open coding was accomplished with several rounds of close reading of the transcripts. The initial open coding was conducted interview question by question to explore shared responses.
and patterns. When the initial coding of the transcripts was complete, a more focused coding was conducted to synthesize the initial codes and cluster them into broader categories based on the relationship among the codes. The categories were further analyzed and used to generate some umbrella themes via axial coding, such as test stakeholders’ view of the placement decisions and the impact of the decisions on the individual students, course instructors as well as academic advisors.

**Figure 3.5.** Steps for interview data analysis.

### 3.5.6. Paired-samples t-tests of ESL course performance data (RQ3)

Research question 3 addresses the impact of placement decisions on ESL students’ English learning. ESL students’ progress in English learning was gauged through a comparison of their performance at the beginning and the end of the 2014 fall semester in ESL courses (see Figure 3.6). For reading and listening courses, the same items were used in the in-class tests and the final exams and therefore, students’ performance were compared using paired sample t-tests.
For the academic English writing classes, the essays written in the in-class test at the beginning of the course were compared to the essays written in the final exam in terms of syntactic complexity, lexical complexity, and grammatical accuracy. An ESL learner corpus was formed in this study, consisting of 68 essays written by the ESL students at two time points, namely the beginning of the course and the end of the course. These linguistic features were quantified using *Lexical Complexity Analyzer (LCA)* and *L2 Syntactic Complexity Analyzer (L2SCA)* (see Appendices L and M for the linguistic features reported in *LCA* and *L2SCA*). Descriptive statistics of these quantified variables were reported. Paired sample t-tests were conducted to investigate students’ progress in English writing. To assess the readiness of the Engl101B students in academic writing, independent sample t-tests of the lexical and syntactic complexity variables were used to compare the essays written by Engl101B students in the final exam and the essays written by Engl101C students in the first timed essay task.

To complement the quantitative analysis of lexical and syntactic complexity of the essays, the essays written in the final exams from one Engl101B class and one Engl101C class (n
were rated by three experienced ESL instructors using the EPT scoring rubric for the same three levels, namely Engl101B, Engl101C, and Pass. The EPT rubric was used by the human raters to make sure the ratings were comparable to the EPT placement. To ensure the rating quality, I led a one-hour rater training with the three raters, following the guidelines in Weigle (2002). In the rating session, a set of benchmark essays were rated by the raters and the ratings were calibrated to the benchmark grade through discussions. After the raters demonstrated a good match of their ratings with the benchmark grades, they rated three final exam essays separately to check their rating reliability during the training session. The rest of the essays were rated independently by the raters. The intra-class correlation, which functions as an inter-rater reliability index when more than two raters are involved, was .711. The final rating was determined based on the two agreed grades. For example, an essay was marked as “Pass” by Adalet, “Engl101C” by Sabriye, and “Pass” by Hee-Sook, and then the final grade for this essay was “Pass.”

3.5.7. Confirmatory factor analysis (RQ4)

Research question 4 addresses the relationship between ESL students’ initial English proficiency measured by the EPT and their academic achievement, as mediated by motivational factors. Another set of confirmatory factor analysis was used to verify the factor structures of the questionnaire of motivation factors as suggested in previous studies on these constructs. The model fit of the hypothesized models were judged using the same set of model fit indices as described in 3.5.1 for the confirmatory factor analysis of the self-assessment, including chi-square ($\chi^2$), the ratio of chi-square to degrees of freedom ($\chi^2$/df), comparative fit index (CFI), Tucker-Lewis index (TLI) or non-normed fit index, the weighted root mean square residual (WRMR), and the root mean square error of approximation (RMSEA).
3.5.8. Structural equation modeling (RQ4)

The last research question is about the impact of the EPT placement decisions on students’ academic achievement, taking into account individual difference factors. I took a structural equation modeling (SEM) approach to examining the relationship among English proficiency, measured with the EPT (self-assessment and TOEFL), academic achievement operationalized as first-semester GPA, and the individual difference factors, including learning motivation, academic self-efficacy, and self-regulated learning strategies. SEM, as a confirmatory statistical methodology, usually consists of two models, namely a measurement model, which shows the relationship between observed indicators and latent variables, and a structural model, which captures the relationship among latent variables (Byrne, 2012).

Based on the theories of individual differences and empirical findings on the relevant individual difference factors, the following relationships were hypothesized: English proficiency, which is represented by the EPT results on reading, listening, and writing, has a direct impact on learners’ academic self-efficacy and learning motivation. At the same time, academic self-efficacy affects self-regulated learning strategies through learning motivation. The residuals of the two types of motivation (intrinsic goal orientation and extrinsic goal orientation) are assumed to be correlated. Based on empirical studies on academic achievement, it is hypothesized that English proficiency has a direct impact on academic achievement, operationalized as first semester GPA, which is also influenced by students’ self-regulated learning strategies and learning motivation.
Figure 3.7. Hypothesized model of the relationship among the latent variables with GPA. EPT = the English Placement Test, ASE = Academic self-efficacy, MO_EX = Extrinsic motivation, MO_IN = Intrinsic motivation, SRL = Self-regulated learning strategies, GPA = Grade Point Average.

The specified models were identified using the t-rule and correlated errors rule. The t-rule is a necessary, but not sufficient condition for model identification, and it states that the number of freely estimated parameters should be less than or equal to the number of known parameters for model identification (Bollen, 1989). The correlated errors rule is a sufficient, but not necessary condition of identification. It is an expansion of the two-indicator rule and the three-indicator rule and allows for limited correlation among the measurement errors. The assumption of both univariate and multivariate normality of the sample data was checked using multivariate kurtosis together with a critical ratio. The Mahalanobis distance was checked for potential multivariate outliers. Considering the non-normality of the data, a robust version of maximum likelihood estimator, MLR in Mplus 7.0, was used in the confirmatory factor analyses to determine model fit and parameter estimates (Byrne, 2012). Whenever the EPT essay grades were included in a model, the mean and variance adjusted weighted least square estimator (WLSMV) in Mplus 7.0 was used to account for the categorical nature of this variable (Byrne, 2012). The same model fit indices were used to judge the extent to which the data fit the model.
3.6. Chapter summary

This chapter presented the research methodology, including the overall design, the characteristics, and recruitment of participants, the rationales, and procedures for development of the instruments, data collection methods, and analytical techniques used in this study. Table 3.6 summarizes the key points as they relate to each of the research questions investigated in the study.

Table 3.6
Summary of Data Collection and Analysis

<table>
<thead>
<tr>
<th>Research question</th>
<th>Data (Type)</th>
<th>Data collection</th>
<th>Analytic methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extrapolation inference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ1: Relationship between the EPT and external criteria</td>
<td>• Test performance data (the EPT, the TOEFL iBT &amp; Self-assessment of English use) – (Quantitative)</td>
<td>• Test performance data were requested from the EPT office with IRB approval. • Likert-scale based online questionnaire was administered to ESL students.</td>
<td>• Rasch model analysis of the self-assessment items; • Confirmatory factor analysis of the self-assessment items; • Analysis of the MTMM correlation matrix • CFA-based MTMM analysis (correlated trait-correlated uniqueness model)</td>
</tr>
<tr>
<td><strong>Ramification inference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ2: Test stakeholders’ perception of the placement decisions and their impact</td>
<td>• Test stakeholders’ perception – (Qualitative)</td>
<td>• Face-to-face semi-structured individual interviews were conducted.</td>
<td>• Inductive analysis of the interview transcripts</td>
</tr>
<tr>
<td>Research question</td>
<td>Data (Type)</td>
<td>Data collection</td>
<td>Analytic methods</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>RQ3:</strong> Impact of the EPT on English learning</td>
<td>• ESL course performance – (Quantitative and Qualitative)</td>
<td>• A pre- and post-test design was used.</td>
<td>• Paired-samples <em>t</em>-test of students’ performance on the in-class tests and the final exam of ESL Listening courses; • Analysis of syntactic and lexical complexity of the essays written at the beginning and the end of ESL writing classes; • Automated evaluation of grammatical accuracy in the essays; • Human rating of the essays using the EPT essay rating rubric</td>
</tr>
<tr>
<td><strong>RQ4:</strong> Impact of the EPT on academic achievement</td>
<td>• First-semester cumulative GPA – (Quantitative); • ESL students’ characteristics on individual difference variables – (Quantitative)</td>
<td>• The GPA data were requested from the Registrar’s office with IRB approval. • Likert-scale based questionnaire on individual difference constructs were administered to ESL students.</td>
<td>• Rasch model analysis questionnaire items; • Confirmatory factor analysis of the individual different items; • Structural equation modeling analysis of the relationship among the EPT, first-semester GPA, and the individual difference constructs</td>
</tr>
</tbody>
</table>
CHAPTER 4. RESULTS

This chapter is organized by the research questions for which the results of quantitative and qualitative analyses are reported. To answer the first research question on the relationship between the EPT and self-assessment as well as the TOEFL iBT, a multitrait-multimethod analysis was conducted, following descriptive statistics and reliability analysis of the three measures of English proficiency. To answer the second research question regarding the EPT test stakeholders’ view of the placement decisions and the impact of these decisions, the qualitative data from semi-structure interviews were coded and analyzed for themes. To answer the third research question on the impact of placement decisions on English learning, ESL students’ course performance data are analyzed quantitatively within the framework of a pre- and posttest design. To answer the last research question, a structural equation modeling analysis was conducted after the factor structure of each motivational construct had been identified using confirmatory factor analysis.

4.1 The relationship among the EPT and self-assessment and the TOEFL iBT

The first research question investigates relationships of the EPT with other measures of English language proficiency to seek support for the extrapolation inference. These relationships were explored together in a multitrait-multimethod (MTMM) design to assess the strengths of observed relationships relative to those expected based on the constructs tested by the EPT. Prior to investigating this model, I examined the descriptive statistics and reliabilities of the measures to assess their usability for the MTMM analysis, and I verified that the factor structure of the self-assessment questionnaire met theoretical expectations. Overall, the results for the MTMM
analysis supported the expected relationships between the EPT and the two measures, namely the TOEFL iBT and the self-assessment.

4.1.1. Descriptive statistics and reliability analysis of the EPT, the TOEFL iBT and the self-assessment

The descriptive statistics and reliability analysis for the three measures, namely the self-assessment, the EPT, and the TOEFL iBT, are reported one by one in this subsection.

4.1.1.1. The EPT

The descriptive statistics for the EPT are listed in Table 4.1. The mean score of the EPT reading sections was 18.36 out of 35 or 52.5 on a 100-point scale with a standard deviation of 5.21. The mean score of the EPT listening section was 21.09 out of 35 or 62.3 on a 100-point scale with a standard deviation of 5.15. By comparison, the ESL students had a higher average score on the EPT reading section than the listening section. The score data on the EPT reading and listening sections are normally distributed as indicated by the small values in skewness and kurtosis as well as the non-significant p-values of the Shapiro-Wilk’s test (.336 & .413). The results of the EPT writing section are of ordinal nature, consisting of three ordinal levels (Engl101B, Engl101C, and Pass).

Table 4.1
Descriptive Statistics for the EPT (n = 202)

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>M</th>
<th>S.D</th>
<th>Min/Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro-Wilk p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPT</td>
<td>Reading</td>
<td>18.36</td>
<td>5.21</td>
<td>6/34</td>
<td>0.001</td>
<td>-0.341</td>
<td>.336</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>21.09</td>
<td>5.15</td>
<td>5/34</td>
<td>0.015</td>
<td>-0.158</td>
<td>.413</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>2.15</td>
<td>0.60</td>
<td>1/3</td>
<td>-0.060</td>
<td>-0.297</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Note. a. The full scores of reading and listening sections of the EPT were 40.

Cronbach’s alpha was reported as reliability index for the scores on the EPT reading and listening sections. The reliability of the EPT reading section in fall of 2014 was .67 and that for the EPT listening section was also .67. Each essay in the EPT writing section was graded by at
least two human raters. In the case of rating disagreement, a third rater, sometimes a fourth rater, was invited to rate the essay independently and the final rating was determined based on the agreed ratings from two raters. The resultant rating data were the typical sparse data with a large number of missing values. In other words, many essays were rated by a combination of different raters. Therefore, only inter-rating reliability, instead of inter-rater reliability, of the EPT writing section was estimated using Cronbach’s alpha after treating human raters as a random facet and consolidating the rating from different human raters into two to four ratings. The inter-rating reliability of the EPT writing section in fall 2014 was .787 (Cronbach’s alpha, N = 587). Due to a lack of access to the EPT essay ratings of the sampled ESL students, I used the inter-rating reliability for the whole fall 2014 EPT essays in this study and the true reliability of the study sample may be slightly lower than .787 due to smaller sample and the resultant smaller variance.

4.1.1.2 The TOEFL iBT

The descriptive statistics for the TOEFL iBT are listed in Table 4.2. Similar to the EPT, the ESL students had a higher average score on the TOEFL iBT reading section, with a mean score of 22.87 out of 35 or 76.2% and a standard deviation of 4.09. The TOEFL iBT speaking score was the lowest, with a mean score of 20.63 out of 30 or 68.8% and a standard deviation of 2.54. This relatively small standard deviation suggests that the TOEFL iBT speaking scores had less variation. The mean score of the TOEFL iBT listening section was 21.88 out of 30 or 72.9% with a standard deviation of 3.88. The mean score of the TOEFL iBT writing section was 22.33 out of 30 or 74.4% with a standard deviation of 3.05. The mean total score of the TOEFL iBT was 87.83 out 120 or 73.2% with a standard deviation of 9.92.

Since all of the ESL students in this study were fully admitted international students who met the minimum English language requirement, the TOEFL iBT scores were the truncated
portions with a relatively narrower range and less variation. In addition, the significant Shapiro-Wilk p-values (<.05) indicated that the TOEFL section scores and total scores were not normally distributed in this data set.

Table 4.2
*Descriptive Statistics for the TOEFL iBT (n = 202)*

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>M</th>
<th>S.D</th>
<th>Min/Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro-Wilk p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT</td>
<td>Reading</td>
<td>22.87</td>
<td>4.09</td>
<td>10/30</td>
<td>-0.435</td>
<td>-0.457</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>21.88</td>
<td>3.88</td>
<td>9/29</td>
<td>-0.294</td>
<td>-0.091</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>Speaking</td>
<td>20.63</td>
<td>2.54</td>
<td>14/27</td>
<td>0.308</td>
<td>-0.474</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>22.33</td>
<td>3.05</td>
<td>15/28</td>
<td>0.065</td>
<td>-0.888</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>87.83</td>
<td>9.92</td>
<td>60/109</td>
<td>-0.160</td>
<td>-0.903</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Because item-level information of the TOEFL iBT scores was not available in this study, a compromise was made to use the reported reliability information from ETS ([http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Reliability_Generalizability.pdf](http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Reliability_Generalizability.pdf)). Considering the impact of restricted score range in the data in this study, the official reliability information released by the ETS may be higher than the actual reliability of the TOEFL iBT scores of the ESL students in this study. The reported reliability for the TOEFL iBT reading section is .86, for the listening section is .87, for the speaking is .90, and for the writing is .78.

4.1.1.3. The Self-assessment

The descriptive statistics show ESL students’ general performance and score distributions on the self-assessment based on the 6-point Likert scale (see Table 4.3). The descriptive statistics of the self-assessment revealed that the means of the responses to the self-assessment items ranged from 3.52 to 4.89 with standard deviation from 1.10 to 1.40. The descriptive statistics suggested that the ESL students tended to positively rate their English proficiency in the specific academic contexts and there were large variances in ESL students’ evaluation in each self-
assessment items as shown by the magnitudes of standard deviations. Significant Shapiro-Wilk 
$p$-values in Table 4.3 showed that the responses on all of the self-assessment items were not 
normally distributed, with the majority of the items negatively skewed or having high scoring 
responses.

Table 4.3

<table>
<thead>
<tr>
<th>SA</th>
<th>Items</th>
<th>M</th>
<th>SD</th>
<th>Min/Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro-Wilk p value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>Rd1</td>
<td>4.27</td>
<td>1.22</td>
<td>1/6</td>
<td>-0.44</td>
<td>-0.39</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Rd2</td>
<td>4.52</td>
<td>1.22</td>
<td>1/6</td>
<td>-0.67</td>
<td>-0.21</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Rd3</td>
<td>4.29</td>
<td>1.27</td>
<td>1/6</td>
<td>-0.41</td>
<td>-0.72</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Rd4</td>
<td>4.89</td>
<td>1.16</td>
<td>1/6</td>
<td>-1.09</td>
<td>0.91</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Rd5</td>
<td>4.12</td>
<td>1.19</td>
<td>1/6</td>
<td>-0.28</td>
<td>-0.59</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Listening</td>
<td>Lsn1</td>
<td>4.44</td>
<td>1.26</td>
<td>1/6</td>
<td>-0.66</td>
<td>-0.03</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Lsn2</td>
<td>4.30</td>
<td>1.22</td>
<td>1/6</td>
<td>-0.35</td>
<td>-0.53</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Lsn3</td>
<td>3.97</td>
<td>1.38</td>
<td>1/6</td>
<td>-0.31</td>
<td>-0.66</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Lsn4</td>
<td>4.13</td>
<td>1.25</td>
<td>1/6</td>
<td>-0.29</td>
<td>-0.47</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Lsn5</td>
<td>3.52</td>
<td>1.40</td>
<td>1/6</td>
<td>0.01</td>
<td>-0.78</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Speaking</td>
<td>Spk1</td>
<td>4.03</td>
<td>1.25</td>
<td>1/6</td>
<td>-0.38</td>
<td>-0.36</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Spk2</td>
<td>4.03</td>
<td>1.29</td>
<td>1/6</td>
<td>-0.37</td>
<td>-0.36</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Spk3</td>
<td>4.31</td>
<td>1.24</td>
<td>1/6</td>
<td>-0.39</td>
<td>-0.62</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Spk4</td>
<td>4.01</td>
<td>1.23</td>
<td>1/6</td>
<td>-0.32</td>
<td>-0.39</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Spk5</td>
<td>4.02</td>
<td>1.23</td>
<td>1/6</td>
<td>-0.28</td>
<td>-0.39</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Writing</td>
<td>Wrt1</td>
<td>4.17</td>
<td>1.10</td>
<td>1/6</td>
<td>-0.23</td>
<td>-0.14</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Wrt2</td>
<td>4.31</td>
<td>1.17</td>
<td>1/6</td>
<td>-0.50</td>
<td>-0.18</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Wrt3</td>
<td>3.81</td>
<td>1.21</td>
<td>1/6</td>
<td>-0.20</td>
<td>-0.29</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Wrt4</td>
<td>4.12</td>
<td>1.20</td>
<td>1/6</td>
<td>-0.43</td>
<td>-0.26</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>Wrt5</td>
<td>4.20</td>
<td>1.19</td>
<td>1/6</td>
<td>-0.31</td>
<td>-0.47</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

The reliability for each of the self-assessment sections was high, as shown in Cronbach’s 
alpha (.90 for reading, .89 for listening, .93 for speaking, and .92 for writing). Since factor scores 
of the self-assessment sections will be used in the multitrait-multimethod (MTMM) analysis, the
scale reliability or factor rho coefficient for factor scores of the self-assessment sections under the confirmatory factor analysis framework will be reported later in this section, using the formula suggested by Kline (2011).

**4.1.2. Rasch model analysis and confirmatory factor analysis of the self-assessment**

To evaluate the quality of the self-assessment, a Rasch measurement analysis using the rating scale model was conducted for each of the subscales. All the valid responses to the self-assessment were used in the Rasch model analysis to increase the modeling accuracy ($N = 347$). The decision to run separate Rasch analyses for each subscale was made based on the results of the pilot study exploratory factor analysis that was used to check the unidimensionality assumption. This decision was further confirmed with the results of the Rasch principal component analysis of residuals. As shown in the two rightmost columns in Table 4.4, the proportion of variance explained by the first dimension is substantial (from 75.2% to 79.9%) and the eigenvalue for the first contrast is less than 2.0 in all four subscales (Linacre, 2011). This configuration of principal components supported the unidimensionality assumption since one component accounted for the majority of the variance.

Table 4.4

*Results of Rasch Measurement Analysis: Reliability and Unidimensionality Check for the Self-assessment (N = 347)*

<table>
<thead>
<tr>
<th>Self-assessment Subscale</th>
<th>Person Separation</th>
<th>Person reliability</th>
<th>Item Separation</th>
<th>Item reliability</th>
<th>Variance explained</th>
<th>Eigenvalue for 1st contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading (k=5)</td>
<td>2.99</td>
<td>.90</td>
<td>8.35</td>
<td>.99</td>
<td>79.6%</td>
<td>1.6</td>
</tr>
<tr>
<td>Listening (k=5)</td>
<td>2.83</td>
<td>.89</td>
<td>8.23</td>
<td>.99</td>
<td>78.1%</td>
<td>1.9</td>
</tr>
<tr>
<td>Speaking (k=5)</td>
<td>3.55</td>
<td>.93</td>
<td>3.55</td>
<td>.93</td>
<td>79.9%</td>
<td>1.5</td>
</tr>
<tr>
<td>Writing (k=5)</td>
<td>3.35</td>
<td>.92</td>
<td>5.12</td>
<td>.96</td>
<td>75.2%</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 4.4 also shows the reliability information of the four subscales. The person reliability, an equivalent of Cronbach’s alpha in the Rasch measurement analysis, ranged from
0.89 to 0.93, indicating a high reliability in each self-assessment section. The person separation index of the self-assessment sections ranged from 2.83 to 3.55, indicating that the extent to which the respondents were spread in standard error units on the measured variables (Bond & Fox, 2007; Iramaneerat, Smith, & Smith, 2008). A higher person separation value suggests a wider spread of person on the scale. Bond and Fox (2007) maintain that both item and person separation index should be at least two. Another useful reference for person separation index value is to use the formula Strata = (4 × Person separation index + 1) / 3 for the number of statistically distinctive levels of person ability. For example, a value of 2.99 in person separation index shows an existence of four discernible levels of person ability. The item reliability in the Rasch measurement analysis refers to the replicability of the item ordering if the items are administered to a group of subjects with similar abilities. The item reliability of the subscales ranged from 0.93 to 0.99 and the item separation index ranged from 3.55 to 8.35. The item separation index can be interpreted in the same way as the person separation index. Overall, the four subscales of the self-assessment tool exhibited high reliability.

To better understand the psychometric quality of the self-assessment items, Rasch item measures, and item-level fit statistics were reported for each subscale in Table 4.5. The item measure or endorsability of the self-assessment statement is estimated on the common scale with the logit as the measurement unit. Overall, the self-assessment items showed a relatively narrow range in item measure (-1.25 to 1.04), meaning that the self-assessment items were not very different in terms of item difficulty.
<table>
<thead>
<tr>
<th>SA</th>
<th>Items</th>
<th>Item measure (logit)</th>
<th>S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Point-biserial coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rd1</td>
<td>0.39</td>
<td>0.08</td>
<td>0.95</td>
<td>0.92</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Rd2</td>
<td>-0.25</td>
<td>0.08</td>
<td>0.78</td>
<td>0.75</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>Rd3</td>
<td>0.35</td>
<td>0.08</td>
<td>1.00</td>
<td>0.98</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>Rd4</td>
<td>-1.25</td>
<td>0.08</td>
<td>0.99</td>
<td>0.94</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td>Rd5</td>
<td>0.76</td>
<td>0.08</td>
<td>1.26</td>
<td>1.28</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>Lsn1</td>
<td>-0.71</td>
<td>0.07</td>
<td>0.99</td>
<td>0.94</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>Lsn2</td>
<td>-0.43</td>
<td>0.07</td>
<td>1.33</td>
<td>1.38</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>Lsn3</td>
<td>-0.09</td>
<td>0.07</td>
<td>0.63</td>
<td>0.65</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>Lsn4</td>
<td>0.20</td>
<td>0.07</td>
<td>0.81</td>
<td>0.78</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Lsn5</td>
<td>1.04</td>
<td>0.07</td>
<td>1.19</td>
<td>1.21</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>Spk1</td>
<td>0.14</td>
<td>0.09</td>
<td>1.32</td>
<td>1.30</td>
<td>.84</td>
</tr>
<tr>
<td></td>
<td>Spk2</td>
<td>0.15</td>
<td>0.09</td>
<td>0.89</td>
<td>0.87</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>Spk3</td>
<td>-0.66</td>
<td>0.09</td>
<td>1.02</td>
<td>1.02</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>Spk4</td>
<td>0.19</td>
<td>0.09</td>
<td>0.81</td>
<td>0.81</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>Spk5</td>
<td>0.17</td>
<td>0.09</td>
<td>0.91</td>
<td>0.91</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>Wrt1</td>
<td>-0.14</td>
<td>0.09</td>
<td>0.98</td>
<td>0.99</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>Wrt2</td>
<td>-0.52</td>
<td>0.09</td>
<td>1.19</td>
<td>1.19</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td>Wrt3</td>
<td>0.86</td>
<td>0.09</td>
<td>0.91</td>
<td>0.91</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>Wrt4</td>
<td>0.01</td>
<td>0.09</td>
<td>1.06</td>
<td>1.05</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>Wrt5</td>
<td>-0.21</td>
<td>0.09</td>
<td>0.82</td>
<td>0.78</td>
<td>.89</td>
</tr>
</tbody>
</table>

According to Bond and Fox (2007), the infit and outfit mean squares (MNSQ) as unstandardized fit statistics are used to assess whether an item functions as the Rasch model expects. Infit is a weighted fit statistic and is less sensitive to outliers, compared with outfit, an unweighted fit statistic. The expected mean square value of the infit and outfit statistics is 1.0. The value range from 0.5 to 1.5 is generally regarded as an acceptable fit to the Rasch model (Green, 2013). As shown in Table 4.5, all the mean square values of the infit and outfit statistics were within the range of 0.5 to 1.5, indicating that the self-assessment items were functioning as
the Rasch model predicted. Another relevant item quality index is the point-biserial correlation coefficient of each item, which is the discrimination index in the classical test theory (CTT) framework. All the self-assessment items had relatively high point-biserial correlation coefficients (.75 to .90), which indicates that the self-assessment items had a good discrimination among the participants.

Similar to the consideration of item fit in Rasch analysis, the categories of the Likert scale should also exhibit a good model fit. In a rating scale model, an estimation of the threshold value of the Likert scale is more meaningful as it tests the functionality of the scale. According to Bond and Fox (2007), the following four characteristics of a rating scale should be checked: the count frequency, monotonicity of category average measures, threshold or step calibrations, and category fit. It is recommended that the count for each category should be no less than 10, the distance between thresholds should be at least 1.4 logits, but less than 5 logits, and the infit statistics for each category should be acceptable. Tables 4.6 to 4.9 contain the category functioning information for each subscale of the self-assessment items.

The categories 1-6 in Tables 4.6 to 4.9 correspond to the 6-point Likert scale levels from “Strongly Disagree” to “Strongly Agree”. The count information refers to the frequency of each category being selected for all the items by the ESL students. The total number of the counts should be close to 1735 for each self-assessment section (5 items × 347 ESL student respondents) with some invalid or extreme responses automatically removed in the Rasch analysis. The infit and outfit mean square (MNSQ) can be interpreted in the same way as for the item infit and outfit MNSQ. Both fit indices indicate the degree to which each category functions as the model expects, with one being perfectly fit to the model the range from 0.5 to 1.5 being acceptable. Structure calibrations are the estimated difficulty for choosing one category over its
adjacent lower category. The bottom level of the Likert scale (1 Strongly Disagree) does not have a lower category for the transition and therefore marked “None” in structure calibration.

Table 4.6
Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Reading Section (N = 347)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count (%)</th>
<th>Obs. average</th>
<th>Infit/Outfit MNSQ</th>
<th>Structure calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12 (1%)</td>
<td>-4.66</td>
<td>2.01/1.41</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>120 (8%)</td>
<td>-1.73</td>
<td>1.48/1.45</td>
<td>-5.80</td>
</tr>
<tr>
<td>3</td>
<td>264 (17%)</td>
<td>-0.05</td>
<td>0.98/1.02</td>
<td>-1.65</td>
</tr>
<tr>
<td>4</td>
<td>408 (26%)</td>
<td>1.62</td>
<td>0.90/0.89</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>550 (35%)</td>
<td>3.15</td>
<td>0.85/0.80</td>
<td>2.10</td>
</tr>
<tr>
<td>6</td>
<td>231 (15%)</td>
<td>5.02</td>
<td>0.99/0.97</td>
<td>4.88</td>
</tr>
</tbody>
</table>

As shown in Table 4.6, the observed average ability of the participants increased monotonically from -4.66 to 5.02 on the logit scale, which indicates that the participants with higher ability tended to endorse or choose a higher category. For example, category 1 or “Strongly Disagree” on the 6-point Likert scale was selected 12 times for the five self-assessment reading items. The average ability of the ESL students who selected “Strongly Disagree” was -4.66 logits. Both infit and outfit mean squares of each category were within the acceptable range (from 0.5 to 1.5), except for the first item in reading (Infit: 2.01, Outfit: 1.41). The distances between thresholds were greater than 1.4 logits, but less than 5 logits. This clear monotonic change pattern in threshold value can be observed in the category probability curves for the Reading (SA_R1), Listening (SA_L1), Speaking (SA_S1) and Writing (SA_W1) subscales respectively (see Figure 4.1). Different categories are represented in curves in different color in Figure 4.1. For example, the red curves represent category 1 (Strongly Disagree) and the light green curves on the rightmost side of each chart represent category 6 (Strongly Agree). The y-axis of the charts shows the probability of endorsing or selecting a category by a respondent with certain ability and the x-axis shows ability level in logits. Ideally, each curve will have certain overlap with its adjacent curves while maintaining a clear un-overlapped peak, if each
category functions as the model expects. The information from Table 4.6 and the first chart in Figure 4.1 jointly showed that the Likert scale for the reading items functioned well.

![Category probability curves for the subscales of self-assessment](image)

*Figure 4.1. Category probability curves for the subscales of self-assessment. SA_R = reading self-assessment, SA_L = listening self-assessment, SA_S = speaking self-assessment, SA_W = writing assessment.*

As expected, the observed average ability of the participants increased monotonically along the Likert scale. Both infit and outfit mean squares of each category were within the acceptable range from 0.5 to 1.5. The distances between thresholds were greater than 1.4 logits, but less
than 5 logits. This clear monotonic change pattern in threshold can be observed in the category probability curve in Figure 4.1.

Table 4.7
Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Listening Section (N = 347)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count (%)</th>
<th>Obs. average</th>
<th>Infit /Outfit MNSQ</th>
<th>Structure calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62 (4%)</td>
<td>-2.91</td>
<td>1.25/1.37</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>158 (10%)</td>
<td>-1.66</td>
<td>1.03/1.13</td>
<td>-3.22</td>
</tr>
<tr>
<td>3</td>
<td>343 (21%)</td>
<td>-0.38</td>
<td>0.93/0.94</td>
<td>-1.75</td>
</tr>
<tr>
<td>4</td>
<td>459 (28%)</td>
<td>0.68</td>
<td>0.83/0.80</td>
<td>-0.12</td>
</tr>
<tr>
<td>5</td>
<td>428 (26%)</td>
<td>2.12</td>
<td>0.96/0.93</td>
<td>1.43</td>
</tr>
<tr>
<td>6</td>
<td>192 (12%)</td>
<td>3.67</td>
<td>1.16/1.11</td>
<td>3.66</td>
</tr>
</tbody>
</table>

The Likert scale for the speaking and writing items functioned well as shown in Tables 4.8 and 4.7. The statistics about the revised scale are listed on the right side of Tables 4.6 and 4.7. As shown below, the observed average ability of the participants increased monotonically in the revised scale. Both infit and outfit mean squares were within the acceptable range from 0.5 to 1.5. The distances between thresholds were greater than 1.4 logits, but less than 5 logits. The monotonic change patterns in threshold can be observed in the category probability curve in Figure 4.1.

Table 4.8
Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Speaking Section (N = 347)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count (%)</th>
<th>Obs. average</th>
<th>Infit/Outfit MNSQ</th>
<th>Structure calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43 (3%)</td>
<td>-5.26</td>
<td>1.32/1.28</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>137 (9%)</td>
<td>-2.75</td>
<td>1.01/1.00</td>
<td>-5.22</td>
</tr>
<tr>
<td>3</td>
<td>338 (21%)</td>
<td>-0.70</td>
<td>0.96/0.97</td>
<td>-2.54</td>
</tr>
<tr>
<td>4</td>
<td>502 (31%)</td>
<td>1.21</td>
<td>0.83/0.81</td>
<td>-0.13</td>
</tr>
<tr>
<td>5</td>
<td>444 (28%)</td>
<td>3.38</td>
<td>0.99/0.99</td>
<td>2.39</td>
</tr>
<tr>
<td>6</td>
<td>136 (9%)</td>
<td>5.24</td>
<td>1.25/1.17</td>
<td>5.51</td>
</tr>
</tbody>
</table>
Table 4.9
Results of Rasch Measurement Analysis: Rating Scale Category Statistics for the Self-assessment Writing Section (N = 347)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count (%)</th>
<th>Obs. average</th>
<th>Infit/Outfit MNSQ</th>
<th>Structure calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25 (2%)</td>
<td>-4.54</td>
<td>1.16/1.16</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>115 (7%)</td>
<td>-2.54</td>
<td>0.98/0.98</td>
<td>-5.03</td>
</tr>
<tr>
<td>3</td>
<td>339 (21%)</td>
<td>-0.70</td>
<td>1.01/1.04</td>
<td>-2.69</td>
</tr>
<tr>
<td>4</td>
<td>532 (33%)</td>
<td>1.16</td>
<td>0.90/0.88</td>
<td>-0.20</td>
</tr>
<tr>
<td>5</td>
<td>467 (29%)</td>
<td>3.24</td>
<td>0.96/0.96</td>
<td>2.29</td>
</tr>
<tr>
<td>6</td>
<td>122 (8%)</td>
<td>5.30</td>
<td>1.14/1.0/1.0/</td>
<td>5.64</td>
</tr>
</tbody>
</table>

Overall, the Rasch analysis results indicate that the self-assessment items have a good quality in terms of reliability, item discrimination, item fit statistics, and scale functioning.

4.1.3. Factor structure of self-assessment and factor scores

In order to examine the factor structure of the self-assessment tool, confirmatory factor analysis was used to assess how well the data fit three theoretically plausible models, namely a correlated four-factor model, a unitary factor model, and a secondary factor model (see Figure 4.2). These three models were proposed based on different views of the constructs of English proficiency (Sawaki, Stricker, & Oranje, 2009; Song, 2008). In Figure 4.2, the ovals represent the latent variables measured in the self-assessment and the rectangular boxes represent the specific self-assessment items.
Figure 4.2. Models tested as potential explanatory models for response patterns on the self-assessment survey (n = 339) of English proficiency as measured by the self-assessment. SA = Self-assessment, Rd or R = Reading, Lsn or L = Listening, Spk or S = Speaking, Wrt or W = Writing. Right: correlated-factor model, Middle: secondary factor model, Lest: unitary factor model

Model identification in confirmatory factor analysis and structural equation modeling refers to the status of the model whether a unique value of each parameter estimate in the hypothesized model can be obtained with known information from the data (Kline, 2011). A model is over-identified when unique values for unknown or freely estimated parameters are derived computationally. The identification of the specified models in this study was checked using the t-rule first and then two-indicator rule, three-indicator rule, or correlated errors rule, when appropriate. The t-rule is a necessary, but not sufficient condition for model identification because it states that the number of freely estimated parameters or the parameters with unknown values should be less than or equal to the number of known pieces (variances and covariances) in the input data for model identification (Bollen, 1989). The other three rules are sufficient, but not necessary conditions of model identification. The three-indicator rule states that for a single
model to be over-identified there must be three indicators or items loading on the factor and the measurement errors are not correlated. The two-indicator rule states that for the models with more than one factor to be over-identified, two indicators are needed and their measurement errors are not correlated. The correlated errors rule is an expansion of two-indicator rule and three-indicator rule and allows for limited correlation among the measurement errors.

Prior to testing the models, the assumptions of both univariate and multivariate normality of the sample data were checked. It was found that the data did not follow a normal distribution. The critical ratio for univariate skewness ranges from -1.106 to 0.012 and the critical ratio for the univariate kurtosis ranges from -8.313 to 0.091. The multivariate kurtosis is 124.274 with a critical ratio 38.566. The Mahalanobis distance was checked for potential multivariate outliers.

The model fit indices for the three proposed models are listed in Table 4.10. The chi-square values for the three models were statistically significant (ρ < .001), which suggests a rejection of the null hypothesis that the variances and covariance generated by the models are the same as the ones in the sample. In other words, these statistics indicated none of the models fit the data. However, the other model fit indices, including χ²/df, CFI, and RMSEA, all indicated an acceptable to good model fit for all of the three models. The values of CFI were over .90 and the values of the RMSEA were less than .08, while only the χ²/df of unitary factor model was over 2.0.

By comparison, the correlated factor model showed the best model fit with smallest χ²/df (2.29), highest CFI (.953), and lowest RMSEA (.062). Because the MLR χ² yielded in the model estimation is an adjusted value and does not follow χ² distribution, scaled difference in χ² was calculated for nested model comparison (Wang & Wang, 2012). The scaled difference χ² test statistic is
\[ TR_d = (T_0 - T_1)/c_d \]

where \( T_0 \) is the maximum likelihood (ML) \( \chi^2 \) for the \( H_0 \) model, \( T_1 \) is the (ML) \( \chi^2 \) for the \( H_1 \) model, \( c_d \) is the difference test scaling correction. \( c_d \) calculated as

\[ c_d = \frac{(d_0 \cdot c_0) - (d_1 \cdot c_1)}{(d_0 - d_1)} \]

where \( d_0 \) is the degree of freedom in \( H_0 \) model, \( c_0 \) is the scaling correction factor for \( H_0 \) model, \( d_1 \) is the degree of freedom in \( H_1 \) model, \( c_1 \) is the scaling correction factor for \( H_1 \) model.

The scaled difference \( \chi^2 \) tests between the correlated factor model and two other models confirmed that the correlated four factor model fitted the data better (\( \Delta ML \chi^2(2)_{M2-M1} = 21.07, p < .001 \) and \( \Delta ML \chi^2(6)_{M3-M1} = 562.81 p < .001 \)).

<table>
<thead>
<tr>
<th>Table 4.10</th>
<th>Model fit indices of the competing models for the self-assessment (n = 339)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>MLR ( \chi^2 )</td>
</tr>
<tr>
<td>Recommended criterion</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>M1 Correlated factor model</td>
<td>370.228 (162)</td>
</tr>
<tr>
<td>M2 Secondary factor model</td>
<td>392.012 (164)</td>
</tr>
<tr>
<td>M3 Unitary factor model</td>
<td>923.690 (168)</td>
</tr>
</tbody>
</table>

Note: CFI = comparative fit index, TLI = Tucker-Lewis Fit Index, SRMR = Standardized Root Mean Square Residual, RMSEA = the root mean square error of approximation.

Nearly all of the standardized factor loadings in the correlated four-factor model were over .07, except for one listening item (.658). The four subskills represented as four latent constructs showed high correlation coefficients among them. The highest correlation is between speaking and writing (.857), followed by the correlation between listening and speaking (.854) and that between reading and writing (.810). Reading had a relatively lower correlation with speaking (.715) and listening (.692).
Once the best fitting model to the self-assessment data was decided, I calculated the scale reliability of the latent measurement or factor rho coefficient of each subscale as alternatives to Cronbach’s alpha, which was reported in Table 4.4 as Person reliability under the Rasch modeling framework. Cronbach’s alpha may not be a dependable estimate of scale reliability, especially with the presence of correlated measurement errors (Raykov, 2001). Scale reliability of the latent measurement or factor rho coefficient is the “ratio of explained variance over total variance that can be expressed in terms of CFA parameters,” which is calculated using the following formula (Kline, 2011, p. 242):

\[
\hat{\rho} = \frac{\left(\sum_{i=1}^{n} \lambda_i \right)^2 \phi}{\left(\sum_{i=1}^{n} \lambda_i \right)^2 \phi + \sum_{i=1}^{n} \theta_{ii} + 2 \sum_{i=1}^{n} \theta_{ij}}
\]

Where \( \hat{\rho} \) is the factor rho coefficient, \( \lambda_i \) is the estimated unstandardized factor loading of the \( i_{th} \) item, \( \phi \) is the estimated factor variance, \( \theta_{ii} \) is the error variance of the \( i_{th} \) item, \( \theta_{ij} \) is the covariance between the \( i_{th} \) item and \( j_{th} \) item if the errors are correlated. All the constructs showed high reliability: The listening section factor rho coefficient is .88, the reading section factor rho coefficient is .91, the writing section factor rho coefficient is .90, and the speaking section factor rho coefficient is .93. The magnitudes of the factor rho coefficients are very close to Cronbach’s alphas in Table 4.4.

Table 4.11
Descriptive Statistics for the Factor Scores of the Self-assessment (\( n = 202 \))

<table>
<thead>
<tr>
<th>Test</th>
<th>Section</th>
<th>M</th>
<th>S.D</th>
<th>Min/Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro-Wilk</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-assessment (^a)</td>
<td>Reading</td>
<td>4.04</td>
<td>0.93</td>
<td>0.94/5.65</td>
<td>-0.616</td>
<td>0.185</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td>3.54</td>
<td>0.94</td>
<td>0.89/5.34</td>
<td>-0.202</td>
<td>-0.246</td>
<td>.107</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speaking</td>
<td>3.82</td>
<td>1.03</td>
<td>0.97/5.79</td>
<td>-0.302</td>
<td>-0.200</td>
<td>.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>3.76</td>
<td>0.89</td>
<td>0.93/5.56</td>
<td>-0.312</td>
<td>0.237</td>
<td>.086</td>
<td></td>
</tr>
</tbody>
</table>

Note. a. The scores of the self-assessment sections were standardized factor scores.
Once the factor structure was determined, a factor score for each subskill was calculated for each participant and then used in the MTMM correlation analysis. In this study, standardized regression-based factor scores were calculated. The descriptive statistics of the factor scores on the self-assessment are listed in Table 4.11.

Overall, the general score pattern in the self-assessment was similar to that of the TOEFL iBT. The highest mean score was on the self-assessment reading section, with a mean of 4.04 out of 6 (67.3%) and a standard deviation of 0.93. The lowest mean score was on the self-assessment listening section, with a mean of 3.54 out 6 (59%) and a standard deviation of 0.94. The mean score of the self-assessment speaking section was 3.82 out of 6 (63.7%) with a standard deviation of 1.03. The mean score of the self-assessment writing section was 3.76 out of 6 (62.7%) with a standard deviation of 0.89. The Shapiro-Wilk test $p$-values associated with the self-assessment indicated that the scores in listening and writing were normally distributed whereas the scores in reading and speaking were not.

To sum up, the results from confirmatory factor analysis indicate that the self-assessment covered four correlated constructs as measured by the items designed to tap four subskills.

4.1.4. Multitrait-multimethod correlation coefficient matrix

Correlation coefficients among the three measures were calculated to form a multitrait-multimethod (MTMM) correlation coefficient matrix. Spearman’s $\rho$ was reported for the correlations that involve the EPT writing grades because the EPT writing grades are ordinal in nature with three levels, namely Engl101B, Engl101C, or Engl101D, and Pass. Pearson’s $r$ was used for other correlations because other scores were interval data.

The multitrait-multimethod matrix with the correlation coefficients for the fall 2014 semester is shown in Table 4.12, which is color-coded for three types of correlations: heterotrait-
monomethod correlations in yellow, heterotrait-heteromethod correlations in blue, and monotrait-heteromethod correlations in green. In addition, the reliability information, if available, is listed on the diagonal line. The monotrait-heteromethod correlations were studied because they revealed the relationship among different measures of the same subskills. In other words, they represent the convergent evidence in MTMM analysis if found to be strong. In the rest of this section, convergent validity evidence (monotrait-heteromethod cells) is reported first, followed by the information about discriminant evidence (heterotrait-heteromethod cells) and test method effect (heterotrait-monomethod cells).

4.1.4.1. Convergent evidence for the EPT validity argument

As shown in Table 4.12, the correlation coefficients between the sections of the EPT and the self-assessment that were intended to measure the same trait in the fall 2014 semester were statistically significant, but only exhibited relatively low to moderate magnitudes in the monotrait-heteromethod cells, ranging from .224 to .375. Since correlation coefficient is affected by the reliability of the measures, disattenuated correlations, or correlations corrected for attenuation, show a more meaningful relationship between two measures without the effect of error (Carr, 2011). A disattenuated correlation is calculated as $r_{xy} / \sqrt{r_{xx} \times r_{yy}}$, where $r_{xy}$ is the correlation coefficient, and $r_{xx}$ and $r_{yy}$ are the respective reliabilities of the measures. After taking account of the reliability of the EPT (.67 for the EPT reading and listening sections, .90 and .89 for the self-assessment reading and listening section), the correlation coefficients in the monotrait-heteromethod cells were slightly smaller than their disattenuated counterparts (.224 versus .288 after correction for attenuation for reading subskill and .318 versus .412 after correction for attenuation for the listening subskills). The highest correlation coefficient was between the writing subskills measured by the EPT and the self-assessment (Spearman’s $\rho =$
.375, or .440 after correction for attenuation), which was followed by the correlation coefficient for listening subskill \( r = .318 \).

Overall, the magnitude of these correlation coefficients suggests that the EPT sections had a relatively weak to moderate relationship with the corresponding self-assessment sections, and that the self-assessment items may have measured somewhat different constructs compared with the corresponding EPT sections.

The monotrait-heteromethod correlations of the same subskills between the EPT sections and the TOEFL iBT sections in the fall 2014 semester ranged from .458 to .519. The disattenuated correlation coefficients for the reading and listening subskills are .684 and .600, respectively. The highest correlation between the EPT and the TOFEL iBT is reading (.519, or .684 after correction for attenuation), followed by writing (Spearman’s rho: .474 or .604 after correction for attenuation). This correlational pattern between the EPT and the TOEFL iBT demonstrated that the EPT tested some shared constructs with the TOEFL iBT and the relationship between these two measures are moderate to strong in Cohen (1988)’s guidelines for interpreting effect size.

The monotrait-heteromethod correlations of the same subskills between the self-assessment sections and the TOEFL iBT sections in the fall 2014 semester ranged from .195 to .368. The highest correlation between the self-assessment and the TOEFL iBT is speaking (.368 or .402 after correction for attenuation), followed by listening (.297 or .330 after correction for attenuation). The lowest correlation between the two measures is writing (.195 or .230 after correction for attenuation). Participants’ self-assessment exhibited statistically significant correlations with the TOEFL iBT, although the overall relationship between them was relatively
weak. This is similar to the correlation coefficient pattern between the EPT and the self-assessment, but with lower coefficients.

A closer look at the monotrait-heteromethod cells, for example, reading skills measured by the EPT and the TOEFL iBT, revealed some convergent evidence. The EPT in fall 2014 had relatively strong relationships with the TOEFL iBT (.458 to .519 or .600 to .684 after correction for attenuation), and weak to moderately strong relationships with the self-assessment sections (.224 to .375 or .228 to .440 after correction for attenuation).
Table 4.12  
*Multitrait-multimethod Correlation Matrix for the Fall 2014 Data (n = 202)*

<table>
<thead>
<tr>
<th></th>
<th>EPT</th>
<th>Self-assessment (SA)</th>
<th>TOEFL iBT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rd</td>
<td>Lsn</td>
<td>Wrt</td>
</tr>
<tr>
<td>EPT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lsn</td>
<td>.514**</td>
<td>(.67)</td>
<td></td>
</tr>
<tr>
<td>Wrt</td>
<td>.498**</td>
<td>.458**</td>
<td>(.79)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * *p < .05, ** p < .01. a. Factor score for each subscale of the self-assessment was used in the correlation analyses. b. The correlation coefficients associated with METP writing are Spearman’s rho and the other correlation coefficients are Pearson’s r. All the correlation coefficients are correlations without correction for attenuation. c. The reliability of each test, if available, is listed on the diagonal.
4.1.4.2. Discriminant evidence for the EPT

The discriminant evidence for the EPT was shown by the heterotrait-heteromethod correlations among the EPT sections and the sections in other measures. Since different measures were used to assess different constructs, the heterotrait-heteromethod correlations are expected to be relatively lower than the monotrait-heteromethod correlations. The correlation coefficients in the heterotrait-heteromethod cells between the EPT and the self-assessment sections in the fall 2014 semester were relatively lower, ranging from .151 (.191 after correction for attenuation) to .338 (.403 after correction for attenuation) with a mean of .277, than those in the monotrait-heteromethod cells between these two measures, ranging from .224 to .375 with a mean of .306.

It is noteworthy that the correlation coefficients (Spearman’s $\rho$) between the writing section of the EPT and other sections of the self-assessment were all statistically significant (.290 to .375). This, to some extent, lends support to the speculation that the self-assessment items at least partially reflected ESL students’ proficiency in English writing. Overall, students’ self-assessment showed significant correlation with the EPT, even though the overall relationship between them was only weak to moderate.

The heterotrait-heteromethod correlations between the EPT and the TOEFL iBT were also smaller in general than the monotrait-heteromethod correlations (.272 to .499 with a mean of .398 for the former versus .458 to .519 with a mean of .484 for the latter). The heterotrait-heteromethod correlations between the corresponding subskills in the self-assessment and the TOEFL iBT ranged from $< .001$ to .361 with a mean of .189, which are smaller in general compared with those in monotrait-heteromethod cells (.195 to .368 with a mean of .265). The correlation coefficients in heterotrait-heteromethod cells were in general smaller than their
counterparts in monotrait-heteromethod cells, which indicates acceptable discriminant evidence for the EPT.

4.1.4.3. Method effect evidence for the EPT

The test method effects were investigated through an analysis of the heterotrait-monomethod correlations within a particular measure. The heterotrait-monomethod correlations among the EPT sections were moderate, ranging from .458 (.630 after correction for attenuation) to .514 (.767 after correction for attenuation), which are moderate to large in correlation magnitude. The heterotrait-monomethod correlations among the TOEFL iBT sections varied from .149 to .512. However, the heterotrait-monomethod correlations among the self-assessment sections were very high, ranging from .741 to .903, which showed a strong method effect. This test method effect associated with the self-assessment in this study is in line with the findings regarding the self-assessment instrument in Bachman and Palmer (1981).

Overall, the EPT had relatively weak to moderate correlations with the self-assessment and moderate correlations with the TOEFL iBT in this study. Compared with the correlation magnitudes in previous studies on the relationship between English placement tests and other measures of English proficiency, this study shows that the relationships between the EPT and the TOEFL iBT as well as the self-assessment are in line with the general findings. The MTMM analyses results suggest that the EPT exhibited some convergent evidence as it measured some shared constructs tapped by the TOEFL iBT and the self-assessment. In addition, the EPT demonstrated certain degree of discriminant evidence as shown by the smaller correlation coefficients in the heterotrait-heteromethod cells than those in the monotrait-heteromethod cells. Furthermore, there was no significant method effect in the EPT. In this sense, the results on the
EPT can be partially extrapolated to similar measures of English proficiency in academic contexts.

Several factors can affect the relationship among the three measures. Firstly, the reliability of each measure affects the strength of correlations between two measures as its square root value sets the theoretical upper limit of the correlations (Bachman, 2004). In this study, the reliabilities of the EPT reading and listening sections were .67, which means that .82 would be the highest possible correlations between the EPT reading or listening sections and other measures. Secondly, constructs measured with the self-assessment may have varied overlaps with the EPT in terms of the measured constructs. All the can-do statements in the self-assessment were phrased with a clear reference to specific academic activities at the university, such as reading textbooks and maintaining discussions with classmates. On the other hand, the constructs defined in the specifications of the EPT and the TOEFL iBT may be somewhat different in terms of target domain descriptions. What is more, the EPT items as well as the TOEFL iBT items may have tested specific linguistic skills in a more analytic way such as the items measuring the skill of identifying detailed information in the listening section on the EPT or the TOEFL iBT.

4.1.5. Correlated trait-correlated methods (CTCM) model and correlated trait-correlated uniqueness (CTCU) model

To further study the relationship between the three measures, a confirmatory factor analysis (CFA) -based MTMM analysis was employed with a correlated trait-correlated method (CTCM) model and a correlated trait-correlated uniqueness (CTCU) model being proposed and tested. However, the analysis of the CTCM model showed two major errors: the residual covariance matrix is not positive definite and the latent variable covariance matrix is not positive
definite. Therefore, the estimated parameters of the CTCM model were problematic, despite its seemingly acceptable model fit shown in Table 4.13. These errors are common in the CTCM model analysis due to its model complexity and these phenomena have been known as a typical issue with the CTCM model (Brown, 2006). In the rest of this subsection, I focused on the results of the CTCU model analysis.

As shown in Figure 4.3 below, the measures of the same subskill are loaded on the corresponding latent trait variable on the left side of the figure and the unique variance or measures or measurement errors are correlated on the right side of the figure (Brown, 2006; Pae, 2012). Measurement errors are usually assumed to be random and their magnitudes determine measurement reliability. The CTCU model includes correlations of the measurement errors under the assumption that the indicators are influenced by the shared measurement methods and certain proportion of the measurement errors can be accounted for by this method effect. One of the advantages of using the CFA-based approach is that it comes up with model fit indices so that researchers can empirically evaluate the extent to which the model fits the data. In this study, the CTCU model was empirically identified, meaning that a unique estimate of model parameter was obtained. Overall, the CTCU model showed a marginally acceptable model fit (see Table 4.13). The WSLMV $\chi^2$ was 62.019 with a degree of freedom of 23 and it was statistically significant ($p < .001$), which suggests that the model did not fit the data well. The value of TLI was .834, below the recommended value of .95. The value of RMSEA was .092 with a 90% confidence interval of 0.065 and .119, which is higher than the recommended values as well. The value of WRMR was 0.696, smaller than one. The value of CFI was .931.
Table 4.13
Model Fit Indices of the Correlated Trait-Correlated Uniqueness (CTCU) and Correlated Trait-Correlated Method (CTCM) Model (n = 202)

<table>
<thead>
<tr>
<th>Model</th>
<th>WSLMV $\chi^2$ (df)</th>
<th>$p$ value</th>
<th>WLSMV $\chi^2$/df</th>
<th>CFI</th>
<th>TLI</th>
<th>WRMR</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended criterion</strong></td>
<td></td>
<td>&gt;.05</td>
<td>&lt;2.0</td>
<td>&gt;.95</td>
<td>&gt;.95</td>
<td></td>
<td>&lt;.08</td>
</tr>
<tr>
<td>CTCU model</td>
<td>62.019 (23)</td>
<td>&lt;.001</td>
<td>2.69</td>
<td>.931</td>
<td>.834</td>
<td>.696</td>
<td>.092 (.065, .119)</td>
</tr>
<tr>
<td>CTCM model</td>
<td>73.855 (24)</td>
<td>&lt;.001</td>
<td>3.08</td>
<td>.913</td>
<td>.800</td>
<td>.513</td>
<td>.100 (.074, .126)</td>
</tr>
</tbody>
</table>

Note. WSLMV = mean and variance-adjusted weighted least square estimator, CTCU model = Correlated trait-correlated uniqueness model, CTCM model = Correlated trait-correlated method model, CFI = comparative fit index, TLI = Tucker-Lewis Index, WRMR = Weighted Root Mean square Residual, RMSEA = the Root Mean Square Error of Approximation.

Figure 4.3. Correlated trait-correlated uniqueness (CTCU) model for the MTMM data.

Table 4.14 shows the standardized parameters estimated from the correlated trait-corrected uniqueness (CTCU) model. The standardized loadings on the trait factors were all significantly different from zero. The range of standard factor loadings for the self-assessment sections was from .387 to .453. This factor-loading pattern matches the low to moderate monotrait-heteromethod correlations discussed above. The range of standardized factor loadings
for the TOEFL iBT was from .474 to .866. The range of the standardized factor loadings for the EPT was from .749 to .892.

Table 4.14
*Standardized Parameter for the Correlated Trait–Correlated Uniqueness (CTCU) Model*

<table>
<thead>
<tr>
<th>Trait Factor Loading</th>
<th>SMC^b</th>
<th>Correlated Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rd</td>
<td>Lsn</td>
</tr>
<tr>
<td>SA_Rd</td>
<td>.453**</td>
<td>.188</td>
</tr>
<tr>
<td>SA_Lsn</td>
<td>.387**</td>
<td>.150</td>
</tr>
<tr>
<td>SA_Spk</td>
<td>.424**</td>
<td>.180</td>
</tr>
<tr>
<td>SA_Wrt</td>
<td>.438**</td>
<td>.192</td>
</tr>
<tr>
<td>EPT_Rd</td>
<td>.757**</td>
<td>.573</td>
</tr>
<tr>
<td>EPT_Lsn</td>
<td>.749**</td>
<td>.561</td>
</tr>
<tr>
<td>EPT_Wrt</td>
<td>.892**</td>
<td>.796</td>
</tr>
<tr>
<td>TI_Rd</td>
<td>.474**</td>
<td>.225</td>
</tr>
<tr>
<td>TI_Lsn</td>
<td>.648**</td>
<td>.420</td>
</tr>
<tr>
<td>TI_Spk</td>
<td>.866**</td>
<td>.750</td>
</tr>
<tr>
<td>TI_Wrt</td>
<td>.582**</td>
<td>.339</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait Factor Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rd</td>
</tr>
<tr>
<td>Lsn</td>
</tr>
<tr>
<td>Spk</td>
</tr>
<tr>
<td>Wrt</td>
</tr>
</tbody>
</table>

Note: a. Rd = Reading, Lsn = Listening, Spk = Speaking, Wrt = Writing, SA = the Self-assessment, EPT = the English Placement Test, TI = the TOEFL iBT. b. SMC, squared multiple correlation (i.e., \( \lambda^2 \)). * significant at p < .05 level, ** significant at p < .01 level.

The lower part of Table 4.14 shows the correlation between trait factors. All of the trait factor correlations are statistically significant. The highest correlation is between listening and writing (\( r = .870 \)), followed by the correlation between writing and reading (\( r = .815 \)). The lowest correlations were between speaking and reading (\( r = .503 \)) and between speaking and writing (\( r = .688 \)). The magnitudes of trait factor correlation can serve as a piece of moderate discriminant evidence. These inter-factor correlations suggest that the subskills were closely related with each other, but were still distinguishable in a substantial way.
The existence of statistically significant correlations between unique variances (uniqueness) or measurement errors suggests significant method effect(s). As shown in Table 4.14, the correlations among the measurement errors in the EPT were not significant, which indicates no method effect in the EPT. This means that the relationship among the individual EPT sections was not noticeably influenced by the test formats used, despite its relatively low reliabilities in the reading and listening section. On the contrary, the self-assessment sections exhibited positive and statistically high correlation, ranging from .743 to .952. This corresponds to the findings from the MTMM correlation matrix analysis and indicates a strong method effect for the self-assessment. The reading and listening sections in the TOEFL iBT had a large positive correlation ($r = .513$) and the correlation of the residuals between the reading and writing sections was also statistically significant, but with a smaller magnitude ($r = .177$). These significant correlations of measurement errors between the TOEFL iBT sections indicated somewhat method effect of the task types used in these three sections, possibly due to the use integrated tasks in the TOEFL iBT, for example, writing an essay based reading and/or listening materials.

Overall, the CTCU model showed an acceptable model fit to the data. Similar to the findings from the MTMM correlation matrix, the results of CTCU modeling indicated that the EPT had relatively high factor loadings on the trait factors and measured some shared constructs with the TOEFL iBT and a small portion of constructs tapped by the self-assessment. There was no statistically significant method effect in the EPT test.

4.1.6. Summary

The quantitative results reported in this section showed that the EPT had a moderate relationship with the TOEFL iBT sections while it also exhibited a weak to moderate relationship
with students’ self-assessment of English proficiency in academic context. The magnitudes of correlations met my expected strength of correlations between the EPT and other two measures based on the review of similar studies. In this sense, the extrapolation inference could be partially established with the reported correlations in this study.

4.2 Test stakeholders’ perceptions

In this section, I analyze the semi-structured interviews with test stakeholders and report the qualitative results regarding test stakeholders’ perceptions of the placement decisions as well as their views on the impact of the EPT placement decisions on ESL students’ English learning and academic achievement. The test stakeholders include eight ESL students as test-takers, three academic advisors who work with ESL students and guide their academic path, five ESL course instructors who teach ESL students in the English Department, and four content course instructors from two major colleges at Iowa State University (Liberal Arts & Science and Engineering). In the following analysis, pseudonyms are used to protect interviewees’ privacy.

4.2.1. Perceptions of the EPT decisions

To elicit test stakeholders’ perceptions of the EPT placement decisions, the interview questions concerning students’ reactions to the placement results were asked in the semi-structured interviews with three groups of test stakeholders: ESL students, ESL course instructors, and academic advisors. Emerging themes are presented for each group of the interviewees.

4.2.1.1. ESL students

ESL students’ perceptions of the EPT decisions were mainly elicited using two main interview questions: “what was your first reaction when you found out the test results? Why?”
and “Did you feel you had been placed in the right ESL class(es)? Why or why not?” Two main themes were identified in the qualitative analysis of the interview responses: frustration in initial reaction to the placement decisions and gradual appreciation of the placement decisions.

Not surprisingly, a typical reaction to the EPT results from the ESL student interviewees was frustration as expressed with words such as “surprised” or “disappointed.” Hua-Chen, who was a graduate student in Business and was required to take Engl99R, Engl101B, and Engl101D, said, “Oh, my god, I need to take three English courses.” This reaction may be partially related to her previous IELTS score of 6.5, which may have given Hua-Chen some confidence in English use. Similarly, Rasha, an undergraduate student in Microbiology placed into Engl99R, Engl101B, and Engl101C, responded that “I was surprised. I thought I could do better.” Yuan-Feng was a Chinese graduate student in Business and was placed into Engl101B and 101D. Yuan-Feng admitted that the EPT results were worse than she expected because she was told by her friends that the EPT was easy and everyone could pass it. Apparently, Yuan-Feng was misled by this false optimism and thus felt disappointed afterwards.

Hyun-Ki, a Korean Engl101D student in Hospitality management, said, “I was kind of disappointed with my writing results. Other than that, I was OK.” Hyun-Ki had received a master’s degree from an American university and his prior learning experience helped build his confidence in writing. Technically, Hyun-Ki should have been exempt from taking the EPT in the first place because of his master’s degree obtained in the U.S. However, he was told to take the test by his department. Austin was an undergraduate student in design and was placed in Engl99L and Engl101C. When he recalled the moment of finding out the EPT results, he said that he was “quite impressed” in a positive way, because he said “I did not think I can make this kind of result with that time pressure.” Austin’s most recent IELTS score was six, which is the
minimum requirement of English language proficiency for admission purposes. It is possible that
his previous IELTS score lowered his expectation of the EPT results. Peng-long was a Chinese
undergraduate student placed into Engl99L, 99L, and 101C. He appeared to be worried when he
found the EPT results and he said, “It was a little too much compared with my friends.”
Nevertheless, Peng-Long admitted that “maybe it is a chance to get [my English proficiency]
impovered [in ESL courses].” Felicity, an Indonesian undergraduate student in design, seemed to
be less surprised at the EPT results because she did not sleep well before the test and she even
fell asleep during the writing test. Like Peng-Long, Felicity tried to see the EPT results in a
positive way and believed that she “overestimated” herself and got to know her real English
proficiency from the EPT.

Sanjar was an Iranian graduate student in Engineering, who was placed into Engl99R and
101D. Sanjar was a little surprised that he was not required to take the listening class because he
thought, “Americans speak too fast [in the EPT test].” Like Felicity, Sanjar’s EPT performance
was greatly affected by non-English proficiency factors. Sanjar said that he was wet and hungry
during the EPT test after moving out from a temporary housing for new students and moving in
to his new apartment in the morning of a raining summer day. Even he was very confident in his
reading ability, Sanjar failed the reading test and was placed in Engl99R. Later he tested out of
Engl99R with a good performance on the in-class test of reading in that class.

These initial reactions are understandable given the fact that all the test-takers had met
the minimum English proficiency requirement at Iowa State University, which is a TOEFL iBT
score of 71 with minimum score of 17 in speaking and writing, and an IELTS score of 6.0 with
no subscore below 5.5 (http://www.admissions.iastate.edu/intl/requirements.php). In addition,
many students did not expect to take the EPT upon their arrival. For example, Yuan-Feng, a
graduate student in Business, acknowledged that she was not aware of the existence of the EPT before she received her admission letter.

Despite the initial reactions, most of the ESL student interviewees showed gradual appreciation of the placement decisions as they tended to agree that they were placed into the right classes in their responses to the question “Did you feel you had been placed in the right ESL class(es)? Why or why not?” Six out of eight ESL student interviewees showed positive attitudes to the placement decisions. Yuan-Feng, who was placed in Engl101B and Engl101D, said, “The placement of ESL class is reasonable since I got guidance from grammar to passage structure, which I was poor at.” Hua-Chen’s response was short, but affirmative about the placement decisions (Engl99R, 101B, and Engl101D): “yes, I think so, especially writing.” Rasha also only gave a short positive response. Peng-Long, an undergraduate student in Electrical Engineering placed in Engl99R, Engl99L, and Engl101C, showed positive attitudes to the placement and said, “Yes, I think so. In fact, it helped me a lot. Our [listening] teacher said, ‘you are so lucky to be placed in this course. Listening is really important part.’” Based on this comment, it appeared to me that the instructor played an important role in convincing Peng-Long that he needed this course to improve his listening. Austin said, “Yes, for me, I am placed in the right class [Engl101C].” However, he followed with a different view on Engl99L and said “except for Engl99L. I do not have much to learn (in that class).” Felicity showed a positive view of the placement decision, but for a different reason. She said that she did not pass the in-class test in both Engl99R and 101C, which convinced her that the academic English requirement in the U.S. was high and she would need these courses.

Hyun-Ki, who questioned the placement of Engl101D, claimed, “I should be in an advanced writing class. I love to write but Engl101D is focused on how to write academic paper,
put your writing in academic skills.” Apparently, he was not necessarily negative toward the placement itself, but he was bothered with the mismatch between his needs or expectations and the course content in Engl101D. Nevertheless, he admitted later that Engl101D helped him improve writing skills. Sanjar was a special case among the ESL student interviewees. He was waived from taking Engl99R with a passing score on the in-class reading test and he had not taken the required Engl101D class yet at the time of the interview. Nevertheless, Sanjar showed strong interest in Engl101D and high expectation of this course. Sanjar said, “I want to take this course [Engl101D]. … Although I have some [published] papers by my own, but I want to know better in introduction and something like that.”

Overall, the ESL student interviewees found the EPT results somewhat surprising to them. Meanwhile, six out of eight interviewees believed that the placement decisions were correct and the other two showed somewhat positive view of the placement decisions as well.

4.2.1.2. ESL instructors

The relevant interview questions for the five ESL instructors were “How satisfied do you think students are with their EPT placement?” and “How motivated are the students in your ESL courses?” The ESL instructors had provided valuable responses to these two questions and four major themes emerged: satisfaction with the placement accuracy, awareness of potential misplacement, existence of negative views of the EPT and ESL courses among the ESL students, and mixed levels of learning motivation in their classes.

Generally speaking, all five ESL instructors agreed that the placement decisions were largely accurate. A representative comment was from Adalet: “I like the test and I think it placed students into classes correctly, because we are never very very surprised with our students. We do not ask ourselves how come this student has been placed into this class.”
It is noteworthy that ESL instructors were more aware of misplaced students in their classes because of their immediate contact with students. Mengqi, who taught a variety of ESL courses such as Engl99R, Engl99L, Engl101B, Engl101C as well as first-year composition course Engl150, shared a similar view about placement accuracy, but mentioned potential misplacement. Mengqi claimed, “90% of the students in my classes are completely justified. They would really need the remedial class. But, 10% of them may not need the course.” Rachael, who taught Engl99L, Engl99R, and Engl150, was confident in the placement of Engl99L students and said, “Engl99L is catching the right kids.” In other words, the placement decisions were plausible as the students in Engl99L were found lacking the skills in listening. However, she was reserved in giving the same confirmation about Engl99R, partly because she believed that Engl99R was too short to evaluate students’ proficiency and better understand the students’ needs.

With more opportunities to interact directly with students, the ESL instructors also have a better chance to learn about ESL students’ attitude towards the EPT placement decisions. Rachael perceived somewhat negative attitudes from ESL students towards the ESL courses. Rachael, who had taught Engl99L and Engl99R before, commented that “overall, from a lot of the students, maybe just the students I encountered, there is a big negative attitude towards the whole process of taking the EPT and having to take the ESL courses and then continuing on with the courses.” Mengqi shared a similar view and she commented, “in general, nobody is really happy about the fact that they have to take ESL classes. But, once they are convinced that the ESL classes will be beneficial to them, most of them are able to adjust their attitudes.” This perception matched six out of eight student interviewees’ initial reaction to placement decisions and their views of the EPT placement decisions as described above.
The ESL instructors noticed mixed levels of learning motivation in their classes. ESL students’ motivation level could be indicative, to some extent, of their attitude toward the EPT placement decisions. Sabriye compared the students’ motivation in Engl99R, Engl99L, and in Engl101D. Her conclusion was that students in Engl99L were the most motivated students and the students in Engl101D, an academic writing course for graduate students only, were the least motivated. Sabriye believed that a possible reason for this is that Engl99L students knew their weakness better because there were many lectures or listening materials that they could not comprehend in both in their content courses and Engl99L, while Engl101D students had different needs, as mentioned by Hyun-Ki. As observed by Mengqi, three quarters of her students were usually motivated whereas the rest “can just stay astray (or pay little attention to the course) from the beginning to the end.” The same motivation issue was also noticed by Young-Soo, an ESL instructor of Engl101C.

4.2.1.3. Academic advisors

Academic advisors at Iowa State University offer students guidance and suggestions to help students reach their academic and career goals. Academic advisors work closely with undergraduates on plans of study and provide specific advice on course selection and registration. Academic advisors become the test stakeholders when they advise ESL students on the course registration at Iowa State University. At Iowa State University, the EPT results are released on Blackboard Learn to the ESL students who have completed the online registration for the test, usually within two days of the test. At the same time, the test results are also made available to academic advisors through AccessPlus, a campus information system used at Iowa State University. ESL students usually meet with their academic advisors for course selection and registration after the EPT results are available.
To better understand the impact of the EPT placement decisions, I interviewed three academic advisors. The following interview questions were used to investigate academic advisors’ interaction with their international advisors regarding the EPT results: “What reactions do international undergraduate students usually have when they received their placement results?” and “What kind of questions do international students usually ask about the EPT?” One major theme emerged from the qualitative analysis is the positive view of the placement decisions.

With regard to ESL students’ reactions to the EPT results, Mike, an adviser in the College of Liberal Arts and Science, commented that ESL students “have been always receptive [to the EPT results].” The main questions Mike received from ESL students were related to the credit points of the ESL courses because the ESL courses credits do not count towards students’ graduation credits based on the policy at Iowa State University. Mike said he had to explain the situation to many students. Todd of the College of Engineering said, “I do not have anybody who has reacted negatively,” and he added, “In my experience here, I do not have any students who argue that placement.” Meanwhile, he faced similar questions from students about ESL course credits, such as “why should I take the courses?” “I do not get credit for that course? Why should I take it?” Mary of the College of Business talked about students’ reactions to the EPT results, based on her observation. She said that that most of the ESL students “seem to be ok with it [the EPT results],” and she had not received many questions concerning the EPT or ESL course from the students. Of course, no question or complaint received by the academic advisors does not necessarily mean ESL students are 100% satisfied with the EPT placement decisions. Some ESL students may not feel comfortable communicating with their academic advisors in their first few encounters.
Overall, the ESL student interviewees showed a mixed feeling toward the EPT placement decisions. On the one hand, some of them found the placement decisions frustrating or surprising, but, they understood why they were placed in ESL courses and could see the potential values in taking the courses. On the other hand, some found a mismatch between their own needs and the ESL course content. As observed by the ESL instructors, most of the ESL students tended to accept the decisions. Similar acceptance was noticed by the interviewed academic advisors, who were not challenged by their international advisees about the EPT decisions. Nevertheless, ESL students’ learning motivation in the ESL courses seemed to vary to some extent and it became an issue in certain ESL courses, as reported by the ESL instructors.

4.2.2. Summary

The qualitative analysis of semi-structured interviews with test stakeholders was conducted with inductive coding for general themes based on the research questions. The qualitative findings indicate that ESL students tended to be receptive to the EPT placement decisions, even if some of them initially held negative attitudes towards the EPT decisions. Overall, ESL course instructors were satisfied with accuracy of placement, especially in the Engl99L course. Nevertheless, they reported some potentially misplaced cases in their classes. Two ESL instructors reported somewhat negative attitudes towards the ESL courses from the ESL students. The ESL instructors noticed mixed levels of learning motivation in their classes.

4.3. Impact of the EPT placement decisions on ESL learners’ English learning

The third question is concerned with the impact of placement decisions on ESL students’ English learning. In other words, this question focuses on whether ESL students have made progress by taking the required ESL courses. In this section, students’ course performance data
from four Engl99L (listening) classes \((n = 38)\), one Engl101B (writing) class \((n = 18)\) and one Engl101C (writing) class \((n = 16)\) are analyzed. More specifically, the scores of the same in-class test administered in Engl99L classes at the beginning and at the end of the fall 2014 semester are also compared using paired-samples \(t\)-tests. The syntactic complexity features and lexical complexity features of the timed essays written in the Engl101B section and Engl101C section at the beginning and the end of the fall 2014 semester are compared using paired-samples \(t\)-tests. Human ratings of the timed essays written at the end of the semester are used to holistically evaluate students’ progress in writing. The essays were rated by trained EPT raters using the same EPT rubric. Therefore, ESL students’ progress in English writing can be directly made through a comparison of the new rating and the original placement.

4.3.1. ESL students’ progress in Engl99L

The same in-class listening test was administered to the students who were enrolled in Engl99L twice as an in-class test at the beginning and as a final exam at the end of the fall 2014 semester. The listening test was prepared by the Engl99L course coordinator and the cut-off score for course exemption was determined jointly by the course coordinator and the supervising professor in the ESL program. The length of the test was 40-45 minutes. The listening materials used in the test consist of one mini-lecture, one dialogue, and 10 short sentences. The items include 13 multiple choice questions, two open-ended questions, and 10 fill-in-the-blank items. In the fall 2014 semester, the cut-off score for course exemption was 65. Unfortunately, the detailed information about the quality of the in-class test as well as the demographic information of the students was not available at the time of this study. This lack of background information will affect the generalizability of the findings to other ESL classes at Iowa State University.
The test scores from four Engl99L sections were collected. The descriptive statistics, along with the results of paired-samples \( t \)-tests, are shown in Table 4.15.

Table 4.15

| Comparison of Test Scores in Engl99L Classes |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|                            | In-class test Mean (SD)     | Final exam Mean (SD)        | Mean diff                   | SD                          | 95% CI                      | \( t \) (df)                  | \( p \)-value | Cohen’s \( d \) |
| Class A \((n = 13)\)       | 36.5 (11.8)                 | 53.8 (12.5)                 | -17.3                       | 8.64                        | -22.5, -12.1                | -7.22 (12)                   | <.001        | 2.00          |
| Class B \((n = 12)\)       | 41.3 (10.6)                 | 55.6 (12.9)                 | -14.3                       | 10.46                       | -21.0, -7.7                 | -4.75 (11)                   | .001         | 1.37          |
| Class C \((n = 10)\)       | 49.7 (10.8)                 | 66.9 (10.0)                 | -17.1                       | 11.34                       | -25.2, -9.0                 | -4.78 (9)                    | .001         | 1.51          |
| Class D \((n = 3)\)        | 47.7 (6.1)                  | 43.3 (6.7)                  | 4.3                         | 8.50                        | -16.8, 25.5                 | 0.88 (2)                     | .471         | 0.51          |
| All classes \((n = 38)\)   | 42.4 (11.8)                 | 57.0 (13.1)                 | -14.6                       | 11.19                       | -18.3, -10.9                | -8.05 (37)                   | <.001        | 1.30          |

The ESL students in Classes A, B, and C made significant progress in terms of test score change in the fall 2014 semester, with an increase in scores from 14.3 to 17.3. A 95% confidence interval of the score differences and Cohen’s \( d \) were reported as effect size for the changes in test scores. Big effect sizes were observed in three sections, ranging from 1.21 to 1.65. Nine out of 38 students or 24% of the students had a score of 65 or higher in the final exam and met the cut-off score for course exemption. However, there was a special case of a class with three students enrolled. These three students actually had a lower score in the final exam. As the instructor commented in an informal communication, these three students were not motivated in the listening class and did not devote adequate time in class activities and assignments.

Despite the fact that the three students in Class D did not benefit from taking Engl99L, the majority of the students in Classes A, B, and C made significant progress in this course, thus partially supporting the positive effect of placing students into Engl99L based on their performance on the EPT listening test.
4.3.2. ESL students’ progress in ESL writing classes

In this subsection, I compared the lexical complexity features and syntactic complexity features in the essays written by ESL students in one section of Engl101B and one section of Engl101C at two time points, namely at the beginning of the semester in an in-class test and at the end of the semester in the final exam. For Engl101B students, the essay prompt was the same in the two tests, whereas the essay prompts were different for Engl101C students (see appendix J for the essay prompts).

The lexical complexity features were analyzed using *Lexical Complexity Analyzer (LCA)* (Lu, 2010) and the syntactic complexity features were analyzed using *L2 Syntactic Complexity Analyzer (L2SCA)* (Lu, 2011). To supplement the objective features from the computational tools, two experienced EPT raters, who also worked as experienced ESL instructors in the ESL program, rated the essays written in the final exam using the EPT rubric. The rating results are reported and discussed following the results of the analysis of objective features.

4.3.2.1. Lexical complexity analysis of students’ essays

Firstly, students’ changes in lexical complexity from their first timed essay to the last timed essay are presented in Tables 4.16 and 4.17 for Engl101B and Engl101C, respectively. There are several variants of the same type of features in the output of *LCA*. For example, there are five types of type-token ratio (TTR), with four of them being mathematically transformed, and four types of verb variation. The variants were usually created to account for the effect of sample size on these measures. More details about and the formulae of the features can be found in Lu (2012). To avoid redundancy in the analysis, I only included 10 lexical complexity features in this study, including one lexical density feature (LD), two lexical sophistication features (LS1, CVS1), and seven lexical variation features (NDWER, CTTR, LV, CVV1, NV, ADJV, and
ADVV). The full names for these features are as follows: LD = Lexical density, LS1 = Lexical sophistication 1, CVS1 = Corrected verb sophistication1, NDWER = Mean new different words of 10 random 50-word samples, CTTR = corrected TTR, LV = Lexical word variation, CVV1 = Corrected, NV = Noun variation, ADJV = Adjective variation, and ADVV = Adverb variation.

Table 4.16 shows the changes in the lexical complexity features from the first timed essay in the in-class test to the last timed essay in the final exam in a section of Engl101B. There were slight increases in three out of 10 lexical features: Lexical density, corrected verb sophistication, and adjective variation. However, none of the increases in lexical complexity features was statistically significant. The only significant change in the lexical complexity features was noun variation (NV), which demonstrated a significant decrease from the in-class test essays to the final exam essays. There were no statistical differences in other lexical complexity features. The speculation about the cause for this phenomenon is that the primary objective of the Engl101B class is to improve grammatical accuracy and paragraph-level writing while only limited attention is paid to explicit teaching of vocabulary in writing (see Appendix H for a sample Engl101B syllabus).
Table 4.16
ESL Student’s Progress in Lexical Complexity in the Engl101B Class (n = 18)

<table>
<thead>
<tr>
<th>Feature</th>
<th>In-class test Mean (SD)</th>
<th>Final exam Mean (SD)</th>
<th>Mean diff</th>
<th>95% CI</th>
<th>t (df)</th>
<th>p-value</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD</td>
<td>0.48 (0.05)</td>
<td>0.49 (0.04)</td>
<td>-0.01</td>
<td>0.03, 0.02</td>
<td>-0.52 (17)</td>
<td>.610</td>
<td>-0.2</td>
</tr>
<tr>
<td>LS1</td>
<td>0.23 (0.07)</td>
<td>0.22 (0.06)</td>
<td>0.01</td>
<td>0.03, 0.05</td>
<td>0.61 (17)</td>
<td>.550</td>
<td>0.13</td>
</tr>
<tr>
<td>CVS1</td>
<td>0.46 (0.19)</td>
<td>0.5 (0.27)</td>
<td>-0.04</td>
<td>0.19, 0.11</td>
<td>-0.53 (17)</td>
<td>.600</td>
<td>-0.13</td>
</tr>
<tr>
<td>NDWER</td>
<td>37.99 (1.68)</td>
<td>37.45 (2.08)</td>
<td>0.54</td>
<td>0.44, 1.52</td>
<td>1.16 (17)</td>
<td>.260</td>
<td>0.27</td>
</tr>
<tr>
<td>CTTR</td>
<td>5.65 (0.73)</td>
<td>5.51 (0.8)</td>
<td>0.14</td>
<td>0.13, 0.41</td>
<td>1.08 (17)</td>
<td>.290</td>
<td>0.25</td>
</tr>
<tr>
<td>LV</td>
<td>0.66 (0.07)</td>
<td>0.62 (0.07)</td>
<td>0.04</td>
<td>0.01, 0.08</td>
<td>1.75 (17)</td>
<td>.100</td>
<td>0.44</td>
</tr>
<tr>
<td>CVV1</td>
<td>3.1 (0.53)</td>
<td>2.98 (0.55)</td>
<td>0.13</td>
<td>0.17, 0.42</td>
<td>0.9 (17)</td>
<td>.380</td>
<td>0.22</td>
</tr>
<tr>
<td>NV</td>
<td>0.64 (0.09)</td>
<td>0.59 (0.07)</td>
<td>0.06*</td>
<td>0, 0.11</td>
<td>2.1 (17)</td>
<td>.050</td>
<td>0.55</td>
</tr>
<tr>
<td>ADJV</td>
<td>0.12 (0.03)</td>
<td>0.13 (0.03)</td>
<td>-0.01</td>
<td>0.02, 0.01</td>
<td>-1.22 (17)</td>
<td>.240</td>
<td>-0.33</td>
</tr>
<tr>
<td>ADVV</td>
<td>0.08 (0.02)</td>
<td>0.08 (0.02)</td>
<td>0</td>
<td>0.02, 0.01</td>
<td>-0.08 (17)</td>
<td>.940</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: * significant at p < .05 level. LD = lexical density; LS1 = lexical sophistication; CVS = corrected verb sophistication; NDWER = new different words in 10 random 50-word samples; CTTR = corrected type-token ratio; LV = lexical variation; CVV = corrected verb variation; NV = noun variation; ADJV = adjective variation; ADVV = adverb variation.

Table 4.17 shows the changes in the lexical complexity features from the first timed essay in the in-class test to the last timed essay in the final exam in a section of Engl101C. There was a statistically significant increase in lexical density (p = .030, Cohen’s d = -0.6). However, there were eight lexical complexity features exhibiting slight, but non-significant decreases from the in-class test to the final exam: lexical sophistication, corrected verb sophistication, new different words in 10 random 50-samples, corrected type-token ratio (TTR), lexical variation, corrected verb variation, adjective variation.
Table 4.17
*ESL Student’s Progress in Lexical Complexity in the Engl101C Class (n = 16)*

<table>
<thead>
<tr>
<th></th>
<th>In-class test Mean (SD)</th>
<th>Final exam Mean (SD)</th>
<th>Mean diff</th>
<th>95% CI</th>
<th>t (df)</th>
<th>p-value</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD</td>
<td>0.48 (0.04)</td>
<td>0.51 (0.05)</td>
<td>-0.03*</td>
<td>-0.06, 0</td>
<td>-2.43 (15)</td>
<td>.030</td>
<td>-0.6</td>
</tr>
<tr>
<td>LS1</td>
<td>0.2 (0.07)</td>
<td>0.17 (0.04)</td>
<td>0.03</td>
<td>-0.01, 0.07</td>
<td>1.73 (15)</td>
<td>.100</td>
<td>0.43</td>
</tr>
<tr>
<td>CVS1</td>
<td>0.53 (0.26)</td>
<td>0.4 (0.21)</td>
<td>0.14</td>
<td>-0.03, 0.3</td>
<td>1.8 (15)</td>
<td>.090</td>
<td>0.45</td>
</tr>
<tr>
<td>NDWERZ</td>
<td>37.73 (1.8)</td>
<td>37.04 (2.12)</td>
<td>0.69</td>
<td>-0.34, 1.72</td>
<td>1.43 (15)</td>
<td>.170</td>
<td>0.36</td>
</tr>
<tr>
<td>CTTR</td>
<td>5.62 (0.61)</td>
<td>5.32 (0.82)</td>
<td>0.3</td>
<td>-0.04, 0.64</td>
<td>1.9 (15)</td>
<td>.080</td>
<td>0.48</td>
</tr>
<tr>
<td>LV</td>
<td>0.66 (0.08)</td>
<td>0.5 (0.1)</td>
<td>0.16*</td>
<td>0.12, 0.19</td>
<td>9.15 (15)</td>
<td>&lt;.001</td>
<td>2.29</td>
</tr>
<tr>
<td>CVV1</td>
<td>3.17 (0.36)</td>
<td>3.02 (0.59)</td>
<td>0.15</td>
<td>-0.17, 0.46</td>
<td>1 (15)</td>
<td>0.330</td>
<td>0.25</td>
</tr>
<tr>
<td>NV</td>
<td>0.64 (0.1)</td>
<td>0.45 (0.11)</td>
<td>0.19*</td>
<td>0.14, 0.23</td>
<td>8.58 (15)</td>
<td>&lt;.001</td>
<td>2.11</td>
</tr>
<tr>
<td>ADJV</td>
<td>0.12 (0.02)</td>
<td>0.1 (0.03)</td>
<td>0.03*</td>
<td>0.01, 0.04</td>
<td>3.57 (15)</td>
<td>&lt;.001</td>
<td>1</td>
</tr>
<tr>
<td>ADVV</td>
<td>0.08 (0.02)</td>
<td>0.08 (0.03)</td>
<td>-0.01</td>
<td>-0.02, 0.01</td>
<td>-0.68 (15)</td>
<td>.500</td>
<td>-0.33</td>
</tr>
</tbody>
</table>

Note: * significant at p < .05 level. LD = lexical density; LS1 = lexical sophistication; CVS = corrected verb sophistication; NDWERZ = new different words in 10 random 50-word samples; CTTR = corrected type-token ratio; LV = lexical variation; CVV = corrected verb variation; NV = noun variation; ADJV = adjective variation; ADVV = adverb variation.

4.3.2.2. Syntactic complexity analysis of students’ essays

The syntactic complexity features were measured using Lu’s L2SCA. Instead of listing all 14 syntactic complexity features in the output of L2SCA, I reported the results regarding seven of the most discriminating features identified in Lu (2011): Mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), coordinate phrases per T-unit (CP/T), coordinate phrases per clause (CP/C), complex nominals per T-unit (CN/T), and complex nominals per clause (CN/C).

The changes in syntactic complexity features in the two timed essays written by Engl101B students are reported in Table 4.18. The essays written in the final exam were slightly
longer than the essays written in the in-class test (328 words versus 320 words). Five syntactic features saw an increase from the in-class test essays to the final exam essays. However, none of the increases was statistically significant. The only syntactic feature that showed a significant change was coordinate phrase per clause (CP/C), which was lower in the essays written in the final exam.

Table 4.18
*ESL Student’s Progress in Syntactic Complexity in the Engl101B Class (n = 18)*

<table>
<thead>
<tr>
<th>Feature</th>
<th>ESL Student's Progress</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-class test</td>
<td>Final exam</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean diff</td>
<td>SD</td>
</tr>
<tr>
<td>W</td>
<td>320.61 (92)</td>
<td>328.83 (69.25)</td>
<td>-8.22</td>
<td>87.77</td>
<td>-0.397 (17)</td>
<td>.696</td>
</tr>
<tr>
<td>MLS</td>
<td>14.67 (2.98)</td>
<td>15.8 (3.31)</td>
<td>-1.13</td>
<td>2.74</td>
<td>-1.743 (17)</td>
<td>.099</td>
</tr>
<tr>
<td>MLT</td>
<td>13.12 (2.57)</td>
<td>13.85 (2.67)</td>
<td>-0.73</td>
<td>1.83</td>
<td>-1.699 (17)</td>
<td>.108</td>
</tr>
<tr>
<td>MLC</td>
<td>7.64 (1.49)</td>
<td>7.8 (1.26)</td>
<td>-0.16</td>
<td>1.67</td>
<td>-0.406 (17)</td>
<td>.690</td>
</tr>
<tr>
<td>CP/T</td>
<td>1.16 (2.45)</td>
<td>0.36 (0.17)</td>
<td>0.8</td>
<td>2.44</td>
<td>1.388 (17)</td>
<td>.183</td>
</tr>
<tr>
<td>CP/C</td>
<td>2.16 (3.43)</td>
<td>0.21 (0.1)</td>
<td>1.95*</td>
<td>3.47</td>
<td>0.22 (3.68)</td>
<td>2.382 (17)</td>
</tr>
<tr>
<td>CN/T</td>
<td>1.18 (0.39)</td>
<td>1.33 (0.4)</td>
<td>-0.15</td>
<td>0.33</td>
<td>-1.883 (17)</td>
<td>.077</td>
</tr>
<tr>
<td>CN/C</td>
<td>0.69 (0.22)</td>
<td>0.75 (0.2)</td>
<td>-0.06</td>
<td>0.26</td>
<td>-0.927 (17)</td>
<td>.367</td>
</tr>
</tbody>
</table>

Note: * significant at p < .05 level. W = words; MLC = mean length of clause; MLS = mean length of sentence; MLT = mean length of T-unit; C = clause; S = sentence; T = T-unit; CP = coordinate phrase; CN = complex nominal

Table 4.19 describes the changes in syntactic complexity features in the two timed essays written by Engl101C students. The essays written in the final exam were much longer than the essays written in the in-class test (476 words versus 338 words) and the difference is statistically significant. Like the developmental pattern identified in Engl101B class, five syntactic features saw an increase from the in-class test essays to the final exam essays. The difference is that in Engl101C class, three cases of the increases in syntactic features were statistically significant:
Mean length per clause, complex nominal per T-unit, and complex nominal per clause. Similar to the findings in Engl101B class, the frequency of the features involving coordinate phrases decreased from the in-class test essays to the final exam essays. However, the decreases were not statistically significant.

Lastly, I turned to the accuracy aspect of students’ writing development in Engl101B and Engl101C sections. I used Criterion, an automated writing evaluation system developed by the Educational Testing Service (ETS), to generate objective feedback on the essays. Criterion is capable of identifying multiple writing features and rule-based errors in the categories of Grammar, Usage, Mechanics, and Style (Link, Dursun, Karakaya, & Hegelheimer, 2014). It is noteworthy that Criterion has also been criticized for failing to identify or missing certain types of errors (Otoshi, 2005). However, in a recent study conducted by Yang, Link, Li, and
Hegelheimer (2015), it was found that the overall accuracy of the grammar error identification was up to 90.7%. In other words, among the identified errors, only 9.3% were incorrectly identified or not actually errors.

In the current analysis, considering the frequency of errors made by the ESL students, I focused only on five most frequent errors made by the ESL students in Engl101B and Engl101C as identified by Criterion, instead of reporting and comparing all the identified errors on Criterion. The five types of errors are Subject-verb agreement, Ill-formed verb, Missing or extra article, Confused words, and Preposition error. The error rates were normalized using the formula of error count/essay length in words × 100. For example, the mean of normalized error of Subject-verb agreement in the in-class test essays in Engl101B was 0.45, meaning that on average there was 0.45 subject-verb agreement error in every 100 words produced by the Engl101B ESL students in their in-class test. The average length of the in-class test essay in Engl101B was 320 words and therefore, the average error count of subject-verb agreement in Egnl101B in-class essays was more than one (0.45 × 3.2 = 1.44).

Tables 4.20 and 4.21 show the changes in normalized errors rates in five errors from the in-class essay test at the beginning of fall 2014 to the final exam in Engl101B and Engl101C, respectively. In the Engl101B class, the normalized error rates in Subject-verb agreement, Confused words, and Preposition error actually increased in the essays written in the final exam, although the increases were not statistically significant. Meanwhile, there were small and statistically non-significant decreases in the normalized error rates in Ill-formed verbs and Missing or extra article.
Table 4.20

ESL Student’s Changes in Normalized Error Rates in the Engl101B Class (n = 18)

<table>
<thead>
<tr>
<th></th>
<th>In-class test Mean(^a) (SD)</th>
<th>Final exam Mean (SD)</th>
<th>Mean diff</th>
<th>SD</th>
<th>95% CI</th>
<th>t (df)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject-Verb agreement</td>
<td>0.46 (0.45)</td>
<td>0.73 (0.76)</td>
<td>-0.27</td>
<td>0.86</td>
<td>-0.69, 0.16</td>
<td>-1.32</td>
<td>.205</td>
<td>-0.31</td>
</tr>
<tr>
<td>Ill-formed verb</td>
<td>0.30 (0.52)</td>
<td>0.28 (0.33)</td>
<td>0.02</td>
<td>0.51</td>
<td>-0.23, 0.27</td>
<td>0.16</td>
<td>.874</td>
<td>0.04</td>
</tr>
<tr>
<td>Missing/Extra article</td>
<td>0.70 (0.52)</td>
<td>0.63 (0.49)</td>
<td>0.07</td>
<td>0.68</td>
<td>-0.27, 0.41</td>
<td>0.44</td>
<td>.667</td>
<td>0.10</td>
</tr>
<tr>
<td>Confused word</td>
<td>0.13 (0.18)</td>
<td>0.27 (0.35)</td>
<td>-0.14</td>
<td>0.42</td>
<td>-0.35, 0.06</td>
<td>-1.47</td>
<td>.161</td>
<td>-0.33</td>
</tr>
<tr>
<td>Preposition error</td>
<td>0.12 (0.28)</td>
<td>0.14 (0.19)</td>
<td>-0.02</td>
<td>0.37</td>
<td>-0.21, 0.16</td>
<td>-0.243</td>
<td>.181</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

Note: \(a\) the error rates are normalized using the formula (error counts/essay length) \(\times\) 100.

Table 4.21

ESL Student’s Changes in Normalized Error Rates in the Engl101C Class (n = 16)

<table>
<thead>
<tr>
<th></th>
<th>In-class test Mean(^a) (SD)</th>
<th>Final exam Mean (SD)</th>
<th>Mean diff</th>
<th>SD</th>
<th>95% CI</th>
<th>t (df)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject-Verb agreement</td>
<td>0.43 (0.42)</td>
<td>0.33 (0.36)</td>
<td>0.10</td>
<td>0.54</td>
<td>-0.19, 0.38</td>
<td>0.725</td>
<td>.480</td>
<td>0.19</td>
</tr>
<tr>
<td>Ill-formed verb</td>
<td>0.18 (0.36)</td>
<td>0.11 (0.19)</td>
<td>0.07</td>
<td>0.41</td>
<td>-0.15, 0.29</td>
<td>0.720</td>
<td>.483</td>
<td>0.17</td>
</tr>
<tr>
<td>Missing/Extra article</td>
<td>0.72 (0.56)</td>
<td>1.08 (0.47)</td>
<td>-0.36</td>
<td>0.70</td>
<td>-0.75, 0.03</td>
<td>-1.98</td>
<td>.067</td>
<td>-0.51</td>
</tr>
<tr>
<td>Confused word</td>
<td>0.12 (0.17)</td>
<td>0.04 (0.09)</td>
<td>0.08</td>
<td>0.17</td>
<td>-0.01, 0.17</td>
<td>1.92</td>
<td>.074</td>
<td>0.47</td>
</tr>
<tr>
<td>Preposition error</td>
<td>0.16 (0.22)</td>
<td>0.22 (0.25)</td>
<td>-0.06</td>
<td>0.32</td>
<td>-0.23, 0.11</td>
<td>-0.74</td>
<td>.470</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

Note: \(a\) the error rates are normalized using the formula (error counts/essay length) \(\times\) 100.

As shown in Table 4.21, decreases in normalized error rates in Subject-verb agreement, Ill-formed verbs, and Confused words were observed in the Engl101C class, although the decreases were not statistically significant. Meanwhile, there were small and statistically non-significant increases in the normalized error rates in Preposition error and Missing or extra article.
Overall, there were no significant improvements in grammar accuracy in ESL students’ writing, as evidenced by the normalized error rates in five error types identified using Criterion.

4.3.2.3. Human ratings of the final exam essays in Engl101B and Engl101C

The last step in evaluating students’ progress in Engl101B and 101C was to invite Adalet, Sabriye, and Hee-Sook, three experienced ESL instructors/EPT raters, to rate the essays written by ESL students in the final exams. Their demographic information can be found in Table 3.3 in Chapter 3. The EPT rubric was used by the human raters to make sure the ratings were comparable to the EPT placement. A one-hour rater training was conducted to ensure rating quality, and the raters finished the rating independently after the training. The ratings from the three raters were consolidated and the final grades were determined with the agreed grade by at least two raters. The intra-class correlation, which functions as an inter-rater reliability index when more than two raters are involved, is .71, which was not high.

For research question 3, I was concerned about whether the ESL students, after taking the required ESL writing courses, reached the expected levels of writing proficiency as described in the EPT scoring rubric. In other words, the question concerned whether the essays written by Engl101B students in the final exam would be rated as “Engl101C” level, the next level for Engl101B students, and likewise, whether the essays written by Engl101C students in the final exam would be rated as “Pass,” the next level for Engl101C students. Table 4.22 contains this information with regard to the “passing rate” of each class rated by the three raters and the final “passing rate” as a result of adjudication. Corresponding to the relatively low inter-rater reliability (.71), it is obvious that Adalet was a lenient rater who rated 16 out 18 Engl101B students as ready for “Engl101C” and 10 out of 16 Engl101C students as “Pass,” whereas Hee-
Sook was a harsher rater who only granted “Engl101C” decisions to 11 out 18 Engl101B students and “Pass” decisions to 6 out of 16 Engl101C students.

<table>
<thead>
<tr>
<th>Class</th>
<th>Adalet</th>
<th>Sabriye</th>
<th>Hee-Sook</th>
<th>Final grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engl101B</td>
<td>16/18 (88.9%)</td>
<td>13/18 (72.2%)</td>
<td>11/18 (61.1%)</td>
<td>14/18 (77.8%)</td>
</tr>
<tr>
<td>Engl101C</td>
<td>10/16 (62.5%)</td>
<td>10/16 (62.5%)</td>
<td>6/16 (7.5%)</td>
<td>8/16 (50.0%)</td>
</tr>
</tbody>
</table>

When all the ratings were taken together, final grades were determined based on the two agreed ratings and were shown as the final grades in the last column of Table 4.22. Overall, the ‘passing rate’ in Engl101B was 77.8% with 14 Engl101B students deemed qualified for Engl101C class, whereas the ‘passing rate’ in Engl101C was lower with a rate of 50.0%, that is eight out of 16 of the Engl101C students were deemed as ‘met the English’ and were ready for first-year composition courses at Iowa State University.

4.3.3. Qualitative Findings regarding the Impact of the EPT decisions on ESL learners’ English learning

ESL students’ performance and progress in the ESL classes can be explained to some extent with a qualitative analysis of the interview data. The analysis results of the interview data with ESL students and ESL course instructors were reported for their perceptions of the impact of the EPT decisions on ESL students’ English learning. Specifically, students’ progress in ESL courses were explored qualitatively based on 1) student interviewee’s responses and ESL instructor interviewees’ responses.

4.3.3.1. ESL students

The ESL student interviewees were asked the following interview questions: “Since this is the end of your first semester, do you feel the ESL courses (Engl99R, Engl99L, and Engl101B/C/D) have prepared you for your study at ISU in terms of English use, for example,
Engl150 and other content courses in the following semesters? If yes, what helped you to improve English? If no, what kind of help would you need?” Their responses can be summarized with one theme: mixed perception of usefulness of the ESL courses.

The interviewed ESL students reported that they improved their English proficiency in the ESL courses, with academic English writing courses (Engl101B, 101C, 101D) being more favorably perceived than English reading and listening courses (Engl99R and Engl99L). For example, Austin pointed out that he “improved much because of Engl101C” and listed several skills he mastered in that course. After taking Engl101C and Engl99R, Felicity believed her English improved and she understood the expectation of academic English writing. Yuan-Feng, a graduate student taking Engl101B in the 2014 fall semester, thought that her English ability improved through peer review and journal tasks. Specifically, she said that Engl101B “armed me more knowledge of words combination” and “I make less mistakes after taking it.” Likewise, despite his concerns over the appropriateness of the placement of Engl101D, Hyun-Ki spoke highly of the course. Hyun-Ki commented, “Engl101D instructor helped me get a sense of how to write a paper” and he felt “a little more comfortable in writing” than at the beginning of the semester. He even recommended Engl101D to be taken by other students because it “is going to make their life much easier.”

Engl99R and Engl99L were less positively perceived by the ESL student interviewees. Felicity mentioned that she did not really use the reading skills taught in Engl99R in dealing with the reading assignments in content courses because she believed that these skills were “not very relevant.” She further explained that these skills were good as her “personal knowledge” and she “may use them in the future.” Felicity did not mention the reading skills or strategies she used in content courses, though. It seems that Felicity understood the value of the skills taught in
Engl99R, but had not used them yet possibly due to different demands in the content courses in her design major. Rasha, after taking both Engl99R and Engl101B, said, “To be honest, Engl99R is a little bit easy for me. I already know the stuff, and I just did my work. It did not help me that much. But, [through] Engl101B, I did improve my writing.” When comparing his learning experience in Engl101C and Engl99L, Austin thought, “Engl101C is huge and helpful.” This positive perception of Engl101C courses is in line with the findings in Todey (2014), who interviewed 20 ESL students who took both Engl101C and Engl150 at Iowa State University and found that the ESL students acknowledged the positive learning outcomes from Engl101C.

The different perceptions of usefulness of ESL courses may be related to the duration of the courses. Engl99R and Engl99L were much shorter than the ESL writing classes. As explained in the next subsection, two ESL course instructors commented on the negative influence of having short courses such as Engl99R and Engl99L.

4.3.3.2. ESL course instructors

To invite ESL course instructors to evaluate students’ learning in ESL courses, the following interview questions were used: “Overall, to what extent do you think the ESL courses helped students improve English?” “How would you evaluate students’ progress in your ESL courses?” “If the ESL students had a chance to take the EPT after finishing the ESL courses, how many of them would have a chance to pass the EPT?” “To what extent do you think the ESL students are prepared for their content courses after they finished the ESL courses?” Two themes emerged from ESL instructors’ responses: noticeable improvement in the target English subskills and varied pass rates or percentages of satisfactory improvement in different classes.

Sabriye and Adalet had experience teaching multiple ESL courses, including Engl99L, Engl99R, and at least one of the ESL writing courses. They felt more optimistic about or
confident in the effectiveness of the academic English writing courses than the reading and
listening courses, which to some degree matches the perceptions of the ESL students on the
impact of the EPT decisions on English learning. For example, Sabriye, who had experiences in
teaching Engl101D, Engl99R, and Engl99L, estimated that “most of the Engl101D students can
pass the test better because at least they got to know how to get the functional language and
vocabulary, how to use the moves and steps, and they raised awareness.” Adalet believed that the
10 out of 15 to 18 students in her Engl101C classes “definitely would pass the test” and she was
also confident that her Engl101C students should be ready for their content courses.

The estimated passing rate in Young-Soo’s Engl101C class was lower than Adalet’s class
and he felt like about 40% to 50% of the students may pass the EPT after taking Engl101C.
Young-Soo found out about 20% of his Engl101C students were not motivated and did not study.
When asked to evaluate students’ readiness of English proficiency for content courses, Young-
Soo pointed out that the students would have a sense of how to write their essays, but they had to
continue studying English to be well prepared. In comparison with her Engl101D class, Sabriye
was much more reserved when evaluating the students in Engl99R and Engl99L. She said, “For
Engl99R and 99L students, I don’t think anything will change because it is only eight weeks and
we only meet 50 minutes twice a week.” Similar to Sabriye’s concern over the progress of her
Engl99R and Engl99L students, Mengqi expressed a low expectation of her students in Engl99R
and Engl99L. Mengqi’s estimated passing rate in these courses was about 30%. She also
mentioned that about 10% of the students in Engl99R and Engl99L did not learning anything at
all, while the other 90% made progress. She believed that about 50% of the students should “be
fine” after taking the courses. Another low passing rate (25%) was estimated by Rachael as well
with regard to her Engl99R and Engl99L students. She attributed this low passing rate to the fact
that “we cannot give them the amount of help that they need to pass the EPT.” However, Rachael held a positive view of the readiness of her students because she had a strong belief in the beneficial effects of content-based English learning. In other words, she thought her students should be able to learn English in content courses as she had already given them a “pass” or equipped them with some strategies to be used in content courses.

Overall, three interviewed ESL instructors expressed concerns about ESL students’ learning motivation in ESL classes, especially in Engl99R and Engl99L classes, which were assumed to be related to students’ improvement in their classes. Low motivation levels may possibly due to a number of factors such as students’ over-estimation of English proficiency, mismatch between students’ needs and the course content, the nature of the ESL courses (credit-bearing versus non-credit bearing), and so on. When students are less motivated mentioned in learning English, it is not realistic to expect much progress or remarkable learning outcomes from the students.

4.3.4. Summary

The quantitative analysis of course performance data in Engl99L, Engl101B, and Engl101C indicated that overall the required ESL courses helped students improve their English proficiency in the targeting subskills, namely listening and academic writing, even there were some students whose improvements were not very noticeable. ESL students in Engl99L had higher scores on the same test administered at the end of the fall 2014 semester. Nevertheless, the passing rate of the four sections was only 24%. The Engl101B and Engl101C students did not make much progress in their writing in terms of lexical complexity. Engl101C students did show more salient increases in their use of syntactic complexity features than Engl101B students did from the in-class test essays to the final exam essays. Human ratings confirmed the progress
in writing proficiency made by the Engl101B and Engl101C students. The passing rates for these two classes were 61.1% and 31.3%, respectively. The effectiveness of ESL courses, especially the academic English writing courses, was acknowledged by ESL students, while some ESL students were less satisfied with Engl99R and Engl99L courses. Similarly, the ESL instructors were generally optimistic about the positive effects of the ESL courses, especially the academic writing courses.

4.4. Impact of the EPT decisions on ESL learners’ academic achievement

The fourth research question addresses the impact of the EPT results on students’ academic achievement, taking into account motivational constructs as mediating variables. Three motivational constructs, namely academic self-efficacy, learning motivation, and self-regulated learning strategies, were measured using the online questionnaire. The quality of these subscales was analyzed using Rasch models and confirmatory factor analysis (CFA). The relationships among the EPT results, GPA, and the motivational constructs were analyzed using structural equation modeling. The sample size of the data set used for research question 4 was 239 after a listwise deletion of the missing cases. This sample size ($n = 239$) is larger than the sample size ($n = 202$) used in the MTMM analysis because this data set did not include the TOEFL iBT information and the self-assessment responses, for which there were more missing cases.

4.4.1. Descriptive statistics and reliability analysis

The reliability information and unidimensionality assumption check are shown in Table 4.23 for the Rasch model analysis of the three subscales. Person reliability in the Rasch model analysis is the equivalent of Cronbach’s alpha. The subscale of academic self-efficacy (ASE) had a high person reliability (.91) and person separation index (3.19). The item reliability of this
subscale is .90, showing a high replicability of item placements along the difficulty scale if the items were to be administered to another similar sample. The item separation index for this subscale is 3.04, showing that the subscale contained at least four groups of item in terms of difficulty using the formula \( \text{Strata} = \frac{4 \times \text{Item separation index} + 1}{3} \). This subscale met the assumption of unidimensionality with 77% of the variance explained and an eigenvalue of 1.8 for the first principal component.

Table 4.23
Scale Level Rasch Measurement Analysis of the Academic Self-efficacy Scale, Motivation Scale, and Self-regulated Learning Strategy Scale (n = 239)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Person Separation</th>
<th>Person reliability</th>
<th>Item Separation</th>
<th>Item reliability</th>
<th>Variance explained</th>
<th>Eigenvalue for 1st contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE (k=5)</td>
<td>3.19</td>
<td>.91</td>
<td>3.04</td>
<td>.90</td>
<td>77.1%</td>
<td>1.8</td>
</tr>
<tr>
<td>MO (k=8)</td>
<td>2.15</td>
<td>.82</td>
<td>2.04</td>
<td>.81</td>
<td>67.4%</td>
<td>2.4</td>
</tr>
<tr>
<td>SRL (k=10)</td>
<td>1.53</td>
<td>.70</td>
<td>6.02</td>
<td>.97</td>
<td>41.8%</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Note: ASE = Academic self-efficacy, MO = Motivation, SRL = Self-regulated learning strategies

The subscale of learning motivation (MO) had an acceptable reliability (.82) and person separation index (2.15). Both the item reliability (.81) and item separation index (2.04) were acceptable as well. However, the learning motivation subscale, as expected, failed to meet the assumption of unidimensionality, which supports the existence of multiple constructs measured by the items. These results are in line with the factor structure of the motivation subscale as stated in the MSLQ manual.

The subscale of self-regulated learning strategies (SRL) had a moderate reliability (.70), which calls for more attention to the quality of this subscale. This subscale showed a somewhat low person separation index (1.53) and its corresponding strata value was 2.37, suggesting the existence of two distinct ability levels on this subscale. The item reliability of this subscale was very high (.97) and similarly the item separation index was high (6.02), indicating that the items in subscale were widely spread on the difficulty scale. Like the subscale of learning motivation,
the SRL subscale did not meet the assumption of unidimensionality (41.8% variance explained and an eigenvalue of 2.7 for the first principal component). These results suggest that more than one construct was measured in the subscale of self-regulated learning strategies, which is in line with the factor structure of this subscale as stated in the MSLQ manual.

The Rasch model analysis was conducted for each subscale to investigate questionnaire item quality. The item level Rasch model analysis results and descriptive statistics are shown in Table 4.24. The item measure or endorsability of the academic self-efficacy subscale ranged from -1.08 to 0.26 logits with a standard error of 0.13. Nevertheless, only one item had a low measure (-1.08) and the other four items were close to each other in terms of measure or endorsability (0.24 to 0.32). All the academic self-efficacy items showed a good model fit with a range of Infit MNSQ from 0.79 to 1.31. In addition, the academic self-efficacy items had high item discrimination with the point-biserial coefficients ranging from .84 to .90.

The Rasch modeling results were echoed in the descriptive statistics for the raw responses (see the right section of Table 4.24). The mean score of the academic self-efficacy subscale ranged from 4.72 to 5.05 on a 6-point scale and the standard deviation varied from 1.03 to 1.08. The responses were slightly skewed as shown by the negative values of skewness (-1.20 to -0.71) and the kurtosis values ranged from 0.10 to 1.63, suggesting a non-normal distribution of the responses to this subscale.
Table 4.24
Item Level Rasch Measurement Analysis of the Academic Self-efficacy Scale (n = 239)

<table>
<thead>
<tr>
<th>Items</th>
<th>Item measure (logit)</th>
<th>S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>P-b coefficient</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE1</td>
<td>-1.08</td>
<td>0.13</td>
<td>1.31</td>
<td>1.22</td>
<td>0.84</td>
<td>5.05</td>
<td>1.03</td>
<td>-1.20</td>
<td>1.63</td>
</tr>
<tr>
<td>ASE2</td>
<td>0.24</td>
<td>0.13</td>
<td>0.98</td>
<td>1.02</td>
<td>0.88</td>
<td>4.74</td>
<td>1.04</td>
<td>-0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>ASE3</td>
<td>0.32</td>
<td>0.13</td>
<td>0.79</td>
<td>0.78</td>
<td>0.90</td>
<td>4.72</td>
<td>1.05</td>
<td>-0.71</td>
<td>0.42</td>
</tr>
<tr>
<td>ASE4</td>
<td>0.26</td>
<td>0.13</td>
<td>0.88</td>
<td>0.87</td>
<td>0.90</td>
<td>4.74</td>
<td>1.08</td>
<td>-0.78</td>
<td>0.47</td>
</tr>
<tr>
<td>ASE5</td>
<td>0.26</td>
<td>0.13</td>
<td>1.04</td>
<td>1.00</td>
<td>0.88</td>
<td>4.74</td>
<td>1.07</td>
<td>-0.72</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Note. ASE = academic self-efficacy, SE = standard error, MNSQ = mean square, p-b coefficient = point-biserial correlation coefficient

The item measure or endorsability of the learning motivation subscale ranged from -0.49 to 0.41 logit with a standard error of 0.08 to 0.09. All the motivation items showed a good model fit with a range of Infit MNSQ from 0.84 to 1.15. In addition, the motivation items had high item discrimination with the point-biserial coefficients ranging from .74 to .83 (see Table 4.25).

The descriptive statistics for the raw responses to the subscale of motivation are shown in the right section of Table 4.22. The mean score of this subscale ranged from 4.36 to 4.99 on a 6-point scale and the standard deviation varied from 1.07 to 1.28. The responses were slightly skewed as shown by the negative values of skewness (-1.41 to -0.70) and the kurtosis values ranged from 0.32 to 1.86, suggesting a non-normal distribution of the responses to this subscale.
Table 4.25
*Item Level Rasch Measurement Analysis of the Motivation Scale* (*n* = 239)

<table>
<thead>
<tr>
<th>Items</th>
<th>Item measure (logit)</th>
<th>S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>P-b coefficient</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN_MO1</td>
<td>0.41</td>
<td>0.08</td>
<td>1.03</td>
<td>1.07</td>
<td>.79</td>
<td>4.36</td>
<td>1.23</td>
<td>-0.70</td>
<td>0.32</td>
</tr>
<tr>
<td>EX_MO2</td>
<td>0.00</td>
<td>0.09</td>
<td>1.05</td>
<td>1.15</td>
<td>.76</td>
<td>4.85</td>
<td>1.07</td>
<td>-0.99</td>
<td>1.11</td>
</tr>
<tr>
<td>IN_MO3</td>
<td>-0.49</td>
<td>0.09</td>
<td>1.17</td>
<td>1.12</td>
<td>.74</td>
<td>4.83</td>
<td>1.16</td>
<td>-1.29</td>
<td>1.86</td>
</tr>
<tr>
<td>IN_MO4</td>
<td>0.34</td>
<td>0.08</td>
<td>0.85</td>
<td>0.87</td>
<td>.80</td>
<td>4.41</td>
<td>1.20</td>
<td>-0.73</td>
<td>0.51</td>
</tr>
<tr>
<td>EX_MO5</td>
<td>0.20</td>
<td>0.09</td>
<td>0.91</td>
<td>0.93</td>
<td>.82</td>
<td>4.74</td>
<td>1.26</td>
<td>-1.03</td>
<td>0.71</td>
</tr>
<tr>
<td>IN_MO6</td>
<td>-0.26</td>
<td>0.09</td>
<td>1.02</td>
<td>1.01</td>
<td>.77</td>
<td>4.73</td>
<td>1.16</td>
<td>-0.86</td>
<td>0.60</td>
</tr>
<tr>
<td>EX_MO7</td>
<td>-0.31</td>
<td>0.09</td>
<td>1.18</td>
<td>1.11</td>
<td>.76</td>
<td>4.99</td>
<td>1.21</td>
<td>-1.41</td>
<td>1.76</td>
</tr>
<tr>
<td>EX_MO8</td>
<td>0.11</td>
<td>0.09</td>
<td>0.84</td>
<td>0.84</td>
<td>.83</td>
<td>4.79</td>
<td>1.28</td>
<td>-1.07</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Note. EX_MO = extrinsic motivation, IN_MO = intrinsic motivation, SE = standard error, MNSQ = mean square, p-b coefficient = point-biserial correlation coefficient

The item measure or endorsability of the self-regulated learning strategies subscale ranged from -0.58 to 0.67 logit with a standard error of 0.06 to 0.08. All the self-regulated learning strategy items showed a good model fit with a range of Infit MNSQ from 0.82 to 1.30.

By contrast, the self-regulated learning strategy items had lower item discrimination than the items in other two subscales, with the point-biserial coefficients ranging from .59 to .76.

The descriptive statistics for the raw responses to the subscale of self-regulated learning strategies are shown in the right section of Table 4.26. The mean score of this subscale ranged from 3.06 to 4.66 on a 6-point scale and the standard deviation varied from 1.16 to 1.52. The responses were skewed as shown by the skewness values (-0.90 to 0.44) and the kurtosis values ranged from -1.02 to 0.58, suggesting a non-normal distribution of the responses to this subscale.
Table 4.26
Item Level Rasch Measurement Analysis of the Self-regulated Learning Strategy Scale (n = 239)

<table>
<thead>
<tr>
<th>Items</th>
<th>Item measure (logit)</th>
<th>S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>P-b coefficient</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kur-tosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRL1</td>
<td>0.21</td>
<td>0.07</td>
<td>0.89</td>
<td>0.88</td>
<td>.76</td>
<td>4.43</td>
<td>1.25</td>
<td>-0.85</td>
<td>0.58</td>
</tr>
<tr>
<td>SRL2</td>
<td>-0.13</td>
<td>0.08</td>
<td>1.14</td>
<td>1.08</td>
<td>.70</td>
<td>4.66</td>
<td>1.22</td>
<td>-0.90</td>
<td>0.48</td>
</tr>
<tr>
<td>SRL3(R)</td>
<td>-0.34</td>
<td>0.06</td>
<td>0.83</td>
<td>0.86</td>
<td>.73</td>
<td>4.13</td>
<td>1.48</td>
<td>-0.35</td>
<td>-1.00</td>
</tr>
<tr>
<td>SRL4</td>
<td>-0.08</td>
<td>0.08</td>
<td>0.96</td>
<td>0.96</td>
<td>.73</td>
<td>4.63</td>
<td>1.16</td>
<td>-0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>SRL5(R)</td>
<td>0.25</td>
<td>0.06</td>
<td>0.95</td>
<td>0.90</td>
<td>.71</td>
<td>3.48</td>
<td>1.51</td>
<td>0.08</td>
<td>-1.02</td>
</tr>
<tr>
<td>SRL6</td>
<td>0.29</td>
<td>0.06</td>
<td>0.96</td>
<td>0.98</td>
<td>.73</td>
<td>3.96</td>
<td>1.33</td>
<td>-0.27</td>
<td>-0.57</td>
</tr>
<tr>
<td>SRL7(R)</td>
<td>0.67</td>
<td>0.06</td>
<td>1.30</td>
<td>1.30</td>
<td>.59</td>
<td>3.05</td>
<td>1.36</td>
<td>0.44</td>
<td>-0.50</td>
</tr>
<tr>
<td>SRL8</td>
<td>0.06</td>
<td>0.07</td>
<td>0.97</td>
<td>0.94</td>
<td>.71</td>
<td>4.15</td>
<td>1.30</td>
<td>-0.47</td>
<td>-0.52</td>
</tr>
<tr>
<td>SRL9(R)</td>
<td>-0.58</td>
<td>0.06</td>
<td>0.86</td>
<td>0.82</td>
<td>.75</td>
<td>4.37</td>
<td>1.52</td>
<td>-0.61</td>
<td>-0.73</td>
</tr>
<tr>
<td>SRL10</td>
<td>-0.35</td>
<td>0.07</td>
<td>1.05</td>
<td>0.99</td>
<td>.69</td>
<td>4.49</td>
<td>1.23</td>
<td>-0.78</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Note. SRL = self-regulated learning strategies, (R) indicates an reverse coded item, SE = standard error, MNSQ = mean square, p-b coefficient = point-biserial correlation coefficient

The category probability curves were used to visually check the functionality of the 6-point Likert scale used in the subscales (see Figure 4.4). The leftmost graph shows a clear progressive pattern of the curves, which represent each category in the subscale of academic self-efficacy. However, the graph for the motivation subscale shows some overlap of category 1 (strongly disagree) and category 2 (slightly disagree) on the left side of the graph. This suggests that these categories were less distinguishable and collapsing these two categories may yield results that are more accurate. The rightmost graph in Figure 4.4 is for the subscale of self-regulated learning strategy (SRL). It is noticeable that the intervals between categories were fairly narrow, which indicates that the Likert scale in this subscale may not be as distinguishable as expected.
4.4.2. Confirmatory factor analysis of the subscales

This subsection reports the results of confirmatory factor analysis of each subscale. This modeling procedure is also called testing a measurement model in the framework of structural equation modeling.

Based on the literature as well as the original manual of the MSLQ, I hypothesized the factor structure of each subscale in the questionnaire. The hypothesized models can be found in Figure 4.5. For example, five academic self-efficacy items were proposed to load on a single latent variable called academic self-efficacy. The model for the motivation subscale contained two latent variables, representing intrinsic motivation and extrinsic motivation. Four items loaded on each of these two latent variables, respectively. As for the subscale of self-regulated learning strategies, three correlated factors were proposed to represent three sub-constructs: effort-regulation, time management, and help-seeking.
The multivariate normality assumption was violated in the data sets for the three subscales, as shown by statistically significant multivariate kurtosis values. For example, the multivariate kurtosis for the academic self-efficacy subscale was 24.198 with a critical value of 22.450; the multivariate kurtosis for the motivation subscale is 43.786 with a critical value of 26.869; the multivariate kurtosis for the self-regulated learning strategy subscale is 23.707 with a critical value of 11.878.

To account for the non-normality issue in the data, a robust version of maximum likelihood estimator (MLR) was used in Mplus 7.0. Table 4.2 contains the model fit indices for each hypothetical model for the subscales. The proposed single factor model for academic self-efficacy showed a good model fit (MLR $\chi^2 = 3.579$, df = 4, $p = .446$, CFI = 1.000, TLI = 1.002, SRMR = .009, RMSEA = .000, 90% CI: .000, .079). In addition, the standardized factor loadings of the individual items on the latent variable academic self-efficacy ranged from .802 to .893.

The proposed correlated two-factor model for the motivation subscale showed an acceptable model fit, although its chi-square value was statistically significant (MLR $\chi^2 =$...
68.572, df = 17, p < .001), the TLI value was lower than .95 (TLI = .877), and the RMSEA value was slightly over .08 (RMSEA = .096, 90% CI: .073, .120). Other model fit indices supported this model (CFI = .925, SRMR = .053). The standardized factor loadings of the individual items on the latent variable Intrinsic Motivation ranged from .693 to .785, and the standardized factor loadings on Extrinsic Motivation ranged from .657 to .783. The correlation between intrinsic motivation and extrinsic motivation was .768.

Table 4.27

<table>
<thead>
<tr>
<th>Model</th>
<th>MLR $\chi^2$ (df)</th>
<th>$p$ value</th>
<th>MLR $\chi^2$/df</th>
<th>CFI</th>
<th>TLI</th>
<th>SRMR</th>
<th>RMSEA 90% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended criterion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASE unitary-factor model ($n = 331$)</td>
<td>3.579 (4)</td>
<td>&lt;.05</td>
<td>&lt;2.0</td>
<td>&gt;.95</td>
<td>&gt;.95</td>
<td>&lt;.05</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>MO correlated 2-factor model ($n = 331$)</td>
<td>68.572 (17)</td>
<td>&lt;.001</td>
<td>4.03</td>
<td>&gt;.95</td>
<td>.925</td>
<td>.877</td>
<td>.053 (.073, .120)</td>
</tr>
<tr>
<td>SRL correlated 3-factor model ($n = 322$)</td>
<td>320.563 (31)</td>
<td>&lt;.001</td>
<td>10.34</td>
<td>.433</td>
<td>.177</td>
<td>.132</td>
<td>.165 (.148, .182)</td>
</tr>
</tbody>
</table>

Note. ASE = academic self-efficacy, MO = motivation, SRL = self-regulated learning strategies, CFI = comparative fit index, TLI = Tucker-Lewis Index, SRMR = Standardized root mean square residual, RMSEA = the root mean square error of approximation.

A three-factor model for the self-regulated learning strategy scale was proposed based on the manual of the MSLQ (Pintrich, Smith, Garcia, & McKeachie, 1991). However, this model yields an abnormal parameter estimate (a factor correlation value exceeds 1) because the latent variable of covariance matrix is not positive definite. This is usually caused by wrong model specification. In this case, the model fit indices were very poor (MLR $\chi^2 = 320.563, df = 31, p < .001, CFI = .433, TLI = .177, SRMR = .132, RMSEA = .165, 90% CI: .148, .182$). In addition, the parameter estimates were not stable due to the estimation error and thus not reported. These model fit results were surprising for a well-established questionnaire like the MSLQ. This finding may be related to the target population because the MSLQ was originally designed for
native English-speaking college students and the participants in my study were international students at Iowa State University. Furthermore, respondent fatigue and boredom may have affected the quality of the data with certain random responses. This subscale of self-regulated learning strategies was placed at the end of the questionnaire, which consisted of 43 Likert scale-based statements and a number of questions for demographic information. Participants may have been tired or bored after completing the self-assessment items and other two subscales (academic self-efficacy and motivation), thus may have paid less attention to the subscale of self-regulated learning strategies.

Given the model testing results, I decided not to include this self-regulated learning construct in the follow-up structural equation model analysis. Other considerations were also taken into account in this decision. For example, the Rasch model analysis described above also indicated that this subscale had a lower reliability (0.70) and the categories in the Likert scale in this subscale were less distinguishable. I could have revised the model based on the modification index in the output of Mplus or explored the factor structure using exploratory factor analysis (EFA). However, model hunting in this fashion deviates from the practice of CFA because models of interest should be proposed a priori in CFA (Brown, 2006). In addition, I would need a new data set to test the revised model if EFA was used with this data.

4.4.3. Structural equation modeling analysis

To investigate the relationship among the latent variables, I conducted a correlation analysis of these variables. Table 4.2 shows the correlations among the latent variables and GPA. The highest correlation was between academic self-efficacy and intrinsic motivation (.819), followed by that between intrinsic motivation and extrinsic motivation (.680) and that between academic self-efficacy and extrinsic motivation (.560). The EPT results had a
moderation correlation with academic self-efficacy (.293) and with the GPA (.313). The correlations between motivational factors and the GPA were not statistically significant.

Table 4.2
*Correlation Coefficients among the Latent Variables (n = 273)*

<table>
<thead>
<tr>
<th></th>
<th>EPT</th>
<th>ASE</th>
<th>MO_EXT</th>
<th>MO_INT</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASE</td>
<td>.293**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MO_EXT</td>
<td>-.040</td>
<td>.560**</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MO_INT</td>
<td>.182*</td>
<td>.819**</td>
<td>.680**</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>.313**</td>
<td>.044</td>
<td>-.109</td>
<td>.001</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: EPT = the English Placement Test, ASE = Academic self-efficacy, MO_EX = Extrinsic motivation, MO_IN = Intrinsic motivation, GPA = Grade Point Average.

Based on the findings in previous studies on the contributing factors to academic achievement, a simplified model was proposed to reflect the relationship among the EPT results, academic self-efficacy, motivation, and cumulative GPA (see Figure 4.6). The ovals in Figure 4.6 represent latent variables and the rectangles represent observed variables, which are not shown in the figure for the sake of simplicity. It is hypothesized that the EPT results, representing initial English proficiency at the beginning of the semester, has a direct impact on students’ cumulative GPA (4-point scale) because previous studies found that English proficiency affects students’ academic achievement (Cho & Bridgeman, 2013; Lee & Greene, 2007). The EPT results are hypothesized to have a direct impact on students’ academic self-efficacy and their extrinsic motivation because of the relationship between their English proficiency and their GPA. It is also hypothesized that students’ academic self-efficacy, in turn, has direct impact on students’ motivation, which has a direct impact on students’ cumulative GPA.

As described in 4.4.2, the confirmatory factor analysis of the subscale of self-regulated learning strategies indicated that its factor structure or model specified in the MSLQ manual did
not fit the data in this study. Considering its problematic psychometric properties in terms of reliability and scale functionality, this subscale was not included in the structural equation modeling analysis. For this reason, the original proposed model in Figure 4.6 was adjusted accordingly. A fully recursive model, which all the elements are linked without any mutual influence or feedback loop, was tested because it is structurally similar to the original proposed model. The fully recursive model was tested with the data from 273 participants after listwise deletion of missing data. WLSMV estimator was used in model testing because the writing grades on the EPT were categorical data.

As shown in Table 4.29, the fully recursive model had an acceptable model fit to the data (WLSMV $\chi^2 = 233.884$, df = 106, $p < .001$, WLSMV $\chi^2$/df = 2.21, CFI = .867, TLI = .829, WRMR = .773, RMSEA = .066, 90% CI: .055, .078), even some of the model fit indices were not satisfactory.

<table>
<thead>
<tr>
<th>Model</th>
<th>WLSMV $\chi^2$ (df)</th>
<th>$p$ value</th>
<th>WLSMV $\chi^2$/df</th>
<th>CFI</th>
<th>TLI</th>
<th>WRMR</th>
<th>RMSEA</th>
<th>90% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recommended criterion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully recursive model</td>
<td>233.884 (106)</td>
<td>&lt;.001</td>
<td>2.21</td>
<td>.867</td>
<td>.829</td>
<td>.773</td>
<td>.066</td>
<td>(.055, .078)</td>
</tr>
<tr>
<td>Trimmed model</td>
<td>224.522 (108)</td>
<td>&lt;.001</td>
<td>2.08</td>
<td>.879</td>
<td>.847</td>
<td>.777</td>
<td>.063</td>
<td>(.051, .074)</td>
</tr>
</tbody>
</table>

Note: WLSMV = mean and variance-adjusted weighted least square estimator, CFI = comparative fit index, TLI = Tucker-Lewis index, WRMR = Weighted root mean square residual, RMSEA = the root mean square error of approximation, C.I. = confidence interval.

The standardized path coefficients and their associated standard errors for the fully recursive model can be found in Figure 4.6. As expected, the EPT results showed a positive and statistically significant impact on the first semester GPA. The standardized path coefficient is
.301, meaning that we can expect an increase of .301 standard deviation in the GPA with an increase of one standard deviation in the EPT results (factor score). Overall, this model could account for 10.5% of the variance in the first semester GPA. This is understandable given the large number of confounding factors contributing to students’ academic achievement.

Also shown in Figure 4.6 is that the EPT results had a significant impact on students’ academic self-efficacy (standardized path coefficient = .284, \( p < .001 \)). At the same time, the EPT results showed a significant but negative impact on students’ extrinsic motivation (standardized path coefficient = -.208, \( p < .001 \)) and a non-significant and negative impact on intrinsic motivation (standardized path coefficient = -.046, \( p = .433 \)). These parameters suggested that when the ESL students’ EPT performance was good, they tended to have higher confidence in their learning in general. With higher English proficiency, the ESL students’ learning was less performance-oriented (lower extrinsic motivation), however, English proficiency did not affect their intrinsic motivation or goal orientation. Students’ academic self-efficacy did not predict their first semester GPA (standardized path coefficient = -.016, \( p = .906 \))
Figure 4.6. A simplified fully recursive model of the relationship among the EPT, motivational factors, and cumulative GPA, with standardized parameters, ** p < .01, EPT = the English Placement Test, ASE = Academic self-efficacy, MO_EX = Extrinsic motivation, MO_IN = Intrinsic motivation, GPA = Grade Point Average.

Academic self-efficacy had a significant and strong impact on both extrinsic motivation (.625, p < .001) and intrinsic motivation (.823, p < .001), which suggested that higher learning confidence predicted higher learning motivation, but comparatively more impact on intrinsic motivation than extrinsic motivation. This is understandable because students with higher self-confidence in their learning may perform well in classes and focus more on mastery of the content, thus showing less worry about their course performance. However, both extrinsic and intrinsic motivation factors did not have a significant impact on the first-semester cumulative GPA (standardized path coefficient = -.119, p = .114, and .045, p = .751).

Based on the path coefficients in the fully recursive model, I dropped the statistically insignificant path from the EPT results to intrinsic motivation, and kept two other paths from
motivation to GPA for their theoretical significance. The trimmed model showed similar model fit indices and parameter estimates (see Table 4.29 and Figure 4.7). Furthermore, a chi-square difference test using the DIFFTEST option for WLSMV estimator in Mplus indicated that the trimmed did not fit the data significantly better than the fully recursive model ($\Delta \chi^2 = 1.283$, df = 2, $p = .527$), although it was more parsimonious. Therefore, I used the fully recursive model as my final model in this study to show the complete picture of the relationship among the latent variables.

![Diagram](image)

**Figure 4.7.** A simplified and trimmed model of the relationship among the EPT, motivational factors, and cumulative GPA, with standardized parameters. **$p < .01$** level, EPT = the English Placement Test, ASE = Academic self-efficacy, MO_EX = Extrinsic motivation, MO_IN = Intrinsic motivation, GPA = Grade Point Average.
4.4.4. Qualitative Findings regarding the Impact of the EPT decisions on ESL learners’ academic achievement

The relationship between ESL students’ English proficiency as measured with the EPT and their first-semester GPA can be partially accounted for by the interview responses from the ESL students and content course instructors. This subsection reports test stakeholders’ views on the impact of the EPT decisions on ESL learners’ academic achievement. A key question asked to all the test stakeholders concerned the relationship between English proficiency and academic achievement or course grades.

4.4.4.1. ESL students

Two major themes emerged from ESL students’ responses: close connection between English proficiency and academic achievement, and the challenge of technical terminology in content courses.

Not surprisingly, all the interviewed ESL student saw a close connection between English proficiency and academic achievement. Hua-Chen, the business graduate student in Engl101B, admitted that her low listening ability considerably affected her course performance. She said, “I have to listen carefully and follow closely with the instructor. A slip of mind can cause a lot problems.” Hua-Chen added that her English-speaking ability made the situation even worse because when she did not understand the instructor and wanted to ask questions, her English was not good enough to help her express the ideas. Yuan-Feng, who was placed in Engl101B and Engl101D, stressed the impact of low writing ability on her course performance and took it as her major problem. She thought that taking Engl101B helped her prepare written assignments in content courses, which contributed to a higher GPA.
Peng-Long, the undergraduate student in Electric Engineering, noticed that English ability affected his performance in Chemistry because of his unfamiliarity with the technical terms in that course. In addition, he admitted that he had trouble joining class discussions due to “fear” or lack of confidence. With these experiences, Peng-Long was very appreciative of taking ESL courses and he explained, “If you did not get well prepared, you will easily fall behind. So, I prefer that the EPT is good for my future English learning and to get the English ability.”

Alex was a special case found in the interview. As an undergraduate student in Design, Alex believed that his English level did not affect his performance in content courses. However, he did mention that he had difficulty understanding some assignments when a professor gave too many details about the assignments orally in class. Alex ranked Psychology as a course with a high requirement in English whereas courses like Math and Design only required basic skills in English.

Rasha talked about her motivation for taking ESL courses in the interview. She said, “I was really happy because I already have an A in that course [Engl101B]. In other courses, I got a B-. This grade [A] will help me a lot.” Like Rasha, Hua-Chen as a graduate student regarded the ESL courses as a helpful way from which she could earn a good grade, which could help boost her overall GPA.

Three the interviewed ESL students mentioned that their challenges in content sometimes were the technical terminology. Felicity, an undergraduate student in Design placed in Engl99R and Engl101C shared her experience with “big word” in classes. She said “If they [content course instructors] use very very big words and some words that I’ve never heard of, sometimes, I think ‘what the hell is this word’. … … For example, in my elective course, I do not know what it means, I would go ask my feelings or maybe ask my supervisors.” Peng-Long, an
undergraduate in Engineering placed into Engl99R, Engl99L, and Engl101C, had similar experiences. He said, “It [English proficiency] affects the chemistry course. Like some specific words, there are so many difficult words, like elements in Chemistry.” In addition, Peng-Long talked about how English proficiency limited his participation in classes: “I think I had trouble joining class discussion, some fears. I am not confident in talking to Americans.”

4.4.4.2. ESL instructors

One theme that was emerged from ESL instructors’ responses was similar to the one from ESL students, that is, a recognition of the close connection between English proficiency and academic achievement. In addition, another theme was about the threshold effect of English proficiency on academic achievement.

The close relationship between ESL students’ English proficiency and their academic achievement was also acknowledged by the ESL instructors in the interviews. Adalet believed that “the more proficient they [the ESL students] are, the more successful they would be. Because you express yourself with language in every assignment, so there is a very high correlation.” Echoing this, Rachael said, “I do think there is a connection. My [Engl99L] students who cannot follow the syllabus and cannot follow the lecture, they are foundering and they even did not know that they had to come to class.”

Mengqi shared a slightly different opinion on the relationship between English proficiency and academic achievement, saying, “If their proficiency isn’t ‘appallingly’ low, that is, unable to comprehend written or oral input in classes I actually would like to positively conjecture that they will actually be fine with their academic achievement.” Mengqi continued, “May not for their first few semesters, but after that they should be OK.”
4.4.4.3. Content course instructors

The input from content course instructors also shed light on the relationship between ESL students’ English proficiency and academic achievement. Content course instructors are usually in a better position to comment on the relationship between ESL students’ English proficiency and their course performance because the instructors have direct contact with both ESL students and native English-speaking students. The major themes from the content course instructors’ responses were the important role of English proficiency and possible disciplinary difference in the levels of English proficiency requirement.

Feng-Chun, a non-native English-speaking instructor in the Department of Mathematics, said, “Actually, I don’t think my math classes are demanding in English. English for us is just a tool. Most times, we use numbers.” Feng-Chun also commented, “But, we also need to communicate” and “if they [ESL students] had better English skills, it would be easier for them to follow the lectures. They can get the idea fast. … It takes a little bit time for them to follow.” This advantage of having better English proficiency was also mentioned by the instructors of engineering-related courses and computer science courses. Joseph, a post-doctorate associate who teaches fundamental courses in Civil Engineering, observed, “International students usually struggle more” and commented, “if they had a better English proficiency, they would probably do better. For example, on a test, they may understand the questions better. If they are not proficient, they may have a harder time.” Joseph also believed that “as long as they [ESL students] try hard and apply themselves to the work, it should not be an issue. It just seems like how much effort students are willing to put to work.” Ganliang, a non-native English-speaking associate professor in Computer Science, regarded his computer science courses to have a low English requirement because they were just computer programming and did not have writing.
However, like Feng-Chun, Ganliang also highlighted the essential role of communication in his classes because students needed to talk to teaching assistants and classmates. In terms of the relationship between English proficiency and academic achievement, Ganliang’s view was less affirmative. Ganliang said, “I do not see a quite correlation even I can see English can hinder students’ understanding English or doing assignments well.” This may be related to his observation in classes that personality affected students’ performance in a more salient way especially when students had similar English proficiency levels.

The close relationship between English proficiency and academic achievement was revealed in the Psychology courses taught by Kathleen, who claimed that these two were “pretty correlated because it [psychology] is such a verbal area.” She said, “A student who is pretty dedicated can do a lot in her own time, even [if] her English proficiency is not high. … It is definitely going to be harder, and it takes longer.” With regard to students’ course grades, Kathleen assumed that “probably the [ESL] students could not get an A. They may receive a B or possibly a C, instead. So, there is a relationship between these two areas.”

As shown above, the interviewed content course instructors acknowledged the close connection between ESL students’ English proficiency and their course performance. On the other hand, the content course instructors noticed other characteristics of the ESL students in general that can help them succeed in classes. For example, Feng-Chun, the mathematics instructor, praised the international students in her classes and said, “I think they are more hardworking. Compared with American students, Asian students and international students are self-paced. They barely miss classes, but they are quiet.” This characteristic was also mentioned by Kathleen, the instructor in Psychology. Kathleen mentioned that “General strength is their [ESL students’] dedication. I wish all the students are as dedicated as my international students.”
Taking the mentioned factors into account, it is less surprising to see a weak, but statistically significant relationship between ESL students’ EPT performance and academic achievement.

**4.4.4.4. Academic advisors**

Academic advisors are another group of test stakeholders who may have an indirect evaluation of the relationship between English proficiency and academic achievement since they have access to students’ grades and assist students in course registration. Like most of the content course instructor interviewees, academic advisors also assumed a noticeable connection between the two and two academic advisor reported particular strategies in guiding ESL students in need of ESL instruction.

Mike of the College of Liberal Arts and Science said, “Obviously, your ability to read and understand the language is going to have a huge impact in your other academic areas. I think, for some students, the effects are universal. It [the content] is not as language-based, but its explanation is English.” The same opinion was expressed by Mary of the College of Business. She said, “it [English proficiency] really does [affect students’ performance in content courses] because for them to understand what the instructor is saying, to keep up with the readings, colleges have a lot reading.” Todd of the College of Engineering had some representative comments:

I would just reinforce the fact that their English proficiency is one of the key elements to determine their success. I know it costs a lot of money to come here. Nobody wants to take extra time. Please take some extra time to become English proficient and that will help you become successful here.

Considering the importance of English in students’ academic achievement, two academic advisors mentioned general strategies used for students to enroll in appropriate courses in terms of English requirement. For example, Mike explicitly explained how he recommended his advisees to select courses based on their English proficiency:
If we see a student taking a reading course for example, we might guide them away from general education requirement that has a lot reading at least initially in the first semester until they completed Engl99R. So, it is not only helpful with English placement classes, it is also helpful to guide students to classes that they will be more successful initially as well.

This strategy is supposedly beneficial to ESL students because it may help mitigate the impact of English proficiency on academic performance through steering the students away from certain linguistically demanding courses in the first few semesters. A similar strategic course-taking was reported by Fox (2009) who identified a strong relationship between this type of strategic decisions and ESL students’ academic performance in a university-level EAP program in Canada.

4.4.5. Summary

The questionnaire items for academic self-efficacy and motivation showed acceptable to good reliability in the Rasch analyses. However, the items for self-regulated learning strategies did not perform as expected and thus were excluded from the structural equation modeling (SEM) analysis. The SEM analysis of the relationship among the EPT results, GPA, and the motivational constructs revealed that the initial English proficiency as measured by the EPT had a direct impact on undergraduate students’ first semester GPA while it did not affect graduate students’ first semester GPA. Undergraduate students’ EPT results did not affect their academic self-efficacy whereas graduate students’ academic self-efficacy tended to be affected by their EPT results. Academic self-efficacy was found to have a strong relationship with motivational factors, which in turn, did not exert much influence on first-semester GPA. The close relationship between ESL students’ English proficiency and academic achievement was recognized by nearly all the interviewed test stakeholders (ESL students, ESL course instructors,
content course instructors, and academic advisors), despite the frequently mentioned notion that different content courses may have different demands of English proficiency.

4.5. Chapter summary

This chapter presented the analysis results in the order of research questions. The first research question regarding the relationship between the EPT and two external criteria was answered with the results in the multitrait-multimethod (MTMM) analyses. The EPT was found to have moderate relationships with the TOEFL iBT and weak to moderate relationships with the self-assessment. The EPT showed some convergent evidence as well as discriminant evidence based on the MTMM correlation coefficient matrix as well as the parameters in the correlated trait-correlated uniqueness (CTCU) model. The second research question on test stakeholders’ perception of the EPT placement decisions was answered through the qualitative analyses of the interviews with four groups of test stakeholders. The qualitative findings suggested that ESL student interviewees and ESL course instructor interviewees held a positive view of the placement decisions. ESL student interviewees showed their appreciation of the ESL courses in helping them improve English proficiency. All the test stakeholders were cognizant of the close connection between English proficiency and academic achievement at Iowa State University. The third research question about the impact of the EPT placement decisions on ESL students’ English learning was answered through analyzing the data from a pre- and posttest design, i.e. ESL course performance data collected at the beginning and the end of the course in Engl99L, Engl101B, and Engl101C classes. ESL students in Engl99L made statistically significant progress in terms of score gain on the same listening test administered at two time points. However, only nine out of 38 ESL students made satisfactory progress with reference to the
course standard. Students in Engl101B and Engl101C did not show much progress in terms of lexical complexity, syntactic complexity, and grammatical accuracy, while the Engl101C students on average wrote longer essays at the end of the course. Nonetheless, the English 101B and 101C students showed different levels of satisfactory progress in these two classes. The last research question regarding the relationship between the EPT performance and academic achievement was answered through the SEM analysis. It was found that ESL students’ EPT performance had significant and direct impact on their academic achievement. What’s more, students’ EPT performance predicted their academic self-efficacy and affected extrinsic goal orientation. However, these motivational factors did not have direct impact on academic achievement.
CHAPTER 5. CONCLUSIONS AND DISCUSSION

This chapter starts with a summary of the findings regarding each research question and discussions of these findings in light of relevant studies, then proceeds with implications for English placement tests in general and the EPT in particular, and it ends with an recognition of the limitations in this study and possible future studies on the EPT. The findings are interpreted within the framework of the validity argument for the EPT.

5.1. Summary of primary findings to the research questions

5.1.1. Research question 1: The relationship between the EPT and two external criteria of English proficiency

The first research question is “To what extent are students’ EPT performance related to their TOEFL iBT scores and self-assessment of English use?” The relationship between the EPT and two external criteria of English proficiency was investigated through an analysis of the multitrait-multimethod (MTMM) correlation matrix and an analysis of the correlated-trait and correlated uniqueness (CTCU) model. The analyses provided both convergent and discriminant evidence about the EPT. The MTMM correlation matrix indicated that the EPT had a moderate relationship with the TOEFL iBT (monotrait-heteromethod correlation coefficients: .458 to .519) and a weak to moderate relationship with ESL students’ self-assessment of English proficiency (monotrait-heteromethod correlation coefficients: .224 to .375). The magnitudes of the monotrait-heteromethod correlations between the EPT and other two measures are moderate and are comparable to other studies on English placement tests, thus lending support to the convergent evidence to the extrapolation inference in the validity argument.
The findings from the confirmatory factor analysis-based MTMM analysis matched the results of the MTMM correlation matrix analysis. Specifically, the correlated-trait and correlated uniqueness (CTCU) model fitted data in an acceptable way. The EPT measured some shared constructs as the TOEFL iBT and students’ self-assessment of English, as evidenced by the moderate to strong factor loadings on each trait or subskill. In addition, there were no noticeable method effects for the EPT, while a strong method effect of the self-assessment was evidenced with significant correlations among the uniqueness or residuals of the self-assessment sections.

5.1.2. Research question 2: Test stakeholders’ perceptions of the EPT placement decisions and their impact

The second research question is “How did the EPT test stakeholders, including ESL students, content course instructors, ESL course instructors, and academic advisors, perceive the placement decisions and their impact on students’ English learning as well as academic achievement?” Test stakeholders’ perceptions of the EPT placement decisions and the impact of the placement decisions were explored through semi-structured face-to-face interviews with eight ESL students, five ESL course instructors, four content course instructors, and three academic advisors for undergraduate students at Iowa State University. The qualitative analysis of the interviews showed that the interviewed ESL students in general understood why they were placed into ESL courses and appreciated the benefits of taking the required courses, especially ESL writing courses. Furthermore, the ESL students acknowledged the close connection between English proficiency and their academic achievement at Iowa State University. Two of the ESL student interviewees believed taking ESL courses would help improve their GPA if they could earn a good grade in the ESL course. Overall, the ESL student interviewees demonstrated positive perceptions of the EPT placement decisions.
The ESL course instructors reported that overall they were satisfied with the placement accuracy, even though occasionally they identified a few cases of potentially misplaced students in the ESL courses. Some instructors also mentioned that some ESL students tended to have some skeptical attitude towards ESL courses at the beginning of the courses partially because they may have overestimated their English proficiency relative to their placement test results or they felt that the course placement slowed down their course-taking schedule in their own majors. The ESL instructors also believed that English proficiency was important to ESL students’ academic success. The ESL instructors were mostly confident in the effectiveness of the ESL courses in preparing ESL students for their content courses and academic success. Meanwhile, two ESL course instructors commented on the limitations of the 8-week Engl99R and 99L courses and felt that these courses may be too short to help ESL students master the reading and listening skills.

The interviewed content course instructors observed a somewhat close connection between students’ English proficiency and their course performance. Their interview responses revealed that instructors believed that different courses had varied levels of requirement for English proficiency. Overall, the content course instructors wanted ESL students to improve their speaking ability and be more involved in their classes. The interviewed undergraduate academic advisors showed positive perceptions of the EPT and the placement decisions. They also reported observing that the majority of the ESL advisees were receptive to the EPT placement decisions. Considering the potential influence of English proficiency on students’ academic performance, two academic advisors mentioned that they would make course registration suggestions based on ESL students’ English proficiency. For example, they would recommend linguistically less demanding courses to the ESL students who were concurrently enrolled in ESL courses.
5.1.3. Research question 3: The impact of the EPT on English learning

The third research question is “To what extent did the ESL courses help ESL students improve their English abilities?” Specifically, ESL students’ performance in the ESL listening class (Engl99L), and two ESL writing classes (Engl101B and Engl101C) were investigated through a within subjects pre- and posttest design. In Engl99L classes, 38 students’ performance on the first in-class test were compared with their performance on the same test as the final exam at the end of the course. On average, the students in Engl99L classes improved their test scores significantly. However, only nine out of 38 students or 23.7% of the total students received a score of 65% or higher in the final exam, which was the cutoff score for course exemption used in the same in-class test at the beginning of the semester. Referencing the final exam scores with the cutoff score (65%) revealed that Engl99L students sampled in this study still had a long way to go to reach the satisfactory level even though they demonstrated improvement in listening skills with the increases in test scores.

Students’ timed essays written for an in-class test at the beginning of the 2014 fall semester and for the final exam were collected from one Engl01B section and one Engl101C section. The essays written by the same students were compared in terms of lexical complexity, syntactic complexity, grammatical accuracy, and fluency. In addition, the essays written in the final exams were rated by three experienced EPT raters using the EPT rubric to judge whether the students were ready to advance by being promoted to or being exempt from the next level of ESL courses.

The analyses showed that the Engl101B students did not make significant improvement in lexical complexity, syntactic complexity, grammatical accuracy, and fluency. Nevertheless, 14 out 18 of the final exam essays written by the Engl101B students were rated as indicating that
students were qualified for next level of ESL writing class, Engl101C. The positive evaluations from the raters suggested that at least the 14 Engl101B students made much progress in paragraph-level writing.

The students in the Engl101C class showed significant progress in lexical density development as well as in three syntactic complexity features. The timed essays written by Engl101C students in the final exam were significantly longer than the timed essays written at the beginning of the class, indicating an improved fluency in writing in timed essay writing tasks. However, there was no significant improvement in terms of grammatical accuracy. Only eight out of 16 of the final exam essays were rated as qualified or ready for next level, Engl150 (the first-year composition course at Iowa State University), indicating that the majority of the sampled Engl101C students may still need some help in order to succeed in Engl150.

5.1.4. Research question 4: The impact of the EPT on academic achievement

The last research question is “to what extent was students’ English proficiency as measured with the EPT related to their academic achievement in light of individual difference in motivational constructs and self-regulated learning strategies?” The data used in this research question included ESL students’ EPT performance data, first-semester GPAs as well as their responses to the comprehensive online questionnaire about motivation, academic self-efficacy, and self-regulated learning strategies. The items of self-regulated learning strategies were not included in the structural equation modeling analysis because it showed relative low reliability and its measurement model did not fit the data. The final structural equation modeling analysis of the relationship among ESL students’ English proficiency, first-semester GPA, and motivational factors revealed several findings. First, the EPT performance was a significant predictor of ESL students’ first semester GPA and its standardized coefficient was .306. In other words, an
increase of one standard deviation unit in ESL students’ EPT performance (factor score of the EPT) can predict an increase of 0.306 of a standard deviation on students’ first-semester GPA on a 4-point scale. Meanwhile, ESL students’ EPT performance also affected their academic self-efficacy positively, which in turn influenced students’ extrinsic and intrinsic motivation as well as their extrinsic motivation negatively. This pattern suggests that the ESL students with higher English proficiency as measured with the EPT tended to be more confident in their class performance and they also tended to be less performance-oriented or worry less about their grades in classes. This confidence level would also affect ESL students’ learning motivation, especially intrinsic motivation. However, the motivational factors did not have a significant impact on students’ first-semester GPA and the motivational factors did not mediate the relationship between ESL students’ EPT performance and their academic achievement. This finding may be explained by the complicated learning process involved for university students (Ota, 2013; Young, Sercombe, Sachdev, Naeb, & Schartner, 2013). Overall, the SEM analysis results demonstrated the positive impact of ESL students’ EPT performance on their academic achievement and academic self-efficacy as well as its negative impact on their extrinsic motivation.

5.2. The validity argument for the EPT

The validity argument for the EPT can be developed with the summary of primary findings from the research questions, which were developed to investigate the extent to which backing could be found for assumptions in the interpretation and use argument described in Chapter 2. In this section, I discuss the extent to which results can serve as backing for the respective assumptions underlying the warrants for both extrapolation inference and ramification
inference. A summary table of the validity argument for the EPT concerning the extrapolation inference and ramification inference is illustrated at the end of this section.

5.2.1. The extrapolation inference

The answers to the first research question provided partial support for the two assumptions underlying the warrant for the extrapolation inference. The warrant is that the scores on the EPT reflect learners’ actual English proficiency in academic contexts at Iowa State University. The first assumption for this warrant is that the constructs of academic English assessed in the EPT account for students’ scores on the TOEFL iBT. This assumption was partially supported with the moderate correlations between the EPT and the TOEFL iBT, which were investigated through correlational analyses of the MTMM data and a CFA-based correlated trait-correlated uniqueness (CTCU) model. The correlation coefficients between the two measures ranged from .458 to .519, and the disattenuated correlation coefficients ranged from .604 to .684 after controlling for the effect of reliability. Given the possible similarity in operationalization of academic English as the constructs as well as the obvious differences in intended uses between the EPT and the TOEFL iBT, the correlation coefficients between the EPT and the TOEFL iBT were within the expected range as shown in other studies on English placement tests. Actually, the correlation coefficients reported in this study were slightly higher than those in other similar studies on the relationship between the TOEFL iBT and a local English placement test. For example, Manganello (2011) reported that the correlation coefficients between the EPT used at Iowa State University and the TOEFL iBT ranged from .317 to .433 for the EPT administrations from fall 2009 to spring 2011. In Fulcher (1997)’s study on an English placement test used in the English Language Institute at the University of Surrey, UK, 33 students took both the old version of TOEFL and the placement test and the correlations
between the two tests varied from .34 in writing to .63 in reading, and the correlation between the total scores was .64. It is noteworthy that Fulcher’s study was conducted in a different academic context in the UK and the TOEFL in his study was the paper-based version. Kokhan (2012) found that the correlations between the English placement test used at the University of Illinois at Urbana-Champaign and the TOEFL iBT varied and the time gap between these two tests affected the correlations. In Kokhan (2012), the shorter the time gap was, the higher correlations were found between the two tests. However, the highest correlations between the two tests were still below .4, even with the minimum time gap. In this dissertation study, I did not control for the time gap of the TOEFL test to the EPT as variable in the correlation analyses of these two tests because of some data preparation issues. This un-differentiating use of the TOEFL iBT scores may have some impact on the relationship between these two tests.

The second assumption for the warrant of the extrapolation inference is that the constructs of academic English assessed in the EPT account for students’ self-assessment of English use. This assumption was only partially supported with weak to moderation correlations between the EPT and the self-assessment. I hypothesized low to moderation correlations between the two measures with the same considerations about construct coverage or representativeness discussed above applied to the relationship between the EPT and the self-assessment. That is, these two tests may have measured something in common as well as something different. Another consideration in interpreting the relationship between the EPT and the self-assessment is about the nature of self-reporting data, which have been found to be affected by a number of factors such as personality, previous experiences, and so on. (Zell & Krizan, 2014). Therefore, I interpret the observed correlation coefficients between these measures (Reading: .224, Listening: .317, Writing: .375) and their disattenuated counterparts (Reading: .288, Listening: .411,
Writing: .440) as in line with other similar studies and comparable to the magnitudes of the relationship between self-assessment of English proficiency and other measures (Ross, 1998). For example, Enright et al. (2008) observed moderate correlation coefficients between a self-assessment and a prototype measure of the new TOEFL (.45 to .55). In a large scale study on the speaking and writing sections of the Test of English for International Communication (TOEIC) test, Powers et al. (2009) reported the correlations between the TOEIC and test-takers’ self-assessment to be .54 and .52, respectively. The correlations between individual items and the TOEIC scores ranged from .30 to .49. In a validation study of the TOEFL iBT writing scores, Weigle (2010) reported an example of moderate correlations between 386 test-takers’ self-assessment of English proficiency and their TOEFL iBT writing scores awarded by two human raters (.33 to .43).

The CFA-based MTMM data analyses offered supplementary information about the relationship among the three measures with model fit indices of the CTCU model and parameter estimates. The CTCU model fit the data well. In addition, the significant factor loadings from the three tests on the same sets of traits, namely reading, listening, speaking, and writing subskills, indicated that the three tests measured some shared constructs. Overall, the backing for the assumptions was positive to establish the extrapolation inference, that is, the EPT scores, to some extent, reflected ESL students’ actual English proficiency in academic contexts at Iowa State University. This supported claim can be used as the ground for the next inference, ramification, in the validity argument.

5.2.2. The ramification inference

The ramification inference is about the impact of test score use in a specific context. Three assumptions were made for the ramification inference. The first one was that the decisions
of ESL course placement are justifiable and comprehensible to test stakeholders. This assumption was supported by empirical backing through face-to-face semi-structured interviews with the test stakeholders as exemplified in the answers to research question 2. Five out of eight interviewed ESL students showed that they understood why they were placed in the ESL courses despite their initial strong reactions to the test results. Three other student interviewees did not make direct comments on their perception of the placement decisions. The interviewed academic advisors seemed to be familiar with the placement decisions and were positive about the decisions based on their experiences with their international advisees. That is, their communication with international advisees helped confirm the EPT placement decisions as the students who were placed in ESL courses tended to struggle in their interactions with the academic advisors. The interviewed ESL instructors believed that the placement decisions in general were reasonable; even there were few cases that seemed to be misplaced.

The second assumption was that the decisions are beneficial for learners’ improvement of academic English proficiency. Since the direct consequences of the use of EPT scores are placement of ESL courses, ESL students’ progress in the placed courses would constitute evidence to support this assumption. This support is based on the answers to research question 3. The course performance data from the Engl99L classes indicated that Engl99L students in general made statistically significant gains in the test scores on the same test administered at the beginning of the course and the end of the course. However, the final achievement in Engl99L varied considerably and only nine of 34 students (26.5%) met the cutoff score set for course waivers in Engl99L on the final exam, which was considered as the posttest. The sampled students in Engl101B and Engl101C also showed a different hypothetical passing rate when the same EPT standards were used to evaluate students’ essays in the final exam. The hypothetical
passing rate in the sampled Engl101C was 50%. This problem was noted by the ESL instructors in the interview, who attributed this to the influence of a low level of learning motivation in some classes. Therefore, this study did not find adequate support for this assumption.

The third assumption was that EPT performances are predictive of ESL learners’ academic achievement at the university. The backing for the third assumption about the predictive power of the EPT performance was supported with the answers to research question 4. The structural equation modeling analysis showed that the EPT performance had a direct and significant impact on ESL students’ first semester cumulative GPA (standardized regression coefficient = .306), meaning that ESL students with higher EPT performance tended to achieve a higher GPAs in the first semester. Previous studies on the factors contributing to students’ GPAs have shown that there are a large number of variables that can influence ESL students’ academic performance, for example, language proficiency (Stoynoff, 1997; Vinke & Jochems, 1993), motivational factors (Ning & Downing, 2012; Phakiti, Hirsch, & Woodrow, 2013; Richardson et al., 2012), learning strategies or academic preparation (Mathews, 2007), social resources (Ota, 2013; Young et al., 2013). This study only included a small number of the potential predictors of academic achievement. The reported relationship between the EPT and students’ academic achievement was fairly strong given the limited predictors involved in this study.
Table 5.1
Summary of the Validity Argument for the EPT

<table>
<thead>
<tr>
<th>Infer -ence</th>
<th>Warrant</th>
<th>Assumptions</th>
<th>Backing evidence</th>
<th>Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>The scores on the EPT (expected scores) reflect learners’ actual English proficiency in academic contexts at the university (target scores).</td>
<td>1. The constructs of academic language proficiency as assessed by the EPT account for students’ scores on the TOEFL iBT.</td>
<td>Moderate correlations were found between the EPT and the TOEFL iBT; Acceptable model fit and significant factor loadings were found in the correlated trait-correlated uniqueness (CTCU) model.</td>
<td>Assumption partially supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. The constructs of academic language proficiency as assessed by the EPT account for students’ self-assessment of English proficiency.</td>
<td>Weak to moderate correlations were found between the EPT and the self-assessment of English proficiency; Acceptable model fit and significant factor loadings were found in the correlated trait-correlated uniqueness (CTCU) model.</td>
<td>Assumption partially supported</td>
</tr>
<tr>
<td>Ramification</td>
<td>The decisions made based on the EPT scores are conducive to ESL students’ English learning and academic success at the university.</td>
<td>1. The decisions of ESL course placement are justifiable and comprehensible to test stakeholders.</td>
<td>ESL student interviewees understood the placement decisions; Positive perceptions of the EPT placement decisions reported by the majority of the interviewed test stakeholders.</td>
<td>Assumption supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. The decisions are beneficial for ESL learners’ improvement of academic English proficiency.</td>
<td>Significant score gains were found between two administrations of the same test in Engl99L; 77.8% of the Engl101B students and 50% of the Engl101C students in the sampled sections made satisfactory progress in the final exam essays in terms of holistic rating, however, not in terms of syntactic complexity, lexical complexity, and accuracy.</td>
<td>Assumption partially supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. The decisions are beneficial for ESL learners’ academic achievement at the university.</td>
<td>The EPT performance was a significant predictor of ESL students’ first semester GPA.</td>
<td>Assumption supported</td>
</tr>
</tbody>
</table>
5.3. Implications for English placement tests in general and the EPT in particular

The implications of the findings are discussed in terms of theoretical, methodological, and practical implications for the future development of the EPT.

5.3.1. Theoretical and methodological implications

This findings in this study can contribute to a better understanding of the score interpretation and use of English placement tests in general in the U.S. higher education as this dissertation project exemplified the use of an argument-based approach to validating the score interpretation and use in a local educational context. While the advantages of the argument-based approach have been demonstrated and discussed in Kane (2006) and Chapelle et al. (2008), this relatively new way to view and conduct validation studies just started gaining popularity in the second decade of the 21st century (Chapelle & Voss, 2013). This study is among the few attempts at applying this approach to studying an English placement test. Even fewer studies have specifically focused on the impact of English placement tests. Therefore, the interpretation and use argument for the EPT in this study as well as the assumptions underlying the major inferences can be relevant and informative to other institutions that use in-house English placement tests.

In addition, the mixed-methods methodology used to investigate the impact of the EPT can be employed in other validations. For example, one of the challenges in establishing the extrapolation inference is to identify and/or construct a proper external criterion representing the language use in the target domain. In this study, self-assessment was used as such a criterion. The psychometric property of this instrument was investigated using the Rasch model in conjunction with confirmation factor analysis and the psychometric properties of this self-assessment tool was found to be robust in these analyses. The methods used in the process of
scale development and validation can be used in other studies with similar validation goals. Since the main concern in the extrapolation inference is to explore the relationship among different measures of a set of similar constructs, the combined use of the MTMM correlation matrix analysis and CTCU model analysis in this study addressed the concern appropriately and provided important information regarding the relationship among different measures. Another important methodological feature of this study is the use of qualitative analysis of interviews with test stakeholders for two purposes: to answer research question 2 which is about test stakeholders’ perceptions of the EPT placement decisions as well as their impact and to help interpret the quantitative findings in other research questions concerning the impact of the EPT placement decisions. For example, ESL students’ interview responses shed light on the relationship among the EPT and the external criteria (the TOEFL iBT and the self-assessment) and also partially explained the impact of the placement decisions on ESL students’ English learning. In addition, making different voices from test stakeholders heard has been very important to test development and validation (Mathew, 2004).

5.3.2. Practical implications for future development of the EPT

Three main implications for the future development of the EPT can be made based on the findings in this study. Firstly, the task formats in the EPT should be revisited with reference to tasks described in the self-assessment and revised, if necessary, in the future to better represent the constructs of academic English as well as to improve test reliability. As expected, the weak to moderate correlations between the EPT and the self-assessment suggest that these two measures tapped somewhat different constructs. The self-assessment was developed to elicit students’ responses to a series of statements about their English use in specific academic activities, whereas the EPT was designed to assess students’ academic English proficiency using essay
writing tasks and traditional multiple-choice questions attached to reading and listening materials. In other words, the correlations demonstrated that ESL students’ feeling or self-confidence in using English in real-life tasks deviated from their EPT performance to some extent. Ideally, ESL students’ self-assessment, if carried out in a bias-free manner, could be treated as a reflection of their actual performance in academic contexts, that is, the target scores. However, as described in Chapter 2, there are a number of factors that can affect the accuracy of self-assessment. Nevertheless, the self-assessment items developed in this study contained the can-do statements about the typical tasks that require using English in academic contexts. If some of these tasks could be simulated in the EPT, the task types in the EPT would be more aligned with the typical academic activities on campus, as suggested by two ESL instructors. In that case, the constructs of academic English could be better represented and the future EPT may enjoy a higher correlation with students’ self-assessment of English in the target domain.

The relatively low reliabilities in the current version of the EPT is worth mentioning in this study. The magnitude of test score reliability (from .67 to .78) indicate that the amount of measurement error in that particular administration of the EPT was noticeable large (Chapelle, 2013). There are a host of factors that could affect test score reliability, for example, distribution of test difficulty, test-takers’ familiarity with the test format, testing environment, test-taker fatigue, motivation, and so on. In addition, the level of score reliability affects the correlation coefficients between the test and other measures, for example, in the multitrait-multimethod analyses, because of the involvement of measurement errors (Bachman, 2004). The EPT office should make efforts to maximize score reliability in the process of test development and administration. One possible action to take is to add more test tasks, which will help improve reliability with more samples elicited from the test-takers. However, the current version of the
EPT lasted about three hours, it may not be very feasible to add more tasks to the test. Other alternative actions include 1) improving item psychometric quality, 2) ensuring consistently comfortable testing environment for test-takers, especially better audio quality in the listening section, 3) monitoring essay rating performance and providing extra training or calibration, when necessary. Of course, new task formats could be designed with a consideration of improving test reliability as well.

Another relevant issue is about the components of the EPT. The interviews with both content course instructors and ESL course instructors highlighted a need for a speaking component in the EPT. The interviewed content course instructors noticed that international students typically felt less comfortable representing ideas and engaging in-class discussions. For example, Feng-Chun of the Mathematics Department commented, “If they (ESL students) can strengthen their communicative skills, that will help them exchange their ideas (in the class).” Feng-Chun also expressed concerns about students’ pragmatic competence in daily communication by sharing examples of seemingly rude conversations with ESL students. She said, “Maybe some students did not know how to use language properly. Some instructors may feel offended.” A speaking test, especially with group oral discussion tasks, would be a valuable addition to the current version of the EPT.

Secondly, effective communications with test stakeholders about the EPT itself and the intended impact of the EPT should be promoted. As mentioned by two interviewed ESL course instructors, some ESL students were less motivated in ESL courses because they believed that they would not need extra ESL assistance. One possible cause is that these students overestimated their English proficiency. In addition, at least one student interviewee said that she was not prepared for the EPT because she knew little about it. This may be an issue for the
students who may experience high test anxiety because of an unfamiliarity with the test. Currently, the EPT website is the only resource containing limited information about the test structure, task format, and other aspects. Therefore, more detailed information about the test structure and task format should be prepared for ESL students before their arrival on campus and during the orientation. If possible, practice materials should be provided to the ESL students to familiarize them with the test. These practice materials could help raise students’ awareness of the challenges in English use on campus and at the same time lower their test anxiety.

In line with this concern is the transparency of placement decision-making and score reporting. The interviews with ESL students revealed the frustrations experienced by some ESL students, who claimed that they did not know their grades or scores except for the final decisions. For example, Peng-Long recalled, “I just looked at the blackboard and it said you did not pass that part and I need to take that course. Because I did not find my score on the website. I just get a level or something.” This limited access to score information may affect students’ attitude to the ESL courses as well as learning motivation in these courses when they could not find a close connection between their need and the content of the courses. This practice calls for more attention in the argument-based approach to validation because more inferences should be investigated before using raw scores for decision-making. In the framework of argument-based validity, evidence is first needed to show that the raw scores are reliable enough (expected score in the generalization inference) and then they should be demonstrated to adequately represent the constructs being measured (construct in the explanation inference) before they can be used for decision-making. These practices could then be communicated to stakeholders through documentation of the EPT validation studies to better communicate with test stakeholders.
Thirdly, the test administration procedures could be improved by addressing some of the concerns over the testing environment, especially the issues regarding the listening tests administered in big auditoriums. One potential solution to this issue is to fully computerize the EPT with new test formats and tasks in the near future and the multi-media tools on computers can help solve this issue. In addition, using new technology can bring about more benefits in the future development of the EPT, such as allowing for a simulation of a real-class scenarios for testing listening and reading, enabling paired or group online oral discussions and creating opportunities to use spelling checker in essay writing. Currently, there is a computer-based version of the EPT on Blackboard Learn that has the same task types as the paper-based EPT. This computer-based version is mainly used for graduate students in distance education programs.

5.4. Limitations and future studies

Limitations of this research should be noted to inform future studies on the EPT. Firstly, the ESL students who participated in this study were mainly the students who took the EPT and were placed into one or more ESL courses. Three groups of ESL students were not fully represented in the study sample, namely the students who were exempt from taking the EPT, the students who passed the EPT, and the students who were waived from taking the ESL courses by excelling on the in-class tests. Given the fact that these three types of students were not affected by the EPT placement decisions in the same way as those who were required to take the ESL courses, the EPT testing experience and academic performance of these three groups of students may be different to some extent from the sampled ESL students. Therefore, the representativeness of the ESL students was limited in a very specific way with such a truncated
sample. Future studies should consider enlarging the sample groups to include more participants from the mentioned three groups of ESL students to study their academic performances after the EPT decisions. In addition, a related topic that is worth investigating is ESL student’s further development of English proficiency along their college journey as well as the longitudinal relationship between their English proficiency and academic performance in their later years at Iowa State University. With this type of longitudinal data, a more comprehensive picture about the impact of the EPT placement decisions could be made. Another sample-related issue is that I treated the whole group of participants as a homogeneous group in my data analysis. Separating the data by other features would decrease the sample size drastically and thus limit the statistical power of the analyses. Finer distinctions among the participants based on their demographic information such as gender, education status, first languages, or disciplines may help yield more informative findings regarding the relationship between the EPT and the external criteria as well as the impact of the EPT placement decisions or performance on students’ English language learning and academic achievement.

In this study, I relied on the questionnaire as a main instrument for both research questions 1 (the relationship between the EPT and two external criteria – the TOEFL iBT and the self-assessment) and 4 (the relationship between the EPT performance and first-semester GPA). It is noteworthy that the quality of self-reported data, such as the responses to statements on Likert scales, could be affected by a number of factors such as method effects, respondents’ social desirability, personality of respondents, and respondent fatigue or boredom (Schwarz, 1999). For the relationship between the EPT and self-assessment, future efforts could be directed to collecting other types of real-life performance data as suggested by Brooks and Swain (2014). For example, ESL students’ speech samples and written samples from content courses could be
analyzed as external criteria of English proficiency. With regard to research question 4, the motivational factors did not perform as expected, especially the items on self-regulated learning strategies. The items on learning motivation and the self-regulated learning strategies were adapted from the MLSQ, which was originally designed for native-English-speaking students in the U.S. This could partially explain why the proposed factor structure of the self-regulated learning strategies items did not hold in this study. Furthermore, respondent fatigue and/or boredom could have contributed to the poor model fit for the data. Future studies could pay more attention to the questionnaire items on the self-regulated learning strategies. For example, wording and readability of the items could be checked with targeting participants, and an investigation of the ease of completion of the questionnaire as a whole used by ESL students should be conducted. Following these steps, the relationship between a revised or modified set of self-regulated learning strategies and ESL students’ academic achievement should be included in the model proposed in this study.

The relationship between the EPT performance and first-semester academic achievement was investigated using the latent variable approach. In other words, English proficiency was operationalized as the factor score of the three components: reading, listening, and writing. Recent studies have shown that ESL students tend to have different skill profiles in English proficiency, which in turn affects academic performance. For example, in a study on the relationship between TOEFL and GPA for Chinese students in an American university, Ginther, Yan, and Potts (2015) identified different correlational patterns between TOEFL and GPA, for example, strong and positive correlations between TOEFL speaking and writing scores and the GPAs, strong and negative correlations between TOEFL reading and listening and the GPAs. Ginther et al. (2015) attributed this correlation pattern to the characteristics or score profiles of a
particular group of students who had high scores on reading and listening, but low scores on speaking and writing. In light of these findings, future studies on the EPT can take a closer look at the influence of different skills profiles on students’ GPA.

5.5. Concluding comments

Findings in this study will help test educators and researchers understand ESL students’ performance on the EPT with a reference to their performance on the TOEFL iBT and self-assessment. In addition, this study is particularly meaningful in that it revealed some positive impact of the placement decisions on ESL students’ English learning and academic achievement. Validation of test score interpretation and test score use is an ongoing process and this dissertation project contributed the validity evidence needed for the two inferences in the validity argument. More systematic efforts should be made within the framework of argument-based validity to construct a more comprehensive and coherent validity argument for the EPT.

The specific findings in this study are not intended to be generalizable to other educational contexts. Nevertheless, other institutions may face similar issues identified in this study if local placement tests are used. In that case, the implications of the findings could be applicable to these institutions. More importantly, the interpretation and use argument for the EPT in this study could be useful roadmap for other institutions to conduct and design their own validation studies on local English placement tests. The same or similar set of assumptions underlying the inferences could be utilized to construct research questions. The mixed-methods approach and the specific research methods employed in this study can be used by other institutions to collect and analyze their own data.
REFERENCES


Bretag, T. (2007). The emperor’s new clothes: Yes, there is a link between English language competence and academic standards. *People and Place, 15*(1), 13–21.


APPENDIX A. SURVEY ABOUT SELF-ASSESSMENT OF ENGLISH USE AND LEARNING MOTIVATION

Dear Student,

You are invited to participate in a research on English as a second language (ESL) learner's self-assessment of English use and learning motivation. This study is being conducted by Zhi Li, a PhD student of Applied Linguistics and Technology, under the supervision of Dr. Carol Chapelle of the Department of English at Iowa State University. You will be asked to respond to several statements on a 6-point scale (Strongly Disagree to Strongly Agree). It will take 5-8 minutes to complete this questionnaire. Your responses will help us improve ESL courses and English Placement Test (EPT) in the future.

To thank you for your time, you have a chance to win one of the five $10 gift cards.

Purpose: The purpose of this study is to investigate how ESL (English as a second language) students self assess their use of English in non-ESL courses or major courses, as well as their general learning motivation.

Participants: In order to be eligible to participate, you must be a non-native English-speaking or ESL students and be 18 years or older. Procedure: You will be asked to respond to several statements regarding your use of English and learning motivation on a 6-point scale (Strongly Disagree to Strongly Agree).

Risk and benefits: There are no physical risks associated with this study. Voluntary participation: Your participation in this study is completely voluntary, and you may choose to withdraw at any time. Withdrawal from this study will not result in any type of penalty.

Confidentiality: All information or your responses will be kept completely confidential. A password protected computer will be used for data analysis. This project has been reviewed and approved by the Institutional Review Board for the Protection of Human Participants at Iowa State University.

If you have any question, you are encouraged to ask at any time during this study. For further information about the study, please contact Zhi Li at zhili@iastate.edu or Dr. Carol Chapelle at carolc@iastate.edu. If you have any questions about the rights of research subjects, please contact the IRB Administrator (IRB@iastate.edu) or Director, (515) 294-3115, Office for Responsible Research, Iowa State University, Ames, Iowa 50011.

If you agree to participate in the study, please print out a copy of this email or page (informed consent) for your own file and start the survey.

Your time and efforts are deeply appreciated.
Sincerely,

Zhi Li
Ph.D. student of Applied Linguistics and Technology
Department of English
226 Ross Hall
Iowa State University
Ames, Iowa, 50011

If you agree to participate in this study and would like to participate in a drawing of five $10 gift cards, please enter your iastate email address in the box below to proceed. Winners of the gift card will be notified by email on or before November 30, 2014.

Self-assessment of English use in your courses 1 (Page 1 out of 5)
Below are some self-assessment statements about English use in and outside your courses. Please rate the following statements based on your own experience. (1 Strongly DISAGREE - 6 Strongly AGREE)
Please rate the following statements about your use of English in your courses based on your own learning experiences at Iowa State University. (Strongly DISAGREE -- Strongly AGREE)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. I can fully understand a lecture, presentation and/or demonstration in English in and outside my classes.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.2. I can make useful notes during a lecture or presentation in my classes.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.3. I can follow what my American classmates/friends are saying in class discussions even when they speak fast.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.4. I can fully understand the English conversations, talks, and reports in films, TV shows, and the radio in and outside my classes.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.5 I can understand jokes and cultural points in conversations with my American classmates/friends in and outside my classes.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.6. I can read and understand complex opinions and implicit or indirect meaning in my textbooks or other readings required in my classes.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.7. I can read to understand the main ideas from complex reports, analyses, and textbook chapters.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.8. I can read quickly to find important details in my textbook, handouts, and technical manuals.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.9. I can read and understand the syllabus, assignment sheets, and equipment/software instructions without any help.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1.10. I can successfully guess the meaning of unfamiliar words or idioms (fixed phrases) in my textbooks and other readings without using a dictionary.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Q1.11. I can write an effective summary with the information from different reading materials in and outside my classes.</td>
<td>Strongly Disagree</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>Strongly Agree</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Strongly Agree</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

1.12. I can write clear and well-organized formal letters or emails (e.g. applications, request, complaints, etc.) in a polite way.

1.13. I can write complex and well-developed (research) articles, reports or essays with good structure and accurate English in my classes.

1.14. I can edit my written work and correct grammar errors in it.

1.15. I can revise my writing and make it better with appropriate vocabulary and expressions in my classes.

1.16. I can give a clear and well-structured presentation in English on a familiar or prepared topic in my classes.

1.17. I can keep up a discussion in English with my American classmates on a wide range of topics in and outside my classes.

1.18. I can talk comfortably in English with university staff (e.g. secretary, librarian) and academic advisers when I need help.

1.19. I can orally summarize my readings or class notes using fluent and correct English in and outside my classes.

1.20. I can express myself fluently in English to my American classmates and instructors without much trouble with vocabulary and expressions in and outside my classes.

General beliefs about academic learning and Learning motivation in major courses (non-English courses) (Page 3 out of 5)

Q2.2. Please rate the following statements about how you see yourself in your major courses or non-English courses. (Strongly DISAGREE -- Strongly AGREE)
<table>
<thead>
<tr>
<th>Q3 3. Please rate the following statements based on your learning motivation in your major courses (non-English courses).</th>
<th>Strongly Disagree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. In my major course(s), I prefer course material that really challenges me so I can learn new things.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.2. I work hard in my major course(s) so that I can improve my GPA with a better grade.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.3. The most satisfying thing for me in my major course(s) is trying to understand the content as thoroughly as possible.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.4. When I have the opportunity in my major course(s), I choose course assignments that I can learn from even if I may not be able to receive a good grade.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.5. Getting a good grade in my major course(s) is the most satisfying thing for me right now.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.6. In my major course(s), I prefer course material that arouses my curiosity or interests me, even if it is difficult to learn.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.7 If I can, I want to get better grades than most of the other students in my major course(s).</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3.8. I want to do well in my major course(s) because it is important to show my ability to my family, friends, adviser, or others.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
### Learning strategies in major courses (Page 4 out of 5)

Q4 4. Below are some statements about the general strategies you may use in your major courses. Please rate the following statements based on your learning experiences. (Strongly DISAGREE -- Strongly AGREE)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. I work hard to do well in my major course(s) even if I don’t like what we are doing.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.2. When I can’t understand the material in my major course(s), I ask other students or friends for help.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.3. When I study for my major course(s), I often quit before I finish what I planned to do because I feel bored or lazy.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.4. I make sure I keep up with the weekly readings and assignments for my major course(s).</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.5. I often find that I don’t have much time for my major course(s) because of other activities.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.6. I ask my teacher during class or privately to clarify or explain more about the concepts I don’t understand well.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.7. Even if I have trouble learning the material for my major course(s), I try to do the work on my own, without help from anyone.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.8. I have a regular time set aside for studying for my major course(s).</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.9. When the assignments in my major course(s) are difficult, I give up or only study the easy parts.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4.10. I often ask myself whether I am reaching my goals in my major course(s).</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

### Something about you (Last Page)

Q6-1 1. What is your gender?
- ○ Female
- ○ Male
- ○ Prefer not to say

Q6-2 2. What is your (major) native language?
Q6-3 3. What is your education status?
- Undergraduate student
- Graduate student
- Other ____________________

Q6-4 4. Which of the following age group do you belong to?
- 18-22
- 23-30
- 31-35
- over 35

Q6-5 5. Which college are you in at Iowa State University?
- Engineering
- Business
- Liberal Arts and Science
- Agriculture and Life Sciences
- Human Science
- Design
- Veterinary Medicine
- Undecided
- Other ____________________
Q6-6 6. How long have you been studying in the United States and other English-speaking countries, such as Australia, Canada, etc.?
- less than 3 months (1)
- 4-12 months
- 1-2 years
- 2-3 years (3)
- more than 3 years (4)
- Other (5) ________________

Q6-7 7. Which of the following English course(s) are you taking in this semester? (Choose all that apply)
- Engl 99R
- Engl 99L
- Engl 101B
- Engl 101C
- Engl 101D
- Engl 150
- Engl 250
- IEOP courses (Intensive English and Orientation Program)
- Engl 180A/B/C/D/E
- None
- Other ________________

Q6-8 8. How many courses are you taking in this semester at Iowa State University?
- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7 or more
APPENDIX B. SAMPLE INFORMED CONSENT FORM

INFORMED CONSENT DOCUMENT
Title of Study: A validation study of the English Placement Test (EPT)
Investigators: Zhi Li, Volker Hegelheimer, Carol Chapelle

This is a research study. Please take your time in deciding if you would like to participate. Please feel free to ask questions at any time.

INTRODUCTION
The purpose of this study is to examine the validity of the English Placement Test (EPT) for non-native English-speaking students by collecting opinions about the EPT and the English course placement decisions from various test stakeholders, including test-takers, academic advisors, and instructors. This study is approved by the IRB office (No. 14-010). You are being invited to participate in this study because you work as an academic advisor at one of the colleges at Iowa State University.
You should not participate if you are under 18.

DESCRIPTION OF PROCEDURES
If you agree to participate, you will be asked to take a short face-to-face interview about your experience in advising international students in course registration and dealing with international students’ questions about the English Placement Test for non-native English-speaking students at Iowa State University. The interview will be audio recorded with your permission. Your participation will last for about 10-30 minutes.

RISKS
There are no foreseeable risks at this time from participating in this study.

BENEFITS
If you decide to participate in this study there may be no direct benefit to you. It is hoped that the information gained in this study will benefit your college and the university by offering better service to international students with more accurate placement decision regarding the English as a second language courses.

COSTS AND COMPENSATION
You will not have any costs from participating in this study. You will not be compensated for participating in this study.

PARTICIPANT RIGHTS
Your participation in this study is completely voluntary and you may refuse to participate or leave the study at any time. If you decide to not participate in the study or leave the study early, it will not result in any penalty or loss of benefits to which you are otherwise entitled. You can skip any questions that you do not wish to answer.

CONFIDENTIALITY
Records identifying participants will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, federal government regulatory agencies, auditing departments of Iowa State University, and the Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy your records for quality assurance and data analysis. These records may contain private information.

To ensure confidentiality to the extent permitted by law, the following measures will be taken: your personal identifiers will be removed and pseudonyms will be. If the results are published, your identity will remain confidential.

QUESTIONS OR PROBLEMS
You are encouraged to ask questions at any time during this study. For further information about the study contact Zhi Li (515-294-6398) and Dr. Volker Hegelheimer (515-294-2282).

If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115, Office for Responsible Research, Iowa State University, Ames, Iowa 50011.

PARTICIPANT SIGNATURE
Your signature indicates that you voluntarily agree to participate in this study, that the study has been explained to you, that you have been given the time to read the document, and that your questions have been satisfactorily answered. You will receive a copy of the written informed consent prior to your participation in the study.

Participant’s Name (printed) __________________________________________________________

(Participant’s Signature) ____________________________ (Date)
APPENDIX C. INTERVIEW QUESTIONS FOR ESL STUDENTS

1. Is this your first semester at ISU? Overall, how was your (first) semester at Iowa State University?

2. What ESL course(s) are you taking this semester? What is your major? How many other courses (non-English courses) are you taking at the same time? What are they (one or two course names will be enough here)?

3. Let’s talk about the English Placement Test (EPT) that you took at the beginning of this semester.
   a. What was your impression of the EPT (in terms of item format, difficulty, etc.)? How well did you do in the EPT?
   b. Do you understand/know why you were placed into the ESL class(es)? [In other words, were you aware of the standards/criterion used for placement in the EPT?]
      i. If you were not sure why you were placed into the classes, did you get some help to know why?
   c. Did you feel you had been placed in the right ESL class(es)? Why or why not?

4. Did you take the in-class test on the first day of your ESL class(es)? (If yes, ask the probing questions. If no, why not?)
   a. What was your impression of the in-class test? Was it a good chance for you to test out of the class?
   b. How would you compare your performances in these two tests (the EPT and the in-class tests)?
   c. Do you remember your TOEFL or IELTS score (a rough range is fine)?
      i. How would you compare your performances in these two tests (the EPT and the in-class tests) and the TOEFL/IELTS test?
      ii. Were these performances similar, better or worse than you expected at the time of these tests?
      iii. Do you think your TOEFL/IELTS score was good enough for you to manage or deal with content courses at ISU?

5. Think about your first few weeks at ISU. Did you feel comfortable using English on campus or in your content courses
   a. Did your content/major courses have a strong requirement of English ability/proficiency? If yes, in what way? If no, why not?
   b. Were you ready for the courses in terms of English requirement at the beginning of the semester?
c. At the beginning of the semester, what were your strengths and weaknesses in English use at ISU (example: grammar, vocabulary, reading for main ideas, writing report, etc.)? (Or, at the beginning of the semester, what kind of challenges did you face in terms of reading /listening /speaking /writing English in your content courses)?

d. Did your English ability/proficiency affect your performance in non-ESL (content) courses, for example, comprehending lectures, taking note, joining in-class discussion, completing written assignments, and etc.? If yes, in what way?

6. It is the end of the semester, how would you evaluate your English? What are your strengths and weaknesses in English at this stage?

a. Do you still have the same challenges after taking the ESL course(s)? If yes, what are they?

b. Since this is the end of your first semester, do you feel the ESL courses (99R, L, 101B/C/D) have prepared you for your study at ISU in terms of English use, for example, Engl150 and other content courses in the following semesters? (Or, do you feel confident in your English at ISU?)

   i. If yes, what helped you to improve English? If no, what kind of help would you need?

   ii. What are the main (reading/listening/writing) skills that you learned in the ESL classes?

   iii. What kinds of (reading/listening/writing/speaking) skills that you learned from the ESL class(es) were (particularly) useful in your content courses?

7. Do you have some suggestions for the EPT to make it a better test? Do you have some suggestions for your fellow international students about how to improve English proficiency and achieve academic success at ISU?
APPENDIX D. INTERVIEW QUESTIONS FOR ESL COURSE INSTRUCTORS

1. Can you tell me something about yourself and your course?
   a. What is your educational background?
   b. What is your professional background? Or, how long have you been working as an ESL instructor (before coming to ISU as well as at ISU)?
   c. What kinds of ESL and English courses have you taught at ISU?
   d. In your ESL courses, what is your teaching style? (Or, how did you teaching these courses?)
   e. In your ESL courses, what are the typical course activities and assignments? (Refer to specific courses)

2. Can you describe your experiences with the English Placement Test at ISU (proctoring, supervising, essay rating, item writing, etc.)?
   a. What is your impression of the EPT (test structure, administration, result reporting, aims and purposes, etc.)?
   b. What are you most satisfied with in terms of the ISU EPT as a placement test? the least?
   c. What do you know about the placement decisions or criteria?
   d. To what extent do you think the EPT is a fair test?

3. In what ways does the EPT have an effect on students’ learning? their college experience? anything else?
   a. How are the placement decisions beneficial for students’ learning?
   b. What are the advantages of being placed into/out of an ESL class? What are the disadvantages?

4. Does the EPT affect your ESL instructions, and if so, it is positive, negative or both? Please describe how it affects instruction. (e.g. what, how, who you teach)

5. What is your general impression of the ESL courses at ISU?
   a. What do you hope students will gain from their ESL coursework?
   b. In your observation, are students in general appropriately placed in your ESL courses?
   c. What are the main challenges that you have faced in teaching ESL courses at ISU?
   d. How do you see the connection or relationship between the ESL courses and content courses?
   e. Overall, to what extent do you think the ESL courses helped students improve English?
   f. How would you evaluate students’ progress in your ESL courses? Suppose these students took the EPT or the in-class test again at the end of a semester, would they be placed into a higher level course or pass the test?
g. To what extent do you think the ESL students are prepared for their content courses after they finished the ESL courses?

6. Did you use the diagnostic or in-class tests in your ESL courses?
   a. How was students’ performance on the writing section evaluated on the EPT?
   b. What reactions do you think students have when they learned their diagnostic or in-class test results?
   c. Did the information from the diagnostic or in-class test inform or affect your teaching?

7. What is your general impression of the English proficiency of ESL students in your ESL classes (Choose Engl99R, 99L, 101B, 101C, or 101D)?
   a. What are the typical weaknesses and strengths of the ESL students in English (English grammar, vocabulary, writing ability, speaking ability, listening ability, reading ability, and etc.)?
   b. How satisfied do you think students are with their EPT placement?
   c. How motivated are the students in your ESL courses?

8. In your opinion, what is the relationship between students’ English proficiency and their performance in your classes?
   a. Overall, will students’ limited English proficiency affect students’ academic achievement (for example, overall GPA)?
   b. Do you have any accommodating strategies to help the international students with relatively low English proficiency?

9. What are some of your suggestions for international students to improve their English and earn a good grade in your classes?

10. What are some of your suggestions for the English Placement Test Office to make a better service for you and your students?
    a. Are there any sections of the EPT would need improvement? If yes, which ones and how to improve?

11. Do you have any final comments?
APPENDIX E. INTERVIEW QUESTIONS FOR CONTENT COURSE INSTRUCTORS

1. Can you tell me something about yourself and your course?
   a. How long have you been working as an instructor at ISU?
   b. What kinds of fundamental courses have you taught at ISU? (Are they a part of the general education program?)
   c. Usually, how many students are enrolled in your classes and how many are international students?
   d. What is your teaching style (or how do you teaching these courses? Can you describe a typical class)? In your fundamental courses, what are the typical course activities and assignments?
   e. What are the typical first-year courses for non-native international students in your college/department?

2. What is your general impression of the English proficiency of your international students?
   a. In your observation, what are the challenges faced by international students in terms of English use and content learning in your classes?
   b. What are the typical weaknesses and strengths of the international students in English (English grammar, vocabulary, writing ability, speaking ability, listening ability, reading ability, and etc.)?
   c. To what extent do you think that these students in general are ready in terms of English language requirement in your class?
   d. In your observation, did some students need more instructions in English, such as Engl101B, Engl99R and 99L?

3. Have you heard about the English Placement Test for international students whose native language is not English at ISU (test structure, administration, result reporting, etc.)? If yes, what is your general impression of the test?

4. Have you heard about the English as a second language courses for international students? If yes, what is your general impression of the ESL courses at ISU?
   a. Are you aware that some of your students are taking ESL courses concurrently?
   b. Would you recommend some of your students to take ESL courses in order to prepare them better for your courses?

5. In your opinion, what is the relationship between students’ English proficiency and their performance in your classes?
   a. Overall, will students’ limited English proficiency affect students’ academic achievement (for example, overall GPA)?
   b. When you graded students’ performances (written and oral), to what extent do you usually pay attention to student’s language issues? Do you penalize students’ assignments for language issues?
c. Would it be helpful if you were given some information about the English proficiency of your international students?

d. Do you have any accommodating strategies to help the international students with relatively low English proficiency?

6. What are some of your suggestions for international students to improve their English and earn a good grade in your classes?

7. What are some of your suggestions for the English Placement Test Office to make a better service for you and your students?

8. Do you have any final comments?
APPENDIX F. INTERVIEW QUESTIONS FOR ACADEMIC ADVISORS

1. Can you tell me a little bit about yourself and your work as an academic adviser? (example: length of work, major duties, number of native English-speaking advisees and non-native English-speaking advisees, challenges/difficulties,)

2. Are you familiar with the English Placement Test for international students whose native language is not English at ISU (test structure, administration, result reporting, etc.)? What is your general impression of the EPT at ISU?
   a. How would compare the EPT with other English tests that students usually take (TOEFL, IELTS)?
   b. Are you aware of the exemption criteria for the EPT? If yes, what kind of comments did you hear from students regarding the criteria?
   c. What reactions do international undergraduate students usually have when they received their placement results?
   d. What kind of questions do international students usually ask about the EPT?
   e. How satisfied do you think students are with their EPT results?
   f. To what extent do you think the EPT is a fair English test to our students?
   g. In what ways does the EPT have an impact on students’ learning? their college experience? anything else?

3. Are you familiar with the English as a second language courses for international students? What is your general impression of the ESL courses at ISU?
   a. Do you understand why some of the students are required to take ESL courses (Engl101B, 99R, 99L, etc.)? (Or, Do you know the criteria used in the EPT?)
   b. In your observation, are those students in general appropriately placed in the English classes, such as Engl101B, Engl99R and 99L?
   c. What are the typical weaknesses of these students (who are required to take ESL courses) in English (English grammar, vocabulary, writing ability, speaking ability, listening ability, reading ability, and etc.)?
   d. What are the challenges, if any, faced by these students in terms of academic English and in their content courses?
   e. Are you aware that there is a diagnostic or in-class test in each ESL course, through which students may test out of the course? If yes, what kind of comments did you hear from your students regarding this diagnostic or in-class test?
   f. What kind of questions do international students usually ask about English courses?
   g. To what extent do you think the English courses will help your students improve English proficiency?

4. What is your general impression of the English proficiency of your students who do not need take ESL courses?
a. What are the typical weaknesses of these students in English (English grammar, vocabulary, writing ability, speaking ability, listening ability, reading ability, and etc.)?

b. To what extent do you think that these students in general are ready in terms of English language requirement at ISU?

c. Have you noticed any challenges faced by these students in terms of academic English and in their content courses?

5. What are the typical first-year courses for non-native international students in your college/department?
   a. When you advise international undergraduate students on course registration, what kind of factors would you consider (for example, EPT records or English proficiency, workload, etc.)?

6. In your opinion, what is the relationship between students’ English proficiency and their performance in content courses?
   a. Will students’ limited English proficiency affect students’ academic achievement (for example, overall GPA)?

7. Do you work with the same students during their academic years at ISU? If so, what kinds of improvement in English do you usually notice in your students?

8. What are some of your suggestions for international students to improve their English and achieve academic success at ISU?

9. What are some of your suggestions for the English Placement Test Office to make a better service for you and your students?

10. Do you have any final comments?
APPENDIX G. COURSE SYLLABUS FOR ENGL99L

Instructor: Office: Landscape Architecture (LA) 203
Email: @iastate.edu Office hours: MTWR 1:00-2:00 pm
Class: Class website:

http://courses.isucomm.iastate.edu

Required Materials

➢ Access to the course website to be used for recording digital audio files.
➢ Earphones/earbuds

Objectives

➢ To improve the comprehension of spoken English
➢ To develop strategies for listening to academic lectures in English
➢ To develop note-taking skills
➢ To develop strategies for listening and speaking in non-academic situations
➢ To acquire vocabulary in English

Grading

➢ 25% Daily Homework
  Automatically Graded Listening Comprehension
➢ 25% Listening Journals
  Practice Listening with Outside Materials
➢ 15% - Midterm Exam
  Exam in 8th Week
➢ 15% - Final Exam
  Exam at end of course
➢ 10% Major Project Presentation
  5-6 minute presentation on chosen video
➢ 10% - Participation in class activities

Policies

Note: In addition to the student disciplinary regulations established by the Dean of Students Office at Iowa State University, which are available online at the following URL: (http://policy.iastate.edu/policy/SDR), the following course-specific policies are in place:

➢ Students must be punctual and bring the necessary materials (book, notebook etc.) to class. Students arriving more than 5 minutes late will be counted as ½ absent for that hour.
➢ Attendance is required. More than TWO absences will negatively affect your grade. Specifically
  o 3 absences: - 5 points from final grade
  o 4 absences: -10 points from final grade
  o 5 absences: automatically fail the course
➢ Plagiarism (copying work from others) is grounds for automatically failing this course.
➢ Computers are provided to support your learning. Using them for other purposes during class time may result in counting you as absent for the day. Texting is not allowed.
Homework

Homework will be posted on the course website. You must check the course website for the assigned homework before each class period. Most homework will be due online through the course website.

Your homework to do before the next class period is to enroll in our course website. To do this:

1. Go to http://courses.isucomm.iastate.edu
2. If you have already created an account in Moodle log-in. If not, please “Create an Account” for yourself. Follow directions given on screen to set up account and log-in.
3. Find our course section, we are 99L Section 01, with Instructor:
4. The enrollment code is isurocks.
5. Once successfully enrolled, take the Introductory Questionnaire.

In the next class period you will be given a tour of the course website.

Special Needs

Please address any special needs or special accommodations with me at the beginning of the semester or as soon as you become aware of your needs. Those seeking accommodations based on disabilities should obtain a Student Academic Accommodation Request (SAAR) form from the Disability Resources (DR) office in Student Services Building, Room 1076 (515-294-6624).

Diversity Affirmation

Iowa State University does not discriminate on the basis of race, color, age, ethnicity, religion, national origin, pregnancy, sexual orientation, gender identity, sex, marital status or status as a U.S veteran. Effective learning environments value and support diversity.
APPENDIX H. COURSE SYLLABUS FOR ENGL101b

English
101B
Academic Writing I

Textbook: Engaging Writing 1: Essential Skills for Academic Writing by Mary Fitzpatrick

Class websites: Instructions for registering and enrolling for these websites will be provided.

- Main course: http://courses.isucomm.iastate.edu.

Objectives: English 101B is designed to help you prepare for the writing requirements of your university work. It covers advanced points of English grammar as they apply to aspects of academic writing, as well as the key writing processes of understanding assignments, developing and organizing ideas, drafting, and revising. You will also learn how to analyze and improve your own writing by working with state-of-the-art writing-support technologies. A recurring journal-writing assignment will help you to reflect on and consolidate your learning on the course. If you successfully complete the course, you will have increased your skills and confidence in writing at the sentence and paragraph levels and be ready to move on to English 101C.

Evaluation

<table>
<thead>
<tr>
<th>Task</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 major paragraph assignments</td>
<td>40%</td>
</tr>
<tr>
<td>1 major essay assignment</td>
<td>15%</td>
</tr>
<tr>
<td>Weekly journal writing</td>
<td>15%</td>
</tr>
<tr>
<td>Homework</td>
<td>10%</td>
</tr>
<tr>
<td>Classwork and participation (including individual conferences)</td>
<td>10%</td>
</tr>
<tr>
<td>Group presentation</td>
<td>5%</td>
</tr>
<tr>
<td>Final timed writing assignment</td>
<td>5%</td>
</tr>
<tr>
<td>FINAL GRADE</td>
<td>100%</td>
</tr>
</tbody>
</table>

Writing assignments: You will have five paragraph-length (200-250 words) writing assignments, each worth 10% of your final grade, which together will account for half of your grade on the course. Because the focus of this course is on writing as a process, each paragraph assignment will be submitted in three stages: first draft, second draft, and final draft, all of which will be graded (see details below). The first and second drafts should be written in GoogleDrive so you and your teacher have access to a detailed record of each composition as it evolves. You can also use the comment feature of GoogleDrive to have “conversations” with your teacher about specific details of content, organization, use of language, etc. The topics and evaluation criteria for these writing assignments will be provided at the appropriate time. Your grade will depend in part on how much effort you put into responding to your instructor’s feedback between the first and final drafts.
Other work: In addition to the paragraph assignments, you will have other in-class work and homework, including reading as preparation for writing, grammar- and vocabulary-related exercises, writing strategy practice, and possibly peer reviews of your classmates’ writing. Your grade will also be based on your participation in class activities and attendance at individual conferences with your instructor. On the first day of class there will be a timed diagnostic writing assignment to help your teacher determine your strengths and areas where you need improvement; on the scheduled final exam day, there will be another timed writing assignment to gauge your progress over the semester.

On-time submission: All work must be completed by the assigned deadline.

Academic honesty: Plagiarism – that is, using someone else’s work and presenting it as your own – is unacceptable in any academic context. Understanding what constitutes plagiarism and academic dishonesty will help you avoid these problems and will strengthen your writing. Plagiarism is a serious legal and ethical issue and it is treated as such by the university. Detecting plagiarism in English 101B is often easy for an instructor who is familiar with your work, and if they find it, they are required by university policy to report it to the Office of the Dean of Students. If you have any questions about plagiarism, see your instructor before turning in an assignment.

Attendance and participation: Classes are conducted in a discussion/workshop format and depend on your active learning. Therefore, regular attendance and productive, courteous participation with classmates and the instructor are important. Absences damage your grade and too many may result in you having to drop the course. Most of our work in English 101B cannot be rescheduled, made up, or accepted late, regardless of your reasons for missing class. To ensure that you stay on track with your attendance and assignments, the following policies, developed by university officials, will be enforced in all sections of English 101B:

Missing more than four classes (MWF) will lower your grade, and excessive absences (more than three weeks; 9 days) will result in a failing grade for the course.

- Specifically, absences after four (MWF) will reduce your class grade by two steps (for example, a B+ becomes a B-; a C becomes a D+), and after a total of nine (MWF) absences, you must drop the course or you will receive an F.
- If you are more than 10 minutes late to class, you will be counted absent. You must let me know at least one hour before class the day of your absence in order to be excused. Valid excuses include: family emergency, short term personal illness, or transportation issues. I will not excuse absences for doctor appointments (unless they are an emergency), incomplete assignments, or lack of preparation for class. If it is necessary for you to miss class, e-mail me or ask a classmate for help so that you can have your assignment ready for the next class period.
- When classes are cancelled for individual conferences, you must be present at the conference time you have chosen. Missing a scheduled individual conference counts as an absence and also lowers your grade in participation grade.
- Excessive absences will result in a low midterm grade report, at which point your advisor will automatically be notified.

Grading and evaluation: While we assume that students admitted to the university can perform satisfactorily most of the time, earning an A on this course will require strong, consistent effort. To make sure you know what is expected, you should read all assignments carefully, especially
the evaluation criteria, and respond thoughtfully and thoroughly to the feedback from your instructor. Start assignments early and work steadily to avoid last-minute rushing. If you do not understand an assignment, the evaluation criteria or feedback, or if you have any questions about your grade on an assignment, visit me during office hours.

Grading criteria for 5 major writing assignments

<table>
<thead>
<tr>
<th>Draft</th>
<th>Weight</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>25%</td>
<td>Timeliness; response to the task</td>
</tr>
<tr>
<td>Second</td>
<td>25%</td>
<td>Timeliness, response to the task; changes from first draft</td>
</tr>
<tr>
<td>Final</td>
<td>50%</td>
<td>Timeliness; response to the task; changes from previous drafts, plus other qualities outlined on the assignment rubric</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Overall course grading scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>94-100%</td>
<td></td>
</tr>
<tr>
<td>A-</td>
<td>90-93%</td>
<td></td>
</tr>
<tr>
<td>A+</td>
<td>87-89%</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>83-86%</td>
<td></td>
</tr>
<tr>
<td>B-</td>
<td>80-82%</td>
<td></td>
</tr>
<tr>
<td>B+</td>
<td>77-79%</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>73-76%</td>
<td></td>
</tr>
<tr>
<td>C+</td>
<td>70-72%</td>
<td></td>
</tr>
<tr>
<td>C-</td>
<td>67-69%</td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>63-66%</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>60-62%</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>59% and below</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>59% and below</td>
<td></td>
</tr>
</tbody>
</table>

Disability accommodation: If you have a disability and require special accommodations, please contact me early in the semester so that your learning needs may be appropriately met.

Email: Please use appropriate and courteous forms of address and expression in your emails to me, and I will do the same. An email message to a university instructor should not read like something you dash off to a friend. I get a lot of email from students, so in your subject line be sure to write “101B” as well as the purpose of the email.

Use of English: Talking in your native language in class is strongly discouraged as it detracts from your learning and that of your fellow students.

Mobile phone policy: Talking on your phone during class is prohibited. Please turn off or silence your cell phone before class begins. I will take a point off from your grade each time I find your phone is ringing or I see you talking on the phone in class.

If you need help, remember …

- Come see me. You can drop by during office hours or make an appointment to see me at another time if you have questions about the course or other issues.
- The Writing and Media Help Center (300 Carver) provides individualized help from trained tutors, and is free for ISU students. You may email them at writectr@iastate.edu or call them at 515-294-5411 to set up an appointment. For more information, go to www.dso.iastate.edu/wmc.
APPENDIX I. COURSE SYLLABUS FOR ENGL101C

Overview
The purpose of English 101C is to prepare undergraduate non-native speakers of English for success in all academic communication assignments with an emphasis on written work. Students who satisfactorily engage in this course will be prepared to enter English 150.

Goals
Upon completion of this course, students will be able to:

✓ Understand the demands of written assignments in their courses
✓ Engage in discussion, provide commentary, and contribute to dialogue and consensus in small and large groups
✓ Think critically; perform analysis, critique, synthesis, and evaluation
✓ Perform close readings of written and multimedia texts
✓ Use the process of multiple drafts and feedback to revise and improve composition
✓ Be independent writers who can identify weaknesses, evaluate effectiveness, and revise compositions
✓ Proofread, edit, and correct drafts for common errors of syntax, mechanics, and word choice

General Requirements (details follow)

➤ All 5 major assignments plus the Final Exam must be completed: Missing any results in failing the course.
➤ Attendance and participation must be maintained: absences or lack of preparation will lower your grade and can result in a failing grade.
➤ The textbook is REQUIRED and each student must have his or her own copy of the textbook to bring to class.
➤ Minor assignments practice the skills needed to fulfill major assignments and are required for passing the course.
➤ The course website contains essential information and must be actively used.
➤ The class format is a workshop style, which means students will arrive prepared to work actively while in class.

Major Assignments
There are five major writing assignments during the semester plus a written final scheduled by the University (see http://www.registrar.iastate.edu/students/exams/fallexams). To complete the first five, you will submit a first draft, participate in writing workshops and exercises for revising
the draft, and submit a final draft for the assignment grade. You may have to do more than one revision before the assignment is complete depending on peer and instructor feedback.

<table>
<thead>
<tr>
<th>Major Assignment Category</th>
<th>Theme</th>
<th>Due</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Expository Description</td>
<td>Role Models</td>
<td>Week 2</td>
<td>10%</td>
</tr>
<tr>
<td>#2 Interview and Analysis</td>
<td>Culture and Identity</td>
<td>Week 4</td>
<td>20%</td>
</tr>
<tr>
<td>#3 Group Film Critique Wiki</td>
<td>Families in the Movies</td>
<td>Week 8</td>
<td>20%</td>
</tr>
<tr>
<td>#4 Summary and Response</td>
<td>Global Economics</td>
<td>Week 11</td>
<td>20%</td>
</tr>
<tr>
<td>#5 Article Synthesis or Evaluation</td>
<td>Economics or Lifestyles</td>
<td>Week 15</td>
<td>20%</td>
</tr>
<tr>
<td>Final Exam – Timed Writing</td>
<td></td>
<td>Dec 15-19</td>
<td>10%</td>
</tr>
</tbody>
</table>

Due dates and detailed requirements of each draft of the major assignments are specified on the assignment sheets. Make sure you have a backup electronic copy of all work before you turn it in to be graded. Major Assignments can be penalized one letter grade (e.g., from B to C) for each class period they are late.

Minor Assignments – Coursework, Attendance, and Participation

Daily coursework and homework are part of what we learn in this course. There is no substitute for doing the work and practicing the skills involved. Coursework consists of:

- **Textbook and other reading assignments:** Readings must be completed before class and reading responses, discussions, and exercises are frequent.
- **Quizzes and other class activities, discussions, or postings:** Your thoughts and commentary are required contributions to the class. Be ready to use the course website or class discussions to interact and contribute. Be prepared for class.
- **Group and Partner Work:** When you are asked to work with your classmates, you are responsible to make it a successful collaboration even when you might prefer to work alone.
- **Attendance:** Much of what we do in English 101 cannot be rescheduled for you individually, made up, or accepted late, regardless of your reason for missing class. Therefore, the Coordinator of 101C mandates that the following policy be enforced in all sections of English 101:
  - Missing more than three classes will lower your grade, and excessive absences can result in a failing grade for the course. Specifically, absences after three will reduce your class grade by a step (a B+ becomes a B; a C- becomes a D+), and after a total of six absences, or if you miss more than three in a row, you must drop the course or you will receive an F. Class meets for the Final Exam period scheduled by the Registrar’s Office the week of December 15-19. See [http://www.registrar.iastate.edu/students/exams](http://www.registrar.iastate.edu/students/exams)
  - Even with a valid reason to miss, you can accumulate so many absences in a semester that your work and classroom experience are too compromised for you to remain in the class. If you have too many absences to remain in English 101, you may be advised to drop the class and take it in a semester when you can attend regularly.

<table>
<thead>
<tr>
<th>Minor Assignment Categories</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textbook Assignments and Readings</td>
<td>50%</td>
</tr>
<tr>
<td>Quizzes and Daily Classwork</td>
<td>15%</td>
</tr>
<tr>
<td>Group Work/Partner Activities</td>
<td>15%</td>
</tr>
<tr>
<td>Attendance and Participation</td>
<td>20%</td>
</tr>
<tr>
<td>Total Weight for Course Grade</td>
<td>30%</td>
</tr>
</tbody>
</table>
Late arrivals less than 10 minutes count as one third of an absence. If you are more than 10 minutes late you should still come to participate, but you might be counted absent.

When conferences are scheduled, missing or not scheduling an individual or group conference counts as an absence.

Your advisor may be notified of attendance or coursework issues that threaten your ability to pass the class or you may receive a poor midterm grade report.

Participation not only includes the above homework and coursework preparation and contributions, but also requires you to use common courtesy, including the following rules:

- NO FOOD is ever allowed in the media classrooms. You may bring drinks only to the regular classroom.
- All electronic devices including cell phones and electronic dictionaries must be turned off and put away throughout the class period. Unauthorized use of electronic devices during class counts as an absence.
- Use of computers in the classroom is strictly limited to the classroom activity only. Use of computers during non-designated times or for non-designated purposes results in an absence for the day.
- You are counted as absent if you do not actively speak, listen, and contribute to class activities IN ENGLISH, or have not done the reading in advance of class, you do not have your textbook, or are engaged in non-course related activities.
- You are the only one responsible for making sure you know what the assignments and due dates are and for keeping track of whether or not you have done the work. If you don’t know or don’t understand, you are the one who must find out where to get the answers. KEEP copies of all your work.

Grading and Evaluation

The work required of you at the university is often more difficult than what you did elsewhere. Expectations and standards are also higher since you are now pursuing a university degree in a language other than your native language. Therefore, earning As and Bs at ISU requires strong, consistent effort.

Your assignment sheets in English 101C include evaluation criteria to help you understand the required work. Be realistic about what it takes to get good grades; start assignments early and work steadily to avoid last-minute rushing; make an appointment with your instructor or the Writing and Media Center for support (http://www.dso.iastate.edu/wmc) before you get into difficulties.

Academic Honesty

Plagiarism is using someone else’s work, turning in work you did not do, or using someone else’s words or ideas and presenting it without citing the source, or using cited sources without sufficient paraphrasing. It is unacceptable and irresponsible. Understanding what constitutes plagiarism and academic dishonesty will help prevent you from committing these acts inadvertently and will strengthen your writing.

Plagiarism is a serious legal and ethical breach, and is treated as such by the university.

<table>
<thead>
<tr>
<th>Grade Scale Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
</tr>
<tr>
<td>100%</td>
</tr>
<tr>
<td>92%</td>
</tr>
<tr>
<td>89%</td>
</tr>
<tr>
<td>86%</td>
</tr>
<tr>
<td>82%</td>
</tr>
<tr>
<td>79%</td>
</tr>
<tr>
<td>76%</td>
</tr>
<tr>
<td>72%</td>
</tr>
<tr>
<td>69%</td>
</tr>
<tr>
<td>66%</td>
</tr>
<tr>
<td>59%</td>
</tr>
</tbody>
</table>
work, and once detected, it is mandatory that the ENGL 101C Coordinator be notified and consulted about consequences.

If you have questions about using outside sources, see your instructor or the Writing and Media Center before you turn in an assignment. The Library also can help you, http://instr.iastate.libguides.com/content.php?pid=10314.

Resources

ISU’s Writing and Media Center: The English Department’s Writing/Media Help Center, located in Carver 300, offers ISU students free, individualized instruction in all aspects of writing. You may visit the center on your own or on the recommendation of a teacher; you may drop in or sign up for a scheduled appointment. Call 294-7430 or check out the website: http://www.dso.iastate.edu/wmc. Business Majors have the College of Business Communications Center just for you: www.business.iastate.edu/communications-center

ISU Library: People at the Library and online are ready to help you – use this valuable resource to your advantage – www.lib.iastate.edu

Student Disability Resources: Please address any special needs or special accommodations with me at the beginning of the semester or as soon as you become aware of your needs. Those seeking accommodations based on disabilities should obtain a Student Academic Accommodation Request (SAAR) form from the Student Disability Resources (SDR) office. SDR is on the main floor of the Students Services Building.

Diversity Affirmation

Iowa State University does not discriminate on the basis of race, color, age, religion, national origin, sexual orientation, sex, marital status, or disability. An effective learning environment not only values but also welcomes and supports diversity and the open discussion of diverse thought. The environment in the classroom is a safe place to discuss any topic: All perspectives must be allowed. Anyone who negatively impacts the comfort or safety of open discussion will be referred to Student Services for diversity training and support. Your instructor promises to help maintain the comfort and safety of all.
APPENDIX J. ESSAY PROMPTS USED IN THE IN-CLASS TESTS IN ENGL101B/C

English 101B/C, Fall 2014 – In-Class Timed Essay
Your essay here will determine if you are correctly placed in this course. Your performance will not be graded, but a very strong essay can pass you into a higher-level course. This writing sample will also serve to diagnose your writing strengths and weaknesses for this course.

A strong essay has:
- a clear thesis stating the main idea of the essay – the main idea answers the prompt.
- body paragraphs are developed with details and support the main idea
- grammar, spelling, and punctuation are correct enough to make reading easy
- approximately 350 words in length

Read the prompt carefully, plan your response, and leave extra time at the end to review and edit. You have 40 minutes to complete this task.

ESSAY TOPIC: Admirable Characteristics

When we admire someone, we value certain qualities or abilities they have. Sometimes, we hope to develop these same qualities in ourselves. This person we admire is a role model for how we want to be. Think of a friend, a family member, a teacher, or someone you know very well who is your role model. Describe a few of his or her characteristics that you admire. Provide examples.

Write your answer here and on additional paper. Include any pre-writing, planning, or outlining. Crossing out, writing over, and other editing marks on the pages are perfectly acceptable. Turn in this sheet on top of your essay.
### APPENDIX K. RATING RUBRIC FOR THE EPT WRITING SECTION

<table>
<thead>
<tr>
<th>Levels</th>
<th>General Descriptions</th>
<th>Organization</th>
<th>Grammar &amp; Vocab</th>
<th>Functional</th>
<th>Mechanics</th>
<th>Comprehensibility</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pass</strong></td>
<td>No need for ESL instruction. A few grammatical, mechanical, and stylistic errors not hindering comprehension. Undergrads can go into first-year composition (ENGL 150/250) and compete adequately with native speakers. Graduates have adequate writing skill for graduate work in their field. Roughly corresponding to Advanced Mid or higher in ACTFL Proficiency Guidelines 2012 - Writing</td>
<td>Utilize a variety of cohesive devices</td>
<td>Adequate use of time frames with moderate control of aspect (e.g., perfect, progressive, habitual, momentary, etc.)</td>
<td>Argue and describe with details</td>
<td>A few mistakes in punctuation and spelling</td>
<td>Readily understood by native speakers not accustomed to non-native speakers’ writing</td>
</tr>
<tr>
<td>101C(UG)/101D(G)</td>
<td>Students can benefit from one semester of ESL instruction. Communication on formal topics is somewhat effective but often involves redundancy or repetition. Corresponding to Advanced Low in ACTFL Proficiency Guidelines 2012 - Writing</td>
<td>Use a limited number of cohesive devices</td>
<td>Moderate use of time frames with fair control of aspect</td>
<td>Narrate with less sophisticated descriptions</td>
<td>Moderate degree of mistakes in spelling and/or punctuation</td>
<td>Can be understood by native speakers not accustomed to non-native speakers’ writing with some effort or patience</td>
</tr>
<tr>
<td>101B</td>
<td>Students need more than one semester of ESL instruction. They will proceed to 101C/D after completing 101B. Communication on formal topics is limited. Corresponding to Intermediate High in ACTFL Proficiency Guidelines 2012 - Writing</td>
<td>Very limited use of cohesive devices</td>
<td>Inadequate/inconsistent use of time frames with lack of control of aspect</td>
<td>Simple communication</td>
<td>Frequent systematic errors or random mistakes in punctuation and/or spelling</td>
<td>Requires a lot of effort to understand</td>
</tr>
</tbody>
</table>

---

**Grammar & Vocab**
- Inadequate/inconsistent use of time frames with lack of control of aspect
- Mostly rely on simple sentence structures
- Connect sentences largely relying on coordinate conjunctions (e.g., *and, but, or*) and common subordinate conjunctions (e.g., *because, if, when*)
- Many grammatical mistakes hindering readers’ comprehension
- Limited use of vocabulary (repetition of the same expressions or redundancy)
- Incomplete sentences

**Functional**
- Simple communication
- Narrate, describe, and/or argue with lack of details
- Ineffective comparison and contrast
- Moderate use of true conditionals, often making mistakes with verb forms
- Inappropriate use of formal/informal expressions

**Mechanics**
- Moderate degree of mistakes in spelling and/or punctuation

---

**Comprehensibility**
- Readily understood by native speakers not accustomed to non-native speakers’ writing
- Little efforts to make to understand
- Ideas fully developed with conspicuous clarity (albeit a few grammatical, mechanical, and stylistic mistakes)
Advanced High: Writers at the Advanced High sublevel are able to write about a variety of topics with significant precision and detail. They can handle informal and formal correspondence according to appropriate conventions. They can write summaries and reports of a factual nature. They can also write extensively about topics relating to particular interests and special areas of competence, although their writing tends to emphasize the concrete aspects of such topics. Advanced High writers can narrate and describe in the major time frames, with solid control of aspect. In addition, they are able to demonstrate the ability to handle writing tasks associated with the Superior level, such as developing arguments and constructing hypotheses, but are not able to do this all of the time; they cannot produce Superior-level writing consistently across a variety of topics treated abstractly or generally. They have good control of a range of grammatical structures and a fairly wide general vocabulary. When writing at the Advanced level, they often show remarkable ease of expression, but under the demands of Superior-level writing tasks, patterns of error appear. The linguistic limitations of Advanced High writing may occasionally distract the native reader from the message.

Advanced Mid (Corresponding to Pass in EPT): Writers at the Advanced Mid sublevel are able to meet a range of work and/or academic writing needs. They demonstrate the ability to narrate and describe with detail in all major time frames with good control of aspect. They are able to write straightforward summaries on topics of general interest. Their writing exhibits a variety of cohesive devices in texts up to several paragraphs in length. There is good control of the most frequently used target-language syntactic structures and a range of general vocabulary. Most often, thoughts are expressed clearly and supported by some elaboration. This writing incorporates organizational features both of the target language and the writer's first language and may at times resemble oral discourse. Writing at the Advanced Mid sublevel is understood readily by natives not used to the writing of non-natives. When called on to perform functions or to treat issues at the Superior level, Advanced-Mid writers will manifest a decline in the quality and/or quantity of their writing.

Advanced Low (Corresponding to 101C/D in EPT) Writers at the Advanced Low sublevel are able to meet basic work and/or academic writing needs. They demonstrate the ability to narrate and describe in major time frames with some control of aspect. They are able to compose simple summaries on familiar topics. Advanced Low writers are able to combine and link sentences into texts of paragraph length and structure. Their writing, while adequate to satisfy the criteria of the Advanced level, may not be substantive. Writers at the Advanced Low sublevel demonstrate the ability to incorporate a limited number of cohesive devices, and may resort to some redundancy and awkward repetition. They rely on patterns of oral discourse and the writing style of their first language. These writers demonstrate minimal control of common structures and vocabulary associated with the Advanced level. Their writing is understood by natives not accustomed to the writing of non-natives, although some additional effort may be required in the reading of the text. When attempting to perform functions at the Superior level, their writing will deteriorate significantly.

Intermediate High (Corresponding to 101B in EPT): Writers at the Intermediate High sublevel are able to meet all practical writing needs of the Intermediate level. Additionally, they can write compositions and simple summaries related to work and/or school experiences. They can narrate and describe in different time frames when writing about everyday events and situations. These narrations and descriptions are often, but not always, of paragraph length, and they typically contain some evidence of breakdown in one or more features of the Advanced level. For example, these writers may be inconsistent in the use of appropriate major time markers, resulting in a loss of clarity. The vocabulary, grammar and style of Intermediate High writers essentially correspond to those of the spoken language. Intermediate High writing, even with numerous and perhaps significant errors, is generally comprehensible to natives not used to the writing of non-natives, but there are likely to be gaps in comprehension.
APPENDIX L. SUMMARY OF THE LEXICAL COMPLEXITY MEASURES IN THE
LEXICAL COMPLEXITY ANALYZER (LCA)

<table>
<thead>
<tr>
<th>Category</th>
<th>Measures</th>
<th>Code</th>
<th>Definition or formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical density</td>
<td>Lexical density</td>
<td>LD</td>
<td>The ratio of the number of lexical words to the total number of words in a text. ( \frac{N_{\text{lex}}}{N} )</td>
</tr>
<tr>
<td>Lexical sophistication</td>
<td>Lexical sophistication-I</td>
<td>LS1</td>
<td>The ratio of the number of sophisticated lexical words to the total number of lexical words in a text. ( \frac{N_{\text{sllex}}}{N_{\text{lex}}} )</td>
</tr>
<tr>
<td></td>
<td>Lexical sophistication-II</td>
<td>LS2</td>
<td>Laufer’s (1994) Lexical Frequency Profile. ( \frac{T_{s}}{T} )</td>
</tr>
<tr>
<td></td>
<td>Verb sophistication-I</td>
<td>VS1</td>
<td>The ratio of the number of sophisticated verbs to the total number of verbs in a text. ( \frac{N_{\text{svverb}}}{N_{\text{verb}}} )</td>
</tr>
<tr>
<td></td>
<td>Corrected verb sophistication-I</td>
<td>CVS1</td>
<td>Corrected version of VS1: ( \frac{T_{\text{svverb}}}{\sqrt{2N_{\text{verb}}}} )</td>
</tr>
<tr>
<td></td>
<td>Verb sophistication-II</td>
<td>VS2</td>
<td>Squared version of VS1 to reduce the effect of sample size on the measure. ( \frac{T_{\text{svverb}}^{2}}{N_{\text{verb}}} )</td>
</tr>
<tr>
<td>Lexical variation</td>
<td>Number of different words (NDW)</td>
<td>NDW</td>
<td>Number of different words</td>
</tr>
<tr>
<td></td>
<td>NDW (first 50 words)</td>
<td>NDWZ</td>
<td>Number of different words in the first 50 words in a text</td>
</tr>
<tr>
<td></td>
<td>NDW (expected random 50)</td>
<td>NDWERZ</td>
<td>Mean of numbers of different words in 10 random 50-word samples in a text</td>
</tr>
<tr>
<td></td>
<td>NDW (expected sequence 50)</td>
<td>NDWESZ</td>
<td>Mean of numbers of different words in 10 random 50-word sequences</td>
</tr>
<tr>
<td></td>
<td>Type-token ratio (TTR)</td>
<td>TTR</td>
<td>The ratio of the number of word type to the number of total words in a text</td>
</tr>
<tr>
<td></td>
<td>Mean segmental TTR (50)</td>
<td>MSTTR</td>
<td>mean TTR of a number of 50-word segments from the same text</td>
</tr>
<tr>
<td></td>
<td>Corrected TTR</td>
<td>CTTR</td>
<td>Corrected version of TTR: ( \frac{T}{\sqrt{2N}} )</td>
</tr>
<tr>
<td></td>
<td>Root TTR</td>
<td>RTTR</td>
<td>Square rooted version of TTR: ( \frac{T}{\sqrt{N}} )</td>
</tr>
</tbody>
</table>
### Bilogarithmic TTR

<table>
<thead>
<tr>
<th>Uber Index</th>
<th>UBER</th>
<th>Bilogarithmic version of TTR: $\frac{\log T}{\log N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D measure</td>
<td>D</td>
<td>A transformation of TTR to reduce sample size effect</td>
</tr>
<tr>
<td>Lexical word variation</td>
<td>LV</td>
<td>The ratio of the number of lexical word types to the total number of lexical words in a text. $T_{\text{lex}}/N_{\text{lex}}$</td>
</tr>
<tr>
<td>Verb variation-I (VV1)</td>
<td>VV1</td>
<td>The ratio of the number of verb types to the total number of verbs in a text. $T_{\text{verb}}/N_{\text{verb}}$</td>
</tr>
<tr>
<td>Squared VV1</td>
<td>SVV1</td>
<td>$T_{\text{verb}}^2/N_{\text{verb}}$</td>
</tr>
<tr>
<td>Corrected VV1</td>
<td>CVV1</td>
<td>$T_{\text{verb}}/\sqrt{2N_{\text{verb}}}$</td>
</tr>
<tr>
<td>Verb variation-II (VV2)</td>
<td>VV2</td>
<td>$T_{\text{verb}}/N_{\text{lex}}$</td>
</tr>
<tr>
<td>Noun variation</td>
<td>NV</td>
<td>$T_{\text{noun}}/N_{\text{lex}}$</td>
</tr>
<tr>
<td>Adjective variation</td>
<td>ADJV</td>
<td>$T_{\text{adj}}/N_{\text{lex}}$</td>
</tr>
<tr>
<td>Modifier variation</td>
<td>MODV</td>
<td>$(T_{\text{adj}} + T_{\text{adv}})/N_{\text{lex}}$</td>
</tr>
</tbody>
</table>

Note. This table is adapted from Lu (2012). T = Type, N = Token
### APPENDIX M. SUMMARY OF THE SYNTACTIC COMPLEXITY MEASURES IN L2 SYNTACTIC COMPLEXITY ANALYZER (L2SCA)

<table>
<thead>
<tr>
<th>Category</th>
<th>Measures</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of production</td>
<td>Mean length of clause</td>
<td>MLC</td>
</tr>
<tr>
<td></td>
<td>Mean length of sentence</td>
<td>MLS</td>
</tr>
<tr>
<td></td>
<td>Mean length of T-unit</td>
<td>MLT</td>
</tr>
<tr>
<td>Sentence complexity</td>
<td>Clauses per sentence</td>
<td>C/S</td>
</tr>
<tr>
<td>Subordination</td>
<td>Clause per T-unit</td>
<td>C/T</td>
</tr>
<tr>
<td></td>
<td>Complex T-units per T-unit</td>
<td>CT/T</td>
</tr>
<tr>
<td></td>
<td>Dependent clauses per clause</td>
<td>DC/C</td>
</tr>
<tr>
<td></td>
<td>Dependent clauses per T-unit</td>
<td>DC/T</td>
</tr>
<tr>
<td>Coordination</td>
<td>Coordinate phrases per clause</td>
<td>CP/C</td>
</tr>
<tr>
<td></td>
<td>Coordinate phrases per T-unit</td>
<td>CP/T</td>
</tr>
<tr>
<td></td>
<td>T-units per sentence</td>
<td>T/S</td>
</tr>
<tr>
<td>Particular structure</td>
<td>Complex nominals per clause</td>
<td>CN/C</td>
</tr>
<tr>
<td></td>
<td>Complex nominals per T-unit</td>
<td>CN/T</td>
</tr>
<tr>
<td></td>
<td>Verb phrases per T-unit</td>
<td>VP/T</td>
</tr>
</tbody>
</table>

Note. This table is adapted from Lu (2010). C = clause, T=T-unit, S = sentence, DC = dependent clause, CN = complex nominal, VP = verb phrases, CP = coordinate phrases
<table>
<thead>
<tr>
<th>Trait Category</th>
<th>Error type</th>
<th>Error examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>Fragments</td>
<td>The last time we went backpacking in the mountains.</td>
</tr>
<tr>
<td></td>
<td>Pronoun Errors</td>
<td>Watching the star of the basketball team shoot free throws made it seem easy.</td>
</tr>
<tr>
<td></td>
<td>Run-on Sentences</td>
<td>As usual, the students are staging a school play this year, one of them also wrote the play.</td>
</tr>
<tr>
<td></td>
<td>Possessive Errors</td>
<td>The instructor's patience for noise waned by the end of the period.</td>
</tr>
<tr>
<td></td>
<td>Garbled Sentences</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Wrong or Missing words</td>
<td>The decided to talk to the administration about the parking problem.</td>
</tr>
<tr>
<td></td>
<td>*Subject-Verb Agreement</td>
<td>The football players is holding a pizza sale to raise money for new uniforms.</td>
</tr>
<tr>
<td></td>
<td>Proof read this!</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>*Ill-formed Verbs</td>
<td>Many students don't know what they going to do over their summer vacations.</td>
</tr>
<tr>
<td>Usage</td>
<td>Determiner Noun Agreement</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Faulty Comparisons</td>
<td>My sister is a gooder runner than my brother is.</td>
</tr>
<tr>
<td></td>
<td>*Missing or Extra Article</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Nonstandard word form</td>
<td>I kinda like dancing, even though I'm too shy to ask anyone to dance with me.</td>
</tr>
<tr>
<td></td>
<td>*Confused Words</td>
<td>First I will stop at the store, and than I will go home.</td>
</tr>
<tr>
<td></td>
<td>Negation Error</td>
<td>I am not going to pay no bills today.</td>
</tr>
<tr>
<td></td>
<td>Wrong Form of Word</td>
<td>I am capability of studying before dinner when I don't procrastinate.</td>
</tr>
<tr>
<td>Category</td>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Wrong Part of Speech</td>
<td><em>I will try to success in my studies.</em></td>
<td></td>
</tr>
<tr>
<td>*Preposition Error</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Wrong Articles</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spelling</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Missing Comma</td>
<td><em>The rise of the Internet has made it easier to do research but much of the information on the Internet is unreliable.</em></td>
<td></td>
</tr>
<tr>
<td>Capitalize Proper Nouns</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Hyphen Error</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Capitalize Initial Capital Letter in a sentence</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Fused Words</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Missing Question Mark</td>
<td><em>What time does the first class start in the morning</em></td>
<td></td>
</tr>
<tr>
<td>Compounds</td>
<td><em>Any thing my mother finds cute, I find hideous.</em></td>
<td></td>
</tr>
<tr>
<td>Missing Final Punctuation</td>
<td><em>That was the best vegetarian pizza I've ever tasted</em></td>
<td></td>
</tr>
<tr>
<td>Duplicates</td>
<td><em>My brother is as strong as a the work horse.</em></td>
<td></td>
</tr>
<tr>
<td>Missing Apostrophe</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Extra Comma</td>
<td><em>I have a cat, and a dog.</em></td>
<td></td>
</tr>
<tr>
<td>Style</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition of Words</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Short Sentence</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Inappropriate Words or Phrases</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Long Sentence</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Sentences Beginning with Coordinating Conjunction</td>
<td><em>But if I don't make the swim team, I won't be too disappointed.</em></td>
<td></td>
</tr>
<tr>
<td>Passive Voice</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td></td>
</tr>
</tbody>
</table>

Note. a. Error subcategories are based on *Criterion* v. 13.1. b. the error examples are taken from the Writer’s Handbook on *Criterion*. * Only the error types marked with asterisk were analyzed and reported for grammatical accuracy in this study.