A test of productive English grammatical ability in academic writing: Development and validation

Yoo-Ree Chung
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, and the Linguistics Commons

Recommended Citation
Chung, Yoo-Ree, 'A test of productive English grammatical ability in academic writing: Development and validation' (2014).
Graduate Theses and Dissertations. 13856.
https://lib.dr.iastate.edu/etd/13856

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.
A test of productive English grammatical ability in academic writing: Development and validation

by

Yoo-Ree Chung

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

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Major Professor
   Dan Douglas
   Volker Hegelheimer
   Frederick O. Lorenz
   Gregory D. Wilson

Iowa State University
Ames, Iowa
2014

Copyright © Yoo-Ree Chung, 2014. All rights reserved.
For my mother Jin-Ok
# TABLE OF CONTENTS

TABLE OF CONTENTS......................................................................................................... iii

LIST OF TABLES .................................................................................................................. vii

LIST OF FIGURES ................................................................................................................. ix

ACKNOWLEDGEMENTS ..................................................................................................... xi

ABSTRACT ........................................................................................................................... xiii

CHAPTER 1. INTRODUCTION ............................................................................................ 1
  1.1. Mandates and Background ............................................................................................ 1
  1.2. Issues from Theoretical and Empirical Perspectives .................................................... 2
  1.3. Goals of the Study ......................................................................................................... 6
  1.4. Significance of the Study .............................................................................................. 7
  1.5. Outline of the Dissertation ............................................................................................ 8

CHAPTER 2. LITERATURE REVIEW ............................................................................... 10
  2.1. Common Acquisitional Orders for Grammar in SLA ................................................. 10
  2.2. Developmental Stages Demonstrated in Language Assessment ................................ 13
  2.3. Validation Frameworks ............................................................................................... 17
      2.3.1. Ontological Conceptions of Validity ................................................................... 17
      2.3.2. Construct Validity as a Unitary Concept ............................................................. 20
  2.4. Chapter Summary ....................................................................................................... 23

CHAPTER 3. INTERPRETIVE ARGUMENT .................................................................... 24
  3.1. Interpretive Argument Framework ............................................................................. 24
  3.2. Research Questions ..................................................................................................... 29
  3.3. Chapter Summary ....................................................................................................... 31

CHAPTER 4. METHODOLOGY ......................................................................................... 33
  4.1. Test Materials.............................................................................................................. 36
      4.1.1. Item Design .......................................................................................................... 36
          4.1.1.1. Selection of target grammatical features ....................................................... 36
          4.1.1.2. Task design ................................................................................................... 39
          4.1.1.3. Writing test items .......................................................................................... 41
          4.1.1.4. Expected developmental stages of the target grammatical features ............. 42
      4.1.2. Scoring Rubric ....................................................................................................... 42
  4.2. Data Collection ........................................................................................................... 45
      4.2.1. Quantitative Data Collection I: Paper-based tests ............................................... 45
          4.2.1.1. Participants .................................................................................................... 45
          4.2.1.2. Data collection procedures ............................................................................ 46
      4.2.2. Quantitative Data Collection II: Computer-based test ......................................... 47
4.2.2.1. Participants ................................................................. 47
4.2.2.2. Data collection procedures ........................................ 47
4.2.3. Interrater Reliabilities .................................................... 48
4.2.4. Qualitative Data Collection ............................................. 49
4.2.4.1. Think-aloud protocols ................................................. 49
4.2.4.2. Individual interviews ................................................... 52
4.2.4.3. Writing sample collection and error analysis .................. 53
4.2.4.4. Individual interview with EPT coordinator ...................... 57
4.2.4.5. Focus group interviews with instructors of ESL writing courses 58
4.3. Data Analysis Methods ..................................................... 58
4.3.1. Quantitative Data Analysis Methods ................................ 59
4.3.1.1. Descriptive statistics (RQ2a) .......................................... 59
4.3.1.2. Two-way ANOVA (RQ 2b) ........................................... 59
4.3.1.3. Reliability analysis (RQ 2c) .......................................... 59
4.3.1.4. Independent two-sample t-test (RQ 2f) ............................ 60
4.3.1.5. ANOVA and a post hoc Scheffé test (RQ 2f & 2h) .......... 60
4.3.1.6. Spearman rho and Pearson correlations (RQs 2g, 2h & 2j) 60
4.3.1.7. Confirmatory factor analysis (RQ 2i) .......................... 61
4.3.1.8. Overlay scatterplot (RQ 2g) ......................................... 64
4.3.1.9. Identification of false positive in EPT exemptions (RQ 5) 64
4.3.2. Qualitative Data Analysis Methods ................................. 65
4.4. Chapter Summary ............................................................ 66

CHAPTER 5. RESULTS .......................................................... 67
5.1. Quantitative Evidence for the Validity Argument ................... 67
5.1.1. RQ 2a: Do the academic grammar test scores effectively discriminate among students of varying proficiency levels? (Warrant 2) 67
5.1.2. RQ 2b: Does gender, as a construct-irrelevant factor, significantly affect examinees’ performances on the academic grammar test? (Warrant 2) 69
5.1.3. RQ 2c: Are the assessment results generalizable across the universe of assessment tasks to an acceptable extent? In other words, is Cronbach’s alpha coefficient equal to or greater than .7? (Warrant 3) 74
5.1.4. RQ 2d: Does the academic grammar test demonstrate similar reliability across different samples from the same population? (Warrant 3) 74
5.1.5. RQ 2e: Are scores from the computer-based academic test comparable to those from the conventional paper-and-pencil test? (Warrant 3) 76
5.1.6. RQ 2f: Do students at different proficiency levels perform distinctively from each other on the academic grammar test? (Warrant 4) 77
5.1.7. RQ 2g: Does students’ performance on the academic grammar test correspond to their performance in a non-testing setting? (Warrant 4) 79
5.1.8. RQ 2h: Does the performance of the academic grammar test items reflect the theoretical assumptions that underlie their development? (Warrant 5) 84
5.1.9. RQ 2i: Do the selected target grammatical features have significant associations with the construct of productive grammatical writing ability? (Warrant 5) 90
5.1.10. RQ 2j: Does examinees’ performance on the academic grammar test correlate with their performance in other writing tests that engage grammatical ability at moderate degrees? (Warrant 5) ................................................................. 93
5.1.11. RQ 5: To what extent does the use of the academic grammar test help test users identify false positive decisions resulting from the decisions on ESL writing placements/exemptions made exclusively on the basis of examinee performance on the EPT essay test? (Warrant 7) ........................................................................................................ 95
5.2. Qualitative Evidence for the Validity Argument ................................................................ 99
5.2.1. RQ1: Do the test tasks on the academic grammar test require any field-specific knowledge of test takers? Does the test properly elicit evidence of examinees’ grammatical ability? (Warrant 2) ............................................................................ 100
5.2.2. RQ3: Does the decision-maker appreciate the value of grammatical ability in academic writing when making decisions on ESL placements/exemptions? (Warrant 6) ............................................................................................................. 104
5.2.3. RQ4: Could academic grammar test scores serve as a tangible ground for decisions on individual students’ ESL placements/exemptions? (Warrant 6) ........................................................................................................... 105
5.2.4. RQ6(a): If the scores on the academic grammar test were used to make decisions on ESL placement, how would students perceive its impact on them with respect to advantages and disadvantages of additional ESL instruction? (Warrant 7) ........................................................................................................... 106
5.2.5. RQ6(b): If the scores on the academic grammar test were used to make decisions on ESL placement, how would students perceive its impact on them with respect to the relationship between grammatical writing ability and academic writing? (Warrant 7) ........................................................................................................... 111
5.2.6. RQ7a: How much do ESL instructors appreciate the value of grammatical ability in academic writing? (Warrant 7) ........................................................................................................... 119
5.2.7. RQ7b: To what extent could ESL instructors benefit from feedback on their students’ grammatical abilities provided by the use of the academic grammar test? (Warrant 7) ........................................................................................................... 122
5.3. Chapter Summary ........................................................................................................... 125

CHAPTER 6. DISCUSSION AND CONCLUSION .................................................................. 126
6.1. Building a Validity Argument ...................................................................................... 126
6.1.1. Domain Description Inference ............................................................................. 127
6.1.2. Evaluation Inference .......................................................................................... 127
6.1.3. Generalization Inference .................................................................................... 129
6.1.4. Extrapolation Inference ...................................................................................... 131
6.1.5. Explanation Inference ....................................................................................... 134
6.1.6. Utilization Inference .......................................................................................... 137
6.1.7. Ramification Inference ....................................................................................... 138
6.2. Further Validation Issues .......................................................................................... 141
6.2.1. Task Design (Evaluation, Generalization, and Explanation) ............................ 141
6.2.2. Developmental Stages of Grammatical Features (Explanation) .................... 147
6.2.3. Curriculum Revision and Pedagogical Improvement (Ramification) .............. 151
6.3. Two Additional Issues on Test Implementation ....................................................... 153
6.4. Conclusion ................................................................................................................. 154
6.4.1. Summary of the Dissertation Study ..................................................................... 154
6.4.2. Limitations .......................................................................................................... 156
6.4.3. Suggestions for Future Research ........................................................................ 157
APPENDIX A: TEST SPECIFICATION ........................................................................... 159
APPENDIX B: ACADEMIC GRAMMAR TEST ............................................................. 163
APPENDIX C: SCORING RUBRIC .................................................................................. 165
APPENDIX D: SAMPLE INTERVIEW QUESTIONS FOR STUDENTS ......................... 169
APPENDIX E: ESSAY WRITING PROMPTS ................................................................. 170
APPENDIX F: EXAMPLES OF GRAMMATICAL ERRORS ......................................... 171
APPENDIX G: INTERVIEW QUESTIONS FOR EPT COORDINATOR ....................... 175
APPENDIX H: INTERVIEW QUESTIONS FOR ESL INSTRUCTORS ............................ 176
REFERENCES ..................................................................................................................... 177
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Interpretive Argument of the Academic Grammar Test</td>
<td>27</td>
</tr>
<tr>
<td>4.1</td>
<td>Target Grammatical Features of the Academic Grammar Test and their Expected Developmental Stages</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Example of a Scoring Rubric</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td>Categories and Subcategories of Grammatical Errors</td>
<td>55</td>
</tr>
<tr>
<td>4.4</td>
<td>Classification of Grammaticality Indices by Accuracy Category</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Contingency Table of the EPT Writing and the Academic Grammar Test Results</td>
<td>65</td>
</tr>
<tr>
<td>5.1</td>
<td>Descriptive Statistics of Academic Grammar Test Scores (F12 &amp; Sp13)</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Means (and Standard Deviations) of the Academic Grammar Test Total Scores by Gender and Proficiency Level</td>
<td>72</td>
</tr>
<tr>
<td>5.3</td>
<td>Tests of Between-Subjects Effects (F12)</td>
<td>73</td>
</tr>
<tr>
<td>5.4</td>
<td>Tests of Between-Subjects Effects (Sp13)</td>
<td>73</td>
</tr>
<tr>
<td>5.5</td>
<td>Cronbach’s Alpha Coefficients of Academic Grammar Test</td>
<td>75</td>
</tr>
<tr>
<td>5.6</td>
<td>Cronbach’s Alpha Coefficients of Academic Grammar Test by Gender</td>
<td>75</td>
</tr>
<tr>
<td>5.7</td>
<td>Descriptive Statistics and Cronbach’s Alpha Coefficient of Computer-based Academic Grammar Test</td>
<td>76</td>
</tr>
<tr>
<td>5.8</td>
<td>Group Statistics of EPT_Pass for F12 and Sp13</td>
<td>78</td>
</tr>
<tr>
<td>5.9</td>
<td>Descriptive Statistics of the Academic Grammar Test Scores by EPT Writing Results (F12 and Sp13)</td>
<td>79</td>
</tr>
<tr>
<td>5.10</td>
<td>Results of Post Hoc Scheffé Test for Differences in the Academic Grammar Test Score among the Three Proficiency Groups by EPT Writing Results (F12)</td>
<td>80</td>
</tr>
<tr>
<td>5.11</td>
<td>Results of Post Hoc Scheffé Test for Differences in the Academic Grammar Test Score among the Three Proficiency Groups by EPT Writing Results (Sp13)</td>
<td>80</td>
</tr>
</tbody>
</table>
Table 5.12. Descriptive Statistics of Interview Participants’ Productivity and Grammatical Accuracy Features ................................................................. 85
Table 5.13. Pearson Product-Moment Correlations between Academic Grammar Test Scores and Grammaticality Indices ......................................................... 86
Table 5.14. Item Statistics and Expected Developmental Stages ......................... 88
Table 5.15. Descriptive Statistics of IF Values by Item Groups (Stages 2, 3, and 4) .......................................................................................................................... 89
Table 5.16. Results of ANOVA Tests of IF values of Item Groups (Stages 2, 3, and 4) ....................................................................................................................... 89
Table 5.17. Results of Post Hoc Scheffé Tests for Differences among Item Groups in Their Mean IF Values (F12 and Sp13) ......................................................... 90
Table 5.18. Summary of Model Fit Indices of the CFA Model in Figure 5.4 ............ 92
Table 5.19. Means and Standard Deviations for TOEFL iBT® writing scores ....... 94
Table 5.20. Means and Standard Deviations of Academic Grammar Test Scores by EPT_WRT_Pass ........................................................................................................ 96
Table 5.21. Contingency Table I on EPT_WRT_Pass and Cut26 (F12 – All) .......... 97
Table 5.22. Contingency Table II on EPT_WRT_Pass and Cut26 (F12 – Unanimous Ratings) ................................................................................................................. 97
Table 5.23. Contingency Table I on EPT_WRT_Pass and Cut26 (Sp13 (Paper-based) – All) .............................................................................................................. 98
Table 5.24. Contingency Table II on EPT_WRT_Pass and Cut26 (Sp13 (Paper-based) – Unanimous Ratings) ............................................................................. 98
Table 6.1. Summary of Validity Argument for Academic Grammar Test ............ 142
Table 6.2. Students’ Perceptions on the Difficulty of the Task Types ................. 145
LIST OF FIGURES

Figure 3.1. Interpretive/validity argument framework for the academic grammar test ................................................................. 26

Figure 4.1. Sequential embedded mixed methods design for the validation study of the academic grammar test ................................................................. 34

Figure 4.2. Multiphase validation procedures of the academic grammar test ........ 35

Figure 4.3. Example of a gap-filling task without key words ......................... 39

Figure 4.4. Example of a gap-filling task with key words provided ..................... 40

Figure 4.5. Example of a jumbled-word-order task ........................................... 40

Figure 4.6. Question 8 on the academic grammar test ...................................... 44

Figure 4.7. Screenshot of the academic grammar test on Blackboard Learn .......... 48

Figure 4.8. Instructions for think-aloud protocols .......................................... 50

Figure 4.9. Academic Grammar Test on Moodle used for Think-Aloud Protocols .... 51

Figure 4.10. Path diagram of the first-order model 1 ......................................... 62

Figure 4.11. Path diagram of the first-order model 2 ......................................... 62

Figure 4.12. Path diagram of the second-order model 1 ..................................... 63

Figure 4.13. Path diagram of the second-order model 2 ..................................... 63

Figure 5.1. Histograms of the total scores of the Academic Grammar Test (F12 and Sp13) ........................................................................................................ 68

Figure 5.2. Q-Q plots of the academic grammar test scores by Gender (F12) ........ 70

Figure 5.3. Q-Q plots of the academic grammar test scores by Gender (Sp13) ....... 70

Figure 5.4. Slopes of marginal means of academic grammar test scores between the two proficiency groups by gender ......................................................... 72

Figure 5.5. 95% CI error bars of mean scores of the three proficiency groups by EPT writing ................................................................. 81
Figure 5.6. Overlay scatter plot on the relationship between E/W and academic grammar test scores (N=27) ................................................................. 82

Figure 5.7. Mean IFs of item groups by target grammatical features’ expected developmental stages ................................................................. 87

Figure 5.8. Best fit CFA model of the construct Grammatical Writing Ability on observed grammatical features .................................................. 94
ACKNOWLEDGEMENTS

I am gratefully indebted to my Major Professor, Dr. Carol A. Chapelle for her excellent guidance and support throughout my study at Iowa State University. Her trust and patience in me helped me obtain academic achievements I had never imagined of before joining the Applied Linguistics and Technology (ALT) program. I am also very grateful to Dr. Dan Douglas for the support and friendship he has shown to me both at school and in church. I would also like to express my endless gratitude to Dr. Volker Hegelheimer, who has provided me with immense opportunities, encouragement, and support, watching me grow as a doctoral student and researcher throughout all the eight years at Iowa State. He was always the first person I could turn to when I needed advice. Dr. Chapelle, Dr. Douglas, and Dr. Hegelheimer, I am extremely honored to have worked with all of you as well as have known you personally.

I would also like to extend my gratitude to the other committee members. Several statistics classes I have taken with Dr. Fred Lorenz greatly helped me develop a good understanding in statistics and other measurement models, many of which were utilized in this dissertation study. He is certainly one of the best teachers in statistics I have ever known. I send my sincere thanks to Dr. Greg Wilson as well, who kindly supported me throughout the entire process of my dissertation research.

My special thanks to Dr. Elena Cotos, who has been a nice friend of mine in the ALT program and also a wonderful boss for the past year. I am looking forward to working with you as a postdoctoral researcher for another year. My heartfelt thanks to peer students in the ALT program, especially all the Korean folks (Heesung, Hyejin, Hye-Won, Jooyoung, Yundeok, Moonyoung, and Yongkook), as well as my friends at St. Thomas Aquinas for
their truthful friendship and support throughout the years in Ames. I may not have survived those lonely years without them.

I also send my greatest thanks to my parents and my brother who have always supported me and kept trust in me. Especially Mom, I owe you much.
This dissertation study is intended to develop a test of productive grammatical writing ability in academic English as a supplement to the essay test of the English Placement Test at Iowa State University. In addition, it attempts to validate the score interpretation and use of the academic grammar test, using a validation framework adapted from both Kane’s (2006) interpretive/validity argument model and Bachman and Palmer’s (2010) Assessment Use Argument (AUA) model. Philosophically grounded in a constructivist-realistic approach to the interpretation of test scores, the test particularly aims to measure non-native English speaking students’ grammatical writing ability exerted in the academic setting. The entire processes of test development and evaluation are guided by an interpretive argument, which consists of seven inferences—Domain Description, Evaluation, Generalization, Extrapolation, Explanation, Utilization, and Ramification. Warrants for these seven inferences and underlying assumptions are identified through the interpretive argument, and the necessary types of backing evidence are also specified, to support these assumptions.

Using the quantitative and qualitative data collected in F12 and Sp13, the study investigates whether each of the seven inferences is sustained by sufficient backing evidence. Overall, the test appears to elicit the target construct of the test effectively from test takers, reflecting the linguistic characteristics of academic writing as a target language use (TLU) domain. Evidence pertaining to score interpretations also supports several claims the examinees’ performance on the academic grammar test corresponds to their ability to use the target construct in both testing and non-testing writing contexts, and satisfies theoretical expectations to some extent. Evidence for the validity of the test use for the intended purpose is sought from the interviews with three different stakeholders (i.e., the EPT
coordinator as a decision-maker, students as test takers, and instructors of ESL writing courses), by asking their perceptions about potential positive and negative impacts of the use of the academic grammar test as well as about the value of grammatical writing ability. Based on these pieces of supporting evidence, a validity argument for score interpretation and use of the academic grammar test is successfully established, despite only weak evidence for a few of the inferences.

Three issues related to the improvement of backing are discussed in regard to task design, the research in the acquisitional order of advanced grammatical features, and the revision of ESL writing curriculum and pedagogical approaches to grammar instruction. This dissertation concludes with a summary of the validity argument and a brief discussion of two logistic issues concerning test implementation, methodological limitations of the study, and suggestions for future research.
CHAPTER 1. INTRODUCTION

1.1. Mandates and Background

The present study is motivated by two internal mandates in relation to the administration of an in-house academic English test (called the English Placement Test (EPT)) at Iowa State University (ISU). All non-native English speaking students entering the university are required to meet this English requirement, and one of the common ways for them to do so is to complete ESL requirements based on a student’s EPT results. Consisting of three parts (i.e., essay writing, reading comprehension, and listening comprehension tests), the biggest concern with the EPT lies in the essay writing test composed of a single essay-type item. Assessing students’ writing skills, based on their performance on a single test item, jeopardizes both the reliability and the validity of the interpretation and use of the test scores, since inference about the students’ writing abilities is grounded in very limited information. Despite this fact, adding one or more essay prompts to the present test format is not a practical solution; this could lead to an increase in test-taking time to a considerable degree and, hence, would cause more intense fatigue from test taking than it does now, adversely affecting test performance in the end.

Another aspect of the mandate stems from a low interrater reliability among the raters of the EPT essays. Once the writing test is completed, the essays are sent to the raters, who also teach ESL courses in the Applied Linguistics program. The written responses are then scored on the basis of a rubric consisting of three proficiency levels, which correspond to the curriculum of ESL writing courses offered by the Department of English—namely, ENGL 101B (lower-level ESL writing), ENGL 101C/D (upper-level ESL writing), and pass. Two
or more raters complete a blind rating for each essay, and the grade most agreed upon among them is determined as the final grade of the relevant essay. One problem behind the scene is a low percentage of agreement among the raters, partly due to practical limitations in rater training and retraining. New raters are usually trained only once, immediately before the actual rating session, and follow-up calibrations among raters through a re-training have rarely occurred. High disagreement rates among individual raters have thus been observed in the EPT essay rating for years. Therefore, the mandate arising from this situation calls for the development of a writing-related test that is short, yet informative and reliable, and makes score interpretations about students’ academic writing ability and placement uses more valid. This is where the assessment of learners’ grammatical abilities can play an important role.

### 1.2. Issues from Theoretical and Empirical Perspectives

Instruction on grammatical ability or grammatical knowledge has often been neglected, since the introduction of the concept of communicative competence in language teaching. Focus on grammatical forms in language instruction is also criticized, based on the opinion of many teachers, that it does not concern how one might use language in a variety of contexts (Purpura, 2004), and communicative teaching concentrates on pushing students to use the language. While a legitimate criticism can be made of the focus of forms, grammatical knowledge of a language is still regarded as more important in real world situations than it is in many communicative language classrooms. For instance, not only linguistic abilities but also cognitive abilities of second language users, are sometimes judged by the grammatical accuracy and complexity of the language they produce in academic contexts. Second language users with underdeveloped grammatical abilities are often
stigmatized as cognitively incompetent as well because they fail to deliver their ideas in a clear, grammatically accurate manner. Researchers cannot publish their work in academic journals of their own disciplines if their writing is grammatically flawed, no matter how valuable their ideas or findings are. Hence, grammatical ability is still an important aspect in language learning. Taking an analogy of building a house, a strong grammatical ability serves as a solid basis and a structural frame of a house; how to use a language (i.e., pragmatic and sociocultural abilities in language use) is rather a matter of how to design and decorate the house. Severely flawed or weak frames will undermine the entire structure of the house, even if it is beautifully designed. Therefore, grammatical skills should be properly taught in academic English classrooms, and the development of a learner’s grammatical ability should also be properly assessed to help them stay on the right track in their second language (L2) development. In this regard, Purpura stresses the necessity of research on appropriate assessment of grammatical knowledge (2004, p. 4):

What is striking, however, in the long-standing debate on grammar and its role in language learning is the relative absence of discussion of how ‘best’ to assess grammatical knowledge or how to determine if grammatical knowledge has been acquired. Even with the sudden increase of research since the mid-1980s on the teaching and learning of grammar, there still remains a surprising lack of consensus on (1) what constitutes grammatical knowledge, (2) what type of assessment tasks might best allow teachers and testers to infer that grammatical knowledge has been acquired and (3) how to design tasks that elicit grammatical knowledge from students for some specific assessment purpose, while at the same time providing reliable and valid measures of
performance. In other words, there is a glaring lack of information available on how the assessment of grammatical ability might be carried out, and how the choices we make in the assessment of grammatical ability might influence the inferences we make about our students’ knowledge of grammar, the decisions we make on their behalf and their ultimate development.

Purpura’s remark above addresses issues on the construct definition of grammatical knowledge and the design of task types necessary for making appropriate inferences about language learners’ grammatical knowledge to be demonstrated in the target domain. First, a clear understanding of grammatical knowledge as a construct is essential in the development of a test. Yet, to the author’s knowledge, few attempts have been made in this regard. Second, traditional grammar tests tend to rely on multiple-choice questions, which evaluate students’ recognition of grammatical structures or errors rather than the actual ability to produce grammatical structures. Assessment of test performance on such recognition-based tasks will not provide sufficient information to aid test users in making inferences about test takers’ abilities to produce grammatically acceptable and developed sentences in writing within a certain target language use (TLU) domain.

The question of what constitutes grammatical knowledge cannot be properly answered without examination of second language acquisition (SLA) research. The value of linking SLA research with L2 assessment research is advocated by Bachman and Cohen (1998), who note the interfaces between SLA and language testing (LT) research are yet to receive further attention of experts from both fields. Particularly, the authors posed several questions that should be investigated in both SLA and LT research, which cover the aspects
of (1) the development of L2 ability, (2) roles of cognitive strategies and processes, and (3) roles of task characteristics in L2 performance. These all relate to the issues raised by Purpura (2004) above directly or indirectly. For example, findings on the developmental sequences of grammatical knowledge feed into the attempts to define the construct of grammatical knowledge. Findings regarding the question of whether the developmental sequences of grammatical knowledge are consistent across different language use contexts can further contribute to making informed predictions of language learners’ performance on grammar tests situated in various TLU domains. Research on the roles of cognitive strategies and processes in language learners’ acquisition, and use of grammatical knowledge can also provide insight into the understanding of what constitutes grammatical knowledge. In addition, it may further offer explanations about the effectiveness of test tasks in eliciting target responses from test takers by enabling the researchers to look into questions like which cognitive and metacognitive strategies the examinees take advantage and how they use those strategies when they complete the tasks.

Topics pertinent to the validity of score interpretations are of great interest in LT research. However, little research has made attempts to validate the uses of test scores and their consequences despite the fact that its importance has often been stressed since Messick (1989). Particularly, in the field of language testing, McNamara (2006) urged LT researchers and practitioners to conduct research to identify value implications underlying a language test and also to examine the impact the use of the test may bring about on different stakeholders of the test. In a similar vein, Bachman and Palmer (2010) proposed an argument-based validation framework called “Assessment Use Argument (AUA)” for score interpretations and uses of language assessments. As indicated by its name, the AUA framework is
designed in such a manner the test developer keeps in mind the intended decisions and consequences of the use of the test scores throughout the entire process of test development and validation. A possible explanation for the lack of research about the validation of test uses is that it requires extensive, often longitudinal, data collection both quantitatively and qualitatively. Tracking test takers’ paths after test administration must also be challenging. Nevertheless, the validation efforts on the aspects of test score uses should not be abandoned. It is especially important to complete when a new test is developed and administered for test users to make medium- or high-stakes decisions.

1.3. Goals of the Study

The present study is intended to respond to the aforementioned internal mandate and validation issues pertaining to test score interpretations and uses. Considering the first issue prompting the mandate of the limitations of a single-item essay test, in particular, this study will start with a development of a test of productive grammatical ability in English academic writing (an ‘academic grammar test’ for short) and attempt to validate its score interpretations and intended uses. In particular, it will investigate the following issues: (1) whether a short academic grammar test can be developed to provide sufficient evidence that helps test users make appropriate interpretations about writing ability and extrapolation of examinees’ test performance with respect to their grammatical writing performance in academic settings; (2) whether the use of the academic grammar test as a supplement to the essay writing in the EPT enables the test users to make appropriate decisions in ESL writing placements or exemptions; and (3) whether different parties of stakeholders (namely, test
takers, test users, and ESL instructors) could benefit from the decisions made by using the academic grammar test.

1.4. Significance of the Study

This study contributes to the validation research in the field of language assessment by making validation efforts not only with respect to inferences on score interpretations, but also on the use of the test and its consequences. The study also suggests a name for the inference pertaining to the claim of consequences of test use (i.e., Ramification Inference). Until now, little research within argument-based validity has been implemented in this regard despite the fact that the importance of validation efforts focusing on the intended and unintended consequences of test use has been stressed (Messick, 1989; McNamara, 2006). The validation research concerning the test use is conducted on a hypothetical scenario in this study to illustrate how even before the test is used, such efforts can shed light into the intended and unintended consequences of the test use and help test developers and evaluators identify room for improvement in the validity of the test use.

This study also extends an attempt to pursue an interface between the second language acquisition (SLA) research and the language assessment, by developing a language test on the basis of theoretical and empirical grounds coming from SLA. The findings obtained in the course of validation attempts will suggest further research issues for investigation in the field of SLA as well. Finally, ESL writing instructors and curriculum developers may benefit from the study in devising or modifying pedagogical approaches to the instruction of English grammar by learning about ESL students’ perceptions about the actual practices in the ESL writing classrooms.
1.5. Outline of the Dissertation

The rest of this dissertation is structured as follows. Chapter 2 begins with the review of literature in second language acquisition research, which investigated the natural order of acquisition of English grammatical features. Then, it is followed by findings from studies in language assessment intended to investigate whether test items developed on the basis of the same SLA literature could successfully measure and indicate test takers’ proficiency levels in written English. The last section of the chapter reviews philosophical backgrounds that underlie three different views on the concept of ‘validity’ in the field of assessment and overviews Messick’s (1989) concept of construct validity as a unitary concept. Primarily grounded in Kane’s (2006) interpretive/validity argument framework, Chapter 3 introduces the validation framework devised for this present study and provides an interpretive argument for the score interpretation and use of the academic grammar test. Research questions raised through the development of the interpretive argument are also presented in this chapter. Chapter 4, Methodology, describes (1) the considerations and procedures of the development of the test and its scoring rubric, (2) the data collection procedures undertaken to investigate the research questions, and (3) the analysis methods employed to answer the research questions. Next, the results of the investigation of the research questions are presented in Chapter 5. Chapter 6, the final chapter, first summarizes backing evidence found for establishment of the validity argument for the score interpretation and use of the academic grammar test for the intended ESL placement purpose. It then discusses three additional issues that need attention to improve the quality of backing evidence for the validity argument and addresses some logistical issues for the academic grammar test.
implementation. This chapter concludes with a summary of the study, its limitations, and suggestions for future research.
CHAPTER 2. LITERATURE REVIEW

To make appropriate interpretations from test scores, it is necessary to understand what theoretical and empirical implications they entail. The test development should thus be preceded by a review of the studies in second language acquisition (SLA) research, which can inform us of the nature of the second language (L2) development with respect to the acquisition of grammatical features of English as the target language. It is also important to examine whether test items targeting these grammatical features can serve as informative proxy indicators of the extent to which test takers have developed their ability to use them correctly in L2 writing. The first two sections of this chapter will review previous studies in these two aspects. The final section of the chapter will tap into the evolution of Messick’s (1989) construct validity as a unitary concept, from which the validation framework of the present study has emerged.

2.1. Common Acquisitional Orders for Grammar in SLA

Much of the SLA research conducted in the 1970s and 1980s tended to explore the question whether second language (L2) acquisition occurs following common developmental sequences across different language learners. Most of these studies provided support for the idea of universal developmental sequences in the acquisition of many grammatical features (e.g., Andersen, 1978; Bailey, Madden, & Krashen, 1974; Dulay & Burt, 1973, 1974; Larsen-Freeman, 1975; Stauble, 1984). Findings from these studies, in particular, suggested that even adult ESL/EFL learners were likely to acquire morphosyntactic/syntactic grammatical features of the English language in similar orders even though some variations among language learners exist. General patterns in the acquisitional order were also found in
functional aspects of English grammar. For example, Bardovi-Harlig (2000) and Bayley (1999) demonstrated that ESL learners followed a common sequence in the acquisition of tense and aspect in relation to verbal lexical aspects (i.e., *state*, *active*, *accomplishment* and *achievement*). That is, ESL learners started the acquisition of the past tense form with *achievement* verbs like ‘hit’ or ‘jump,’ which has features of [+dynamic, +telic, +punctual] in lexical aspect, before other verb classes. On the other hand, the acquisition of imperfect forms began with *state* verbs (e.g., ‘live’ and ‘think’) and then proceeded to active, accomplishment, and achievement in respective order. Bayley (1999) clearly illustrated step-wise acquisitions of English verb tense and aspect by Spanish-speaking children in relation to verb class.

Despite the implications of their findings in light of generative aspects of adult L2 learning, SLA research in universal acquisitional orders was criticized for several reasons. For example, Purpura (2004) raises a few questions on the findings of this generative approach to SLA. One of the vital drawbacks of the developmental order research is that a majority of the SLA research on common developmental sequences was limited in the scope of grammatical features of interest. Most of the studies focused on the acquisition of morphological or morphosyntactic features, which include articles, plural, possessive, copula *be*, 3rd person singular present –s, progressive (-ing), past tense (regular and irregular). In addition, the proficiency levels with which the order of the acquisition of these morphemes was found to correspond tended to range at lower levels of learner Interlanguage (IL) (e.g.,

---

1 Vendler (1967) proposed three features of lexical aspect in verbs. First, ‘dynamic’ is a feature of action. A verb retains a feature of [+telic] when the action implies a final goal. The feature [+punctual] indicates that the action takes place in an instant moment. Verbs can be classified into four classes, depending on the presence/lack of these individual features, which are *state* [-dynamic, -telic, -punctual], *activity* [+dynamic, -telic, -punctual], *accomplishment* [+dynamic, +telic, -punctual], and *achievement* [+dynamic, +telic, +punctual].
beginning through low/mid intermediate levels). It would thus not make sense to evaluate the grammatical ability of higher proficiency learners exclusively on the basis of these foregoing research findings.

Another apparent major issue concerning this generative SLA strand lies in the fact that much of the research on this topic grounded their conclusions in the data arising from the naturalist setting of language acquisition. Purpura (2004) comments on this issue that those studies did not take into account the effect of formal instruction in L2 acquisition that may also affect the sequence in the development of a target grammatical feature. However, this concern may be alleviated given the findings of other studies demonstrating that formal intervention is rarely effective or does not play a crucial role in L2 acquisition unless learners are cognitively ready to take in and process target grammatical features (Pienemann, 2007). For instance, Lightbown (1998) and Pienemann (1984) showed that language instruction was not effective when the instruction took place on grammatical features of non-contiguous developmental stages. Input Processability Hypothesis also suggests that formal L2 instruction will not be effective, if it does not take into account learners’ cognitive processing of input data in making connections between meaning and form (VanPatten, 2007).

Limited effects of instruction on L2 acquisition can be systematically explained within Pienemann’s Processability Theory (PT). Grounded in Levelt’s (1989) speech model and Lexical-Functional Grammar (Falk, 2001), this theoretical SLA framework hypothesizes a cognitive processability hierarchy of linguistic structures ranging from lemmata (i.e., canonical word forms) to phrasal structures to clausal and sentential levels of inter-phrasal structures. This model posits two hypotheses—namely, Topic Hypothesis and Lexical Hypothesis—to explain the acquisition of structures deviant from canonical word orders.
What this model aims to illustrate is “processing is incremental” (Pienemann, 2007, p. 137). In other words, learners cannot properly produce nor comprehend a complex linguistic structure unless they have acquired preceding relevant linguistic features and rules available at previous developmental stages. Hence, PT offers a solution to the problem that the findings of the naturalistic SLA research were limited to the acquisition of grammatical features of low developmental stages. Conceptual models, like the processability hierarchy, enable both SLA researchers and language testers to speculate the developmental stages of a variety of grammatical structures based on their presumable cognitive complexities and to test their hypotheses. PT can also be defended against a criticism that the naturalistic approaches to SLA concerns linguistic forms only, in that the Lexical-Functional Grammar (LFG), upon which PT has been developed, embraces semantic aspects of language in its theoretical framework. LFG hypothesizes a mapping between the semantic/argument and the functional structures (a-to-f mapping) as well as a mapping between the constituent and the functional structures (c-to-f mapping) (Falk, 2001). PT therefore deals with grammatical meaning as well as grammatical form in Purpura’s terms (2004, p. 78). Although the acquisition of pragmatic knowledge is not addressed in this theory, PT successfully explains the mechanism underlying the acquisition of grammatical knowledge—a vital instrument in academic writing.

2.2. Developmental Stages Demonstrated in Language Assessment

Not only in SLA research, but also in language assessment research, has evidence of common developmental stages been found, although only a limited number of studies have been conducted in language testing in this regard. Norris (1996, cited in Norris (2005))
would be one of the earliest language assessment research studies conducted utilizing common developmental sequences. It reported that the degrees of accuracy in the production of developmentally sequenced German word order rules corresponded with learners’ oral proficiency levels determined by the ACTFL Proficiency Guidelines. As predicted in PT, Norris also found that, unlike advanced-level learners, Intermediate-Low and lower proficiency learners failed to produce any evidence of the acquisition of the German word order at the predicted highest developmental stage.

Evidence of the association between proficiency levels and the degree of accuracy in producing the grammatical features at different predicted developmental stages is also reported with respect to English as a second language (ESL) in the written mode. For instance, being grounded in Pienemann’s (1998) Processability Theory, Norris (2005) developed grammar test items that tapped six different syntactic developmental sequences for a quick diagnostic purpose of a commercial online ESL learning module. After piloting a number of test items in several task types (e.g., gap filling, strict jumbled word order tasks and its variation, and cloze tasks) with learners from five different proficiency levels, he adopted the jumbled word order and its variation (‘choose-words-and-put-them-in-order’) tasks for the second pilot tests. Norris then created three 12-item test sets for each of two sampling approaches (one concerning the target population and the other concerning the effective discrimination between Intermediate and Advanced levels) by taking well-performing items that represent each of the target grammatical features of interest. Scoring based on a target-like accuracy criterion successfully allowed each of the test sets to discriminate three broad proficiency levels (i.e., Beginner, Intermediate, and Advanced) with acceptable reliabilities (e.g., alpha coefficients above .7), although the distinctions between
Intermediate and Advanced were found less stable than those between Intermediate and Beginner levels.

Similar results are also reported in Chapelle, Chung, Hegelheimer, Pendar, and Xu (2010), although they took a few different approaches from Norris (2005) in designing and scoring test items. First, Chapelle et al. developed test items targeting not only word-order based syntactic grammatical features but also morphosyntactic and functional features (like article, tense, aspect, and voice). They also classified these grammatical features into three broad, predicted developmental stages, based on findings in the aforementioned SLA research and Norris (2005). In their classification, morphosyntactic and functional features were stipulated relevant to the Beginner and Intermediate levels, whereas syntactic features belonged to the Advanced level. Chapelle et al.’s (2010) grammar test also differed in its scoring system. Unlike Norris’ (2005) scoring scheme based on target-like accuracy, Chapelle et al. (2010) utilized a three-level scoring scheme to take into account both aspects of emergence and mastery present in students’ responses. The second test set of their study consisted of 15 restricted construct-response items and two short free writing items, and demonstrated acceptable reliability (i.e., Cronbach’s alpha = .82). Furthermore, in general, the items’ difficulty indices had a statistically significant strong association with the items’ expected developmental stages. More important, this study illustrated distinctive patterns in examinee performance on the test that clearly corresponded with the students’ proficiency levels. Given that the examinees’ proficiency levels ranged from the upper-intermediate level to the near-native level in Chapelle et al. (2010), this finding implies a promising potential of developmental sequences-based measures in language assessment even among higher proficiency-level learners.
A follow-up study conducted by Chung (2012) supports this implication. With the construct of productive grammatical ability defined within an academic writing setting, she employed restricted constructed-response item types and adopted Chapelle et al.’s (2010) three-point polytomous rating system. Her findings demonstrated almost identical patterns to those reported by Chapelle et al. (2010) both in test item behaviors and in distinction of learners from different proficiency levels. One difference is that the mean item difficulty indices of Developmental Stage 3 items was not significantly different from that for Stage 2 items, which, in turn, led to a moderate correlation between the item difficulty indices and the predicted developmental stages. However, instead of rejecting the potential of developmentally sequenced, grammatical features as useful proficiency indicators, Chung attributed this idiosyncratic phenomenon to restrictive item design employed to narrow the number of possible unexpected student responses to test items. Such constrictive item design appeared to lend itself to the result that test difficulty levels were confounded by the test developer’s restrictions placed on those items. Her study also suggests that further research is needed in regard to the expected developmental stages of lexical or lexicosyntactic features (such as prepositions in collocations or verb phrases).

Despite minor variations in findings, the results of the studies reviewed thus far imply that research on developmental sequences of grammatical features can certainly feed into a reliable measurement of L2 learners’ language abilities. However, grammar-focused tests are often criticized for their limited scope of interest in light of a rather complex system of language learning and use, such as the language use model suggested by Bachman and Palmer (2010). Therefore, the interpretation and use of grammar test scores for assessment purposes need a great deal of supportive evidence to warrant validity.
2.3. Validation Frameworks

In educational measurement, validity has long been conceived in terms of whether a test “measure[s] what it is supposed to measure” (Valette, 1967, p. 30; cited in Chapelle, 2012, p. 22). However, the legitimacy of this question on validity depends on the ontological view regarding the relationship between the test and what it actually measures. As such, this section will start with a brief discussion about different philosophical stances concerning the interpretation of test and non-test performances. The discussion will then lead to Messick’s (1989) construct validity as a unitary concept.

2.3.1. Ontological Conceptions of Validity

The answer to the foregoing question is determined by language testing experts’ ontological perspectives on the issue about ‘what is being measured’: realist vs. constructivist vs. constructivist-realist perspectives (Messick, 1989). First of all, the realist view assumes that what is being measured (i.e., a trait) actually exists in the real world and that the trait is consistent regardless of the characteristics of the context in which it is exhibited. In other words, it is traits and only traits that govern the test performance. From the realist perspective, the test score is thus a direct indicator of the test taker’s trait ability; there is a direct causal relationship between trait and test/non-test behavior. A good test is a test that makes a good observation of examinees’ trait skills. Hence, validity is the property of the test itself; the test validity is “a function of truth” in a sense that what matters in assessment is whether a test captures the true trait as existential reality (Borboom, Cramer, Kievit, Scholten, & Franić, 2009, p. 138).
However, unlike the realist assumption of trait consistency, test takers’ performance may vary depending on various contexts. For example, a test taker’s vocabulary size is found to be different across the receptive and productive language skills, although receptive vocabulary size may be a good indicator of productive vocabulary size (Webb, 2008). An examinee’s speaking ability may also be exhibited very distinctively by the interviewer’s personal attributes or experience as an oral test interviewer (Brown, 2003). These findings allow us to cast doubt on whether an examinee’s trait is consistently demonstrated regardless of the circumstances in which the examinee’s performance is situated. The objection to the realist perspective on ‘what is being measured,’ in turn, brought about the constructivist perspective.

The constructivist perspective assumes that a construct as what is being measured is only defined in terms of its nomological network or a network of relevant theoretical laws (Cronbach & Meehl, 1955). Messick (1989) explains that constructs serve an instrumental purpose. What is measured through an assessment can be understood with respect to the theoretical laws pertinent to the test performance. An examinee’s performance in a certain target domain can thus only be inferred by drawing upon the evidence observed in relation to the nomological network. From this ontological perspective, a logical construction of the entity being inferred through the observation is critical for the validation of a test. Therefore, according to the constructivist view, validity is not an inherent property of the test per se; it is rather a matter of establishing a plausible argument on the basis of solid evidence that pertains to an underlying theoretical network of the construct.

A third perspective on the issue, namely a constructivist-realist approach to the interpretation of examinee behavior, takes both of the preceding ontological stances to some
extent. It assumes the existence of a trait as a target attribute, which cannot be directly measured by test performance but can be understood through a logical construction based on observed evidence. By assuming the existence of a trait, this approach enables assessment experts to explain two apparent conflicting phenomena of (1) oscillation of examinee performance across various contexts and (2) consistency in examinee behavior in different settings or circumstances. Take Webb’s (2008) findings of receptive and productive vocabulary sizes as an example again. His finding of different vocabulary sizes depending on the receptive and productive aspects of language use reflects the oscillation of learner performance in the use of vocabulary knowledge across different language modes. A network of theoretical laws assumed with respect to the relation between language mode and learner performance can help explain this phenomenon. On the other hand, another finding on a proportional relationship between receptive and productive vocabulary sizes implies that there exists a certain trait of vocabulary knowledge that affects the consistency of learner performance across different language modes to some extent. The constructivist-realistic view therefore attempts to understand traits through the interpretation of relevant constructs that encompasses “not only nomological networks but also causal models and other ways of representing constructs and their measures” (Messick, 1989, p. 29).

Both constructivist and constructivist-realistic perspectives on test/non-test behaviors dictate that it is the interpretation and use of test scores rather than test scores that assessment experts should validate, since a test cannot measure the intended trait or construct in a direct fashion. The adoption of either of these two approaches in the mainstream educational assessment has led to attempts to identify evidence that addresses multiple aspects of the validity of score interpretations and uses as illustrated in Sireci (2009, p. 26). These aspects
of validity have often been addressed in terms of content validity, criterion-related validity (or concurrent/predictive validities), and face validity, which Messick integrated together under an overarching concept of ‘construct validity’ (1989, p. 42).

2.3.2. Construct Validity as a Unitary Concept

Messick’s (1988, 1989) proposal of construct validity as a unitary concept is driven by a pondering upon limitations that the aforementioned validities are flawed when they stand alone to support the validity of a score interpretation and use. For example, content validity is sustained by the test content’s relevance to, and its representativeness of, the target domain. However, it is by and large determined by expert judgment, which may fail to sufficiently satisfy the adequacy of two aspects of content validity—content relevance and content coverage (Bachman, 1990). Both Messick (1989) and Bachman (1990) particularly point out the limitation of content relevance that it does not provide evidence for an inference regarding how examinees’ performance on the test relates to their performance in non-testing setting. Similarly, the structural consistencies of test responses in relation to other criterion measures do not necessarily indicate that these tests tap the same construct (for internal validity) or irrelevant constructs (for external validity) in regard to intended score interpretations. Messick (1989) addresses this issue of complementary roles played by content validity and criterion-related validity under a unitary concept of construct validity as follows:

It is clear that content-related evidence cannot stand alone, but we need to examine how it functions in concert with construct-related evidence in a unified validity framework. The critical role of content in test validity
becomes clarified by viewing it in conjunction with the structure of response consistencies, both internally to the test and externally with other measures, as part of the overarching concept of construct validity (p. 42).

Face validity in essence concerns political aspects of test score interpretations and uses (McNamara, 2006). Messick (1989) attends to this political facet of validity in terms of value implications and social consequences. Test scores can be understood with value-laden interpretations, which often makes the score interpretations much more complicated because there may be many different perspectives involved in the conception of a value like ‘justice’ or ‘equity.’ Different value implications tend to affect decision-making on educational policies as well. Even within the SLA field, questions are often raised like whether or not native-like pronunciation is important in L2 learning, whether or not L2 grammar should be explicitly instructed, and so forth. One may find from these examples the theoretical framework of the target construct affects such value-laden decisions on language assessment as well as on language instruction. Critical testing studies also reveal that social values associated with power relations affect the uses of tests and that tests are often used as a means to impose policies (Shohamy, 2001).

As for social consequences of test uses, McNamara (2006) distinguishes two perspectives on the issue: ethical testing practice and critical testing practice. Most language practitioners are concerned with ethical testing practice. According to McNamara, ethical testing practice covers three main areas of responsibility, which are accountability (i.e., responsibility to stakeholders like test takers and test users), washback (i.e., influence on classroom teaching), and test impact (i.e., impact on broad social and educational world) (2006, pp. 43-44). He also introduces two different views within the ethical testing practice
perspective: ‘social responsibility view’ and ‘professional responsibility view’ in the same
discussion on social consequences of test uses. Language testing practitioners from the
‘social responsibility view’ would play a more proactive role than those of the latter view,
such that the former would actively monitor the consequences of a test implementation to
avoid the misuse of the test, while the latter would provide codes of practice for language
testers to help produce quality tests.

As illustrated in the discussions so far, content validity, criterion-related validity, and
face validity each merely taps a single facet of Messick’s (1988, 1989) overarching concept
of construct validity. Messick’s (1989) definition of validity hence delivers a gist of the
foregoing discussions in one sentence: “Validity is an integrated evaluative judgment of the
degree to which empirical evidence and theoretical rationales support the adequacy and
appropriateness of inferences and actions based on test scores or other modes of assessment
[sic]” (p. 13).

Messick’s seminal discussion on construct validity as a unitary concept has
enlightened language testing experts about multi-faceted characteristics of validity.
However, his definition of validity has also frustrated most of them until recently because he
did not offer any practical suggestion that could provide an effective methodological conduit
for the validation of intended score interpretations and uses. It is Kane (1992, 2004, 2006)
and his colleagues (Kane, Crooks, & Cohen, 1999) who first suggested a methodological
framework of validation. Following Toulmin’s (2003) argument model, Kane’s approach to
validation has enabled testing experts and practitioners to address the multifacetedness of
score meanings and uses by making inferences based on evidence pertaining to each aspect
of construct validity. Kane’s suggestion also inspired Bachman and Palmer (2010) to
develop a similar validation scheme called Assessment Use Argument (AUA). The present study will also address the validity of the intended score interpretations and uses of the academic grammar test following Kane’s approach to validation.

2.4. Chapter Summary

In this chapter, the literature was reviewed from three perspectives. The first section of the chapter reviewed findings of SLA research, which focused on the natural order of acquisition of English grammatical features. The Processability Theory was also introduced as a primary theoretical background for the design of the academic grammar test. The second section reviewed preceding studies in language assessment, which developed a grammar test drawing upon the findings of the aforementioned SLA research as an empirical background of the present study. In the final chapter, three ontological conceptions of validity were briefly reviewed, advocating the constructivist-realist perspective of validity as the philosophical background for this study. Finally, this chapter concludes with a review of Messick’s (1989) construct validity as a unitary concept, which brought about the development of two validation frameworks grounded in Toulmin’s argument model (Kane, 2006; Bachman & Palmer, 2010).
CHAPTER 3. INTERPRETIVE ARGUMENT

The establishment of an interpretive argument for test score interpretation and use is essential in test development in that it allows the test developer to identify important elements for careful consideration in test design (Kane, 2006). This section will thus outline an interpretive argument of the academic grammar test and the research questions that guide the collection of backing. The results of the research will be constructed into a validity argument for the score interpretations and uses of the academic grammar test.

3.1. Interpretive Argument Framework

Figure 3.1 depicts the conceptual framework of the interpretive/validity argument for the academic grammar test. In the diagram, each box represents a respective claim (except for ‘Domain’, which serves as a launching grounds for the subsequent claim) to be associated with test development, score interpretations, and uses. As illustrated at the bottom of the figure, the claim in the lower box (e.g., ‘Domain’) serves as grounds for the following claim (e.g., ‘Observation’), and the transition from the ground to the claim is permitted through the establishment of the respective inference (e.g., ‘Domain Description’). Each inference is legitimized by a warrant—namely, “a generally held principle, rule of thumb, or established procedure” (Chapelle, Enright, & Jamieson, 2010, p. 6). Warrants are legitimate only to the extent that their underlying assumptions are supported. Therefore, the interpretive argument must state the warrants and assumptions associated with each inference, and the research should provide backing for the assumptions underlying each warrant in the interpretive argument.
The interpretive argument of the present study was developed detailing the claims, inferences, warrants, and assumptions underlying test interpretation and use of the academic grammar test. The format primarily adapts the argument models proposed by Kane (2006) and used by Chapelle et al. (2008, 2010). Bachman and Palmer’s (2010) Assessment Use Argument (AUA) model was also taken into consideration for the development of the interpretive argument. Whereas they share the common grounds of Toulmin’s framework of argument (Toulmin, 2003) for their validation models, Kane-style models consists of a series of grounds and claims structured in a finer-grained step-wise manner than the AUA model. The former thus makes the relevant aspect of score interpretation at each validation stage much clearer than the latter. In turn, it enables test developers and evaluators to delineate and address the needs for validation research in a straightforward fashion.

Meanwhile, the AUA includes a bidirectional view in the argument structure, which makes it suitable for conceptualizing test development on the one hand and score interpretation and use on the other. This bidirectional framework is illustrated in Figure 3.1. In an effort to ensure that the impact of the test score interpretation and use is beneficial to stakeholders, the interpretive argument starts with intended consequences to be brought about by the intended test use. The intended consequences are then borne in mind through the development of the interpretive argument down to the stage of test task designs. This direction illustrates that the intended test consequences are intended as a basis for decisions about the claims along the inferential chain. This approach is intended to make coherent the entire process of test development, score interpretation, and score use. The detailed interpretive argument of the academic grammar test is delineated in Table 3.1 with necessary
backing for each assumption underlying the respective warrant listed in the rightmost column.

**Figure 3.1.** Interpretive/validity argument framework for the academic grammar test
### Table 3.1. Interpretive Argument of the Academic Grammar Test

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Assumptions</th>
<th>Backing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ramification</strong></td>
<td>W7: Intended consequences of the use of the academic grammar test for making decisions on students’ placements in or exemptions from ESL writing courses are beneficial to groups of stakeholders.</td>
<td>• ESL instructors benefit from the decisions made on ESL placements through the use of the scores from the academic grammar test.</td>
<td>• ESL writing instructors appreciate the value of grammatical writing ability in academic writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Examinee students benefit from the decisions made on ESL placements by the use of the academic grammar test.</td>
<td>• ESL writing instructors benefit from the information on their students’ grammatical proficiency levels in lesson planning or course instruction.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• False positive decisions in exemptions from ESL writing placement are avoided or diminished.</td>
<td>• Students appreciate the benefits of the instruction on grammatical writing ability.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Students appreciate the importance of grammatical ability in academic writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Negative ramifications of the decisions made on students’ ESL placements do not outweigh the positive ramifications.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Use of the test scores on the academic grammar test identifies the number of false positive exemptions from ESL writing placement.</td>
</tr>
<tr>
<td><strong>Utilization</strong></td>
<td>W6: Equitable decisions are made with respect to examinee students’ placement in ESL writing by using the test results along with the results of the writing test.</td>
<td>• Scores on the academic grammar test are used in making ESL placement decisions in conjunction with EPT writing results.</td>
<td>• Decision-maker refers to the grammar scores in making decisions on ESL placements/exemptions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Educational values concerning the instruction of grammar in academic writing are carefully considered in making decisions on ESL placements or exemptions.</td>
<td>• Decision-maker appreciates the value of grammatical ability in academic writing in making decisions on ESL placements/exemptions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Decision made on ESL writing placement/waiver is fair and impartial for every individual student.</td>
<td>• Use of the academic grammar test scores provides an objective ground for the decisions on individual students’ ESL placements/exemptions.</td>
</tr>
<tr>
<td><strong>Explanation</strong></td>
<td>W5: Target scores estimated by the grammar test are attributed to the theoretical construct definition underlying the construct of productive grammatical writing ability in English.</td>
<td>• Examinee performance on test task reflects the cognitive difficulty levels of the target grammatical features in their acquisition.</td>
<td>• Item facility indices in general inversely correspond with the expected developmental stages of their target grammatical features.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Observables of sub-construct target grammatical features are meaningfully associated with the construct of the productive grammatical writing ability in English.</td>
<td>• The averaged scores of the test items tapping into the same grammatical features have significant factor loadings on the test construct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Construct of productive grammatical writing ability in the academic setting is consistently exerted in other criterion measures engaging the same construct, even though learner performances may be affected by other factors in these measures.</td>
<td>• Examinees’ performances on the academic grammar test correlate with their performances on other criterion measures (e.g., EPT writing or TOEFL writing) at moderate degrees.</td>
</tr>
<tr>
<td>Inference</td>
<td>Warrant</td>
<td>Assumptions</td>
<td>Backing</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Extrapolation     | W4: Examinees’ expected scores estimated by the observation of the academic grammar test represent their target scores manifested in the target language use (TLU) domain. | • Examinees performances on the test are consistent with their general proficiency levels determined by another criterion.  
• Examinees’ estimated expected scores are reflective of their performances in other spontaneous writing situations within the academic context. | • Examinee performance at different ESL proficiency levels is significantly distinctive from each other.  
• Examinees’ performances on the academic grammar test are in accordance with their performances in other spontaneous writing activities in actual classroom setting. |
| Generalization    | W3: Observed scores are stable estimates of expected scores in the universal domain of test tasks. | • Test scores have an acceptable level of reliability.  
• Examinees’ performance on the academic grammar test is consistent regardless of test delivery format. | • Cronbach’s alpha coefficient is .7 or higher.  
• Reliability of the test does not vary significantly across different samples from the same population.  
• Scores from the computer-based and paper-based academic grammar test are comparable. |
| Evaluation        | W2: Observations of examinee performance are evaluated to provide observed scores informative of examinees’ task performance. | • Examinees’ performances on the test are not affected by students’ background knowledge relevant to the content of the texts of test tasks or lack thereof.  
• Examinee’s performances on the test are properly evaluated in such a way that generates observed scores reflective of their relevant abilities to produce grammatical sentences in academic writing. | • Completion of test tasks does not require examinees to have content knowledge.  
• Tasks of the academic grammar test target examinees’ grammatical writing ability and properly elicit it through their performance on the test.  
• Scoring rubric is created in such a way to identify different degrees of language acquisition (e.g., emergence vs. mastery) of target grammatical features.  
• Individual responses are given proper scores.  
• Scores discriminate among examinees effectively.  
• Examinee performance on the academic grammar test is not significantly affected by construct-irrelevant factors like gender. |
| Domain Description| W1: Test tasks contain the language of the target domain of academic English. | • Linguistic nature of the tasks in the academic grammar test represents those of academic English. | • Texts used in test tasks are adapted from authentic academic texts.  
• Linguistic features of test tasks reflect those of authentic academic texts. |
3.2. Research Questions

The research questions of the present study are raised to seek backing for each assumption identified in Table 3.1. In this study, the findings from the investigation on each of the following research questions will be evaluated in terms of whether they provide backing for the respective assumption:

(1) Do the test tasks on the academic grammar test require any field-specific knowledge of test takers? Does the test properly elicit evidence of examinees’ grammatical ability? (Warrant 2)

(2) Can students’ performances on the academic grammar test be properly interpreted and extrapolated as reflecting their typical grammatical abilities in academic writing as the target domain?

   a. Do the academic grammar test scores effectively discriminate among students of varying proficiency levels? (Warrant 2)

   b. Does gender, as a construct-irrelevant factor, significantly affect examinees’ performances on the academic grammar test? (Warrant 2)

   c. Are the assessment results generalizable across the universe of assessment tasks to an acceptable extent? In other words, is Cronbach’s alpha coefficient equal to or greater than .7? (Warrant 3)

   d. Does the academic grammar test demonstrate similar reliability across different samples from the same population? (Warrant 3)
e. Are scores from the computer-based academic test comparable to those from the conventional paper-and-pencil test? (Warrant 3)

f. Do students at different proficiency levels perform distinctively from each other on the academic grammar test? (Warrant 4)

g. Does students’ performance on the academic grammar test correspond to their performance in a non-testing setting? (Warrant 4)

h. Does the performance of the academic grammar test items reflect the theoretical assumptions that underlie their development? (Warrant 5)

i. Do the selected target grammatical features have significant associations with the construct of productive grammatical writing ability? (Warrant 5)

j. Does examinees’ performance on the academic grammar test correlate with their performance in other writing tests that engage grammatical ability at moderate degrees? (Warrant 5)

(3) Does the decision-maker appreciate the value of grammatical ability in academic writing when making decisions on ESL placements/exemptions? (Warrant 6)

(4) Could academic grammar test scores serve as a tangible ground for decisions on individual students’ ESL placements/exemptions? (Warrant 6)

(5) To what extent does the use of the academic grammar test help test users identify false positive decisions resulting from decisions on ESL writing placements/exemptions made exclusively on the basis of examinee performance on the EPT essay test? (Warrant 7)
(6) If the scores on the academic grammar test were used to make decisions on ESL placement, how would students perceive its impact on them with respect to (a) advantages and disadvantages of additional ESL instruction and (b) the relationship between grammatical writing ability and academic writing? (Warrant 7)

(7) How much do ESL instructors appreciate the value of grammatical ability in academic writing? To what extent could ESL instructors benefit from feedback on their students’ grammatical abilities provided by the use of the academic grammar test? (Warrant 7)

If the investigations of these research questions produce backing evidence that supports the assumptions of Warrants 1 through 7, this will allow elaboration of a validity argument for the development, score interpretation, and use of the academic grammar test for making ESL placement/exemption decisions.

3.3. Chapter Summary

This chapter started with descriptions of the validation framework chosen for the present study, adapted from both Kane’s (2006) interpretive/validity argument and Bachman and Palmer’s (2010) Assessment Use Argument models. The validation framework also introduced a term for the inference pertaining to the claim of the consequence—namely, Ramification Inference. Table 3.1 provides the inferences of the interpretive argument for the score interpretation and use of the academic grammar test, along with their associated warrants and underlying assumptions. Backing evidence necessary to support each assumption was also identified and is listed Table 3.1. Seven major research questions with
subordinate questions were then devised to guide investigations to collect backing evidence throughout the study.
CHAPTER 4. METHODOLOGY

This study adopted a mixed methods research approach to address the research questions that cover various types of validation research. In other words, this study purports to corroborate the validity of the intended score interpretation and use of the academic grammar test on the basis of evidence resulting from the investigation of the research questions. Among a variety of mixed methods approaches to research, this study, in particular, adopted a sequential design as different types of data were collected and analyzed in three different phases, from which meta-inferences of the results (i.e., the validity argument) would be obtained (Cameron, 2009). The adopted design can also be considered as an embedded design, which provides a primary role to quantitative data in its interpretations (Creswell & Plano Clark, 2011). The quantitative and qualitative strands will interact with each other to some extent, as some findings from quantitative analyses will provide a basis for the formation of survey and interview questions. While most of the research questions will be addressed by the quantitative data collection and analysis, interpretations of the qualitative data will also contribute to the support of the assumptions underlying the warrants of the Evaluation, Extrapolation, and Ramification Inferences in Figure 3.1.

The overall sequential embedded mixed methods design adopted by the present study is structured as Figure 4.1 illustrates. Data collection was completed in three phases as shown in Figure 4.2. The first two phases centered on the quantitative aspects of the validation issues, which arose through the construction of the interpretive argument (Table 3.1). Additionally, the final phase attended to both quantitative and qualitative aspects of the research.
Figure 4.1. Sequential embedded mixed methods design for the validation study of the academic grammar test.
Figure 4.2. Multiphase validation procedures of the academic grammar test
4.1. Test Materials

It is necessary to develop test materials of good quality to assess test takers’ abilities in the target language skills. This section describes the considerations and procedures of item development with respect to the selection of target grammatical features and the design of test tasks.

4.1.1. Item Design

4.1.1.1. Selection of target grammatical features

To sustain the warrant of the Domain Description Inference in the interpretive argument, it is mandated to design test items reflective of the linguistic characteristics of written academic English as stated in the assumption underlying the warrant. One aspect of language closely related to the linguistic characteristics of the target construct constitutes grammatical features commonly used in the academic contexts as the target setting. With regard to this aspect, thirteen grammatical features were first selected as candidate target features of test items primarily drawing upon the findings reported in Chapelle, Chung et al. (2010) and Chung (2012). In their studies, all but one item targeting these features demonstrated effective measurement of test takers’ productive grammatical abilities with item discrimination indices measured in terms of item-total correlation coefficients ranging between .30 and .62, which are acceptable according to the rule of thumb suggested in Carr (2011).

Another reference used to make the decision of target grammatical features reflective of the linguistic features of academic texts was Longman Grammar of Spoken and Written English (LGSWE) (Biber, Johansson, Leech, Conrad, & Finegan, 1999), a collection of
corpus-based extensive descriptions of English grammar. In LGSWE, the authors describe structural and usage patterns of English grammar found in the analysis of a large collection of texts across four major registers they believe are “highly productive varieties of the language and (...) different enough from one another to represent of a wide range of variation” (pp. 15-16). Academic prose is one of these major registers and the descriptions provided in their work were helpful in the selection of target grammatical features for the present research. For example, the grammatical features ‘modal verb + present perfect’ (e.g., should have done) was one of the target features in Chapelle et al. (2010) and Chung (2012), included with an intent to assess ESL learners’ acquisition of an advanced level of Aspect in conjunction with the function of modal verbs. However, it was removed from the target construct features in the present study, given the finding in LGSWE that it is not as frequently used in academic prose as in the other registers like conversation, fiction, and newspaper language. Rather, it illustrates that a form of ‘modal + passive’ (e.g., can be defined) is commonly used in academic prose. Since passive is already included in the target grammatical features of the academic grammar test, the feature ‘modal + present perfect’ was dropped from the inventory of the target features to not only reflect its relative paucity in academic prose in the test, but also to avoid a heavy reliance on a particular grammatical feature over others in evaluating learners’ productive grammatical ability in academic ESL writing.

On the other hand, some very commonly used features like ‘make + NP + Adjective’ were excluded from the inventory of potential target grammatical features of the test, because the use of such features was more pertinent to vocabulary knowledge—knowledge of collocations, precisely speaking—than to the structural aspect of the grammar. Although
many “aspects of English grammar have (…) systematic associations with lexical classes” (Biber et al., 1999, p. 13), the academic grammar test was not intended to measure vocabulary knowledge unless the item particularly pertains to the use of a preposition (e.g., engaged in making). Being grounded in the Processability Theory, the test was rather designed to assess the extent to which a learner’s interlanguage (IL) grammar has developed to produce English syntactic structures of varying complexity commonly found in academic prose with respect to his/her cognitive readiness for second language acquisition.

These considerations on the grammatical characteristics of the academic writing genre brought about minor changes in the definition of the construct of grammatical knowledge in comparison to that in Chapelle, Chung et al. (2010). The construct of the present study lies within the domain of academic language, whereas their study focuses on the grammatical ability in a general setting. The lists of sub-construct grammatical features also slightly differ between the two studies, although the grammatical ability as the target trait remains identical. As a result, nine grammatical features among the 13 candidate features were finally chosen as the target sub-constructs of individual items of the academic grammar test. These target features are articles, prepositions, perfect, passive, gerund, wh-nominal clauses in conjunction with cancellation of SV inversion, relative clauses, SV inversion triggered by a sentence-initial negation, subjunctive (untrue) conditionals, and participle phrases. Most of these features are intended to measure ESL learners’ advanced abilities to produce complex syntactic structures in English because the target population of the test takers are students of higher education and are assumed to have a proficiency level equivalent to the Intermediate High or above of the ACTFL Proficiency Guidelines 2012 – Writing (American Council on the Teaching of Foreign Languages, 2012).
4.1.1.2. Task design

The test tasks of the academic grammar test are classified into three types. The first type is ‘filling in the blank by adding a new word, if necessary.’ In this type of task, each sentence is provided with a blank space and test takers are supposed to add a word to make the sentence complete, if they believe it necessary. This task is designed to evaluate the examinees’ knowledge of articles and prepositions in English, which are often dropped, when needed, or incorrectly used by ESL learners. An example of the gap-filling task is illustrated below in Figure 4.3 with the answer provided within a parenthesis.

![Figure 4.3. Example of a gap-filling task without key words](image)

The second task type is ‘use given words, change word forms if necessary, and add minimally required new words also, if necessary,’ shown in Figure 4.4. Unlike in the tests of Chapelle, Chung et al. (2010), the tasks of ‘changing word forms’ and ‘adding (minimally required) new words’ are combined into one category in the academic grammar test to minimize clues from the directions that test takers may utilize to determine which grammatical features should be used to complete the given sentence in each test item. It will hence allow examinees’ responses to reflect the status of their grammatical development in IL more accurately than the prototype tasks. The target grammatical features of this task type
include those pertaining to morphosyntactic categories of English grammar like tense use, passive, preposition, perfect, and conditional.

**Figure 4.4.** Example of a gap-filling task with key words provided

The final task type is ‘rearranging jumbled words correctly.’ In this task, examinees are not asked to change word forms nor add new words; instead, they should place the given words into the correct order to make a complete sentence. This task type is intended to assess test takers’ abilities to produce cognitively complex syntactic structures of English, such as relative clauses, one or more indirect wh-question (realized in nominal wh-clauses), and participle phrases as reduced structures of relative clauses. Figure 4.5 illustrates an example of the jumbled-word-order task. A detailed specification of the academic grammar test can be found in Appendix A.

**Figure 4.5.** Example of a jumbled-word-order task
4.1.1.3. Writing test items

In the academic grammar test, two test items each addressed seven of the ten target grammatical features. On the other hand, examinees’ acquisition of prepositions was measured by four items, two of them mainly focusing on the construction of Passive. Another test item was designed to assess more or less ‘pure’ knowledge of the preposition associated with a lexical item \((to)\, release\). The other one was created to assess the knowledge of a frequent collocation \((be)\, engage(d)\, in\) with an intent to tap into a property of prepositions that take a nominal complement. Due to limited space in the test booklet, grammatical features of ‘negation followed by SV inversion’ and ‘gerund’ were tapped in a single item. The test included 16 questions with four intended to measure two grammatical features each. Among the 16 questions, three test items took the form of a ‘fill-in-the-gap’ task, and eight items took the second task type. The remaining items took the format of jumbled word order.

Sentences in the test items were adapted from actual college textbooks at an introductory level and the LGSWE to reflect the formality of the academic contents in the test materials as well as the register. However, the test items were designed in such a way that test takers’ performances would not be affected by their background knowledge on the topic or lack thereof. The academic disciplines from which the texts were adapted include sociology, communications, biology, and linguistics. More texts from liberal arts and human sciences were referred to than those from sciences, as the latter tends to require technical or area-specific knowledge for understanding. See Appendix B for the actual materials of the academic grammar test.
4.1.1.4. Expected developmental stages of the target grammatical features

Given the reasoning that the cognitive complexity of a grammar test item would be reflected in item difficulty indices, drawing upon the Processability Theory, test items were first classified into four groups of expected developmental stages—roughly speaking, beginner (Stage 1), lower-intermediate (Stage 2), upper-intermediate (Stage 3), and advanced (Stage 4) levels. This classification was made primarily on the basis of the findings of SLA literature on an acquisitional order of grammatical features (Sections 2.1 and 2.2). Meanwhile, the researcher, drawing upon her judgment recalling her experiences as an ESL teacher as well as an ESL learner, tentatively determined a few items expected in the developmental stages. For example, the expected acquisitional stage of a preposition which is part of a collocation ‘engage in’ was set to be Stage 3 unlike those of other prepositions associated with passive forms (i.e., Stage 2), because its associated verb ‘engage’ is classified as an academic word, which is assumed more difficult than K1 and K2 words, and, therefore, acquired later than those K1 and K2 words. Similarly, the null article (Ø) was expected to be acquired at a later stage than the overt articles (a/an and the) due to its covertness. Table 4.1 shows the target grammatical features of the academic grammar test items and their expected developmental stages, which were tentatively determined.

4.1.2. Scoring Rubric

The rating scheme of the academic grammar test used a three-point scale, scores ranging from 0 to 2, and a partial score (1 point) was given to responses that demonstrated evidence for some level of acquisition of the target grammatical feature but failed in accuracy. Primary and secondary target grammatical features for each test item were also
Table 4.1. Target Grammatical Features of the Academic Grammar Test and their Expected Developmental Stages

<table>
<thead>
<tr>
<th>Item</th>
<th>Grammatical Feature</th>
<th>Stage</th>
<th>Item</th>
<th>Grammatical Feature</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q01</td>
<td>Indefinite article (a)</td>
<td>1</td>
<td>Q11</td>
<td>Present perfect (without an overt cue)</td>
<td>3</td>
</tr>
<tr>
<td>Q02</td>
<td>Null article (Ø)</td>
<td>3</td>
<td>Q12</td>
<td>NEG + SV inversion</td>
<td>4</td>
</tr>
<tr>
<td>Q03</td>
<td>Preposition</td>
<td>2</td>
<td>Q13</td>
<td>Present participle</td>
<td>2</td>
</tr>
<tr>
<td>Q04</td>
<td>Passive</td>
<td>2</td>
<td>Q14</td>
<td>Subjunctive</td>
<td>4</td>
</tr>
<tr>
<td>Q05</td>
<td>Preposition (Passive)</td>
<td>2</td>
<td>Q15</td>
<td>Subjunctive</td>
<td>4</td>
</tr>
<tr>
<td>Q06</td>
<td>Past participle (Passive)</td>
<td>2</td>
<td>Q16</td>
<td>NP + present participle</td>
<td>2</td>
</tr>
<tr>
<td>Q07</td>
<td>Preposition</td>
<td>2</td>
<td>Q17</td>
<td>Cancel SV Inversion (nominal clause)</td>
<td>2</td>
</tr>
<tr>
<td>Q08</td>
<td>Preposition (Collocation)</td>
<td>3</td>
<td>Q18</td>
<td>Multiple WH-questions</td>
<td>3</td>
</tr>
<tr>
<td>Q09</td>
<td>Gerund</td>
<td>2</td>
<td>Q19</td>
<td>Relative clause (with an overt relative pronoun)</td>
<td>3</td>
</tr>
<tr>
<td>Q10</td>
<td>Present perfect</td>
<td>2</td>
<td>Q20</td>
<td>Relative clause (without an overt relative pronoun)</td>
<td>3</td>
</tr>
</tbody>
</table>

included in the rubric so raters could refer to them in assigning scores to students’ responses. The primary features were target construct grammatical features that should be present in a student’s response to receive full credit (2 points) awarded. Secondary features were the linguistic features related to the primary feature or to the target structure of the question whose presence in the response deserved at least a partial score (i.e., 1 point) when the response lacked the primary feature. They were carefully chosen so raters could refer to it when they needed to determine whether a test taker’s answer provided evidence of partial acquisition of the primary grammatical feature. An example of these scoring principles is
presented in Table 4.2, where the primary target grammatical feature of Question 8 is a

collocation \((to)\) engage in. The respective test item is shown in Figure 4.6.

Table 4.2. Example of a Scoring Rubric

<table>
<thead>
<tr>
<th>Q #</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>Primary Feature</th>
<th>Secondary Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>in</td>
<td>Other prepositions (followed by ‘making’)</td>
<td>to (infinitive)</td>
<td>Collocation ((engage in (NP)))</td>
<td>Preposition</td>
</tr>
</tbody>
</table>

Complete the sentence below using all the given words or phrases. You may change word forms if necessary. Also, add minimally required words if necessary.

We are actively engaged \((in\ making)\) sense of TV programs. [make]

Figure 4.6. Question 8 on the academic grammar test

This question is intended to elicit two different grammatical features: (1) the

knowledge of collocation engage in (Question 8) and (2) the gerund form of make (Question

9). In Question 8, the test takers are supposed to add a preposition in associated with the

verb \((to)\) engage. If they acquired this collocation, they would correctly add the preposition

in the blank in the given sentence. On the other hand, if they did not master this collocation

but knew that a preposition is needed in that specific place, they would insert a preposition, possibly other than in, their response would meet the secondary target feature ‘preposition’ and, thus, deserve a partial credit (i.e., 1 point). This response would also provide evidence that the test takers have acquired (or have been acquiring) the target collocation expression,
although not having mastered it yet. On the other hand, if a test taker places an infinitive to
(which is followed by the infinitive form of make), this would serve as evidence that he or
she has not acquired the target collocation. A complete scoring rubric is presented in
Appendix C.

4.2. Data Collection

The data for the present study were collected in three phases. The first data collection
was implemented in Fall 2012 (F12) to address most of the quantitative aspects of the test
validation. The second phase of data collection was implemented during the EPT
administered in Spring 2013 (Sp13). The data for this phase were collected in two strands:
(1) one in a large-group paper-based test and (2) the other in a small-group computer-based
test. While the first two phases paid attention to the quantitative aspects of data collection,
Phase 3 centered on the collection and analysis of qualitative data, most of which would
address the questions on the use of the academic grammar test for making ESL placement
decisions and possible consequences, which could be brought about by the use of the test.

4.2.1. Quantitative Data Collection I: Paper-based tests

4.2.1.1. Participants

Fall 2012. The first set of the quantitative data was collected during F12 with 558
non-native English speaking students, who entered the university as either graduate or
undergraduate students in a variety of academic disciplines at Iowa State University. As a
result, a total of 557 students’ responses were used for the pursuit of the research questions.
Among the 557 students, 206 students were graduate students, and the remainder (N = 351)
was undergraduate students. Three hundred thirty-three students were male and 222 students were female. Two students did not indicate their gender. First language information of the students was not collected.

**Spring 2013.** The second set of the quantitative data was collected during S13 with 179 students attending the regular session of the EPT and taking the paper test. These participants were from the same target population as the F12 samples and admitted to the university, based on the same admission criteria. This participant group was composed of 155 undergraduate and 24 graduate students; 127 students were male and 52 were female. The first language of the Sp13 test takers was as follows: Chinese (73), Korean (60), English (10), Arabic (7), Hindi (4), Vietnamese (4), Malay (3), Tamil (3), Bangali (2), Gujarati (2), Swahili (2), Telugu (2), Ganda (1), Igbo (1), Japanese (1), Persian (1), Portuguese (1), Sinhalese (1), and Spanish (1). It appears that students who indicated English as their first language came from countries where English is one of the official languages or language of instruction.

### 4.2.1.2. Data collection procedures

The academic grammar test was administered during orientation week prior to the beginning of the Fall 2012 and Spring 2013 semesters, along with the English Placement Test at Iowa State University. The students were asked to complete the test within 15 minutes and leave unanswered questions blank in case they failed to finish the test within the time limit. When the EPT test administration and result processing was finished, test takers’ responses on the academic grammar test were entered in an Excel spreadsheet. When data entry was complete, the individual responses were graded based on the rubric, using a semi-automatic rating method. The investigator provided a second-round grading independently to
identify responses that deserve partial credit but failed to receive any for such reasons as
missing a word or a few additional words not provided in the test tasks. In the F12 dataset,
one student’s entry was removed from the analysis due to many missing responses. The
sample size for this set is \( N = 557 \).

4.2.2. Quantitative Data Collection II: Computer-based test

4.2.2.1. Participants

Due to limited lab facilities on campus, the computer-based test was administered to
late-arrival students, who needed to take the make-up session of the EPT a week after the
paper-based test administration. A total of 24 students took the computer-based test. Four
were graduate students. Two-thirds of the test takers (i.e., 16) were male students. Nine of
the 24 students spoke Chinese as their first language (L1), followed by Arabic and Korean
speakers (4 each), and two Portuguese speakers. The first languages of the remaining test
takers were Croatian, French, Hindi, Malay, and Spanish.

4.2.2.2. Data collection procedures

The computer-based test was delivered on Blackboard Learn, a course management
system, since it is available through the university license. After logging into the system,
students began with the academic grammar test and then moved on to the other EPT sections.
Figure 4.7 illustrates the format of the academic grammar test delivered via Blackboard
Learn. Presentation of items grouped by task type per page was not available by the system.
Directions were thus included in every single question. As in the paper-based test, 15
minutes was set for the computerized testing, and the system automatically saved and
submitted test takers’ responses entered in the text boxes and terminated the test session.
when the time elapsed. When the test was finished, students’ responses were transferred into a spreadsheet pre-designed for scoring as used for the scoring of the responses from the paper-based test.

![Figure 4.7. Screenshot of the academic grammar test on Blackboard Learn](image)

### 4.2.3. Interrater Reliabilities

After the investigator graded both F12 and Sp13 responses, one graduate student in Applied Linguistics was invited as the second rater for another independent rating of the responses. After completing a rater training session with the investigator, she graded 20% of the responses randomly sampled from F12 and Sp13 separately ($N_{F12} = 112; N_{Sp13} = 40$). The interrater reliability of the total scores was measured in terms of the Pearson Product-Moment Correlation, $r = .990$ ($p = .000$). The interrater reliability for the rating of individual
responses was also calculated in terms of the Pearson Product-Moment Correlation and Cohen’s *Kappa* coefficient. The former was \( r = .968 \) (\( p = .000 \)), and the *Kappa* coefficient for the agreement between the two raters was \( k = .948 \) (\( p = .000 \)).

### 4.2.4. Qualitative Data Collection

#### 4.2.4.1. Think-aloud protocols

**Participants.** Three non-native English speaking students were recruited by email to participate in the think-aloud protocols. Two were native Korean speakers, and the other spoke Chinese as his first language. One of the Korean speakers was a female graduate student, and the other participants were male undergraduate students. The participants had stayed in the United States for 8 to 18 months by the time of participation. The two Korean participants had experienced learning English in an English-speaking country like the U.S. and Australia before they began studying at Iowa State University. The participants came from different academic disciplines—Apparels Study, Agricultural Biochemistry, and Accounting. All were aged in early their 20s.

**Data Collection Procedures.** The investigator met with each participant in a computer lab where a computer screen recording program, *Camtasia*, was installed on the computers. After obtaining the participants’ official consent for participation, the researcher provided detailed instructions on what they were supposed to achieve, while completing the questions on the academic grammar test on a computer. A standard instruction for the think-aloud protocols is shown in Figure 4.8, adapted from Rosa and O’Neill (1999, cited in Gass & Mackey, 2007, p. 57). The instructions were provided in both English and the participants’ native language.
In this experiment, I am interested in what you think about when you complete the academic grammar test. To find out, I am asking you to THINK ALOUD as you answer each of the questions on the test. You may use English or Korean/Chinese during the task to express your thoughts freely. Please talk CONSTANTLY. I don’t want you to plan out what you say or try to explain to me what you are saying. You can act as if you were alone in this room, speaking to yourself. Please talk clearly and loudly into your microphone. I will not help you with the questions during the experiment. Your performance on the test will not be graded, although I can provide you with some feedback on your responses on the test if you like. The experiment may take approximately 20-30 minutes. You do not need to rush, even if it takes longer than that. Do you have any questions?

Figure 4.8. Instructions for think-aloud protocols

After the instructions, the participants had a short practice session with the investigator to become familiarized with the think-aloud protocol procedures before the actual data collection began. When they felt comfortable with the practice, the actual data collection procedures proceeded. The test was delivered via Moodle, an open-source course management system used by the Department of English (Figure 4.9).

The participants were encouraged to use their native language to speak freely about what they thought during participation. From time-to-time, the investigator prompted them to speak more about their thinking when they did not produce much orally. A graduate student in Applied Linguistics, who also spoke Chinese as his first language, assisted with data collection from the Chinese-speaking participant. The entire think-aloud protocol
sessions were both audio- and video-recorded, and the investigator observed the entire sessions. Their interactions with the test were also recorded using Camtasia.

![Academic Grammar Test on Moodle](image)

**Figure 4.9.** Academic grammar test on Moodle utilized for think-aloud protocols

The recorded data were first transcribed and then translated independently into English by two native speakers of each participant’s first language. The two translated texts from each participant’s data were cross-referenced for appropriate understanding of their cognitive and metacognitive processes during the test tasks. The think-aloud data were
analyzed with a particular focus on two questions: (1) whether the completion of the tasks on the academic grammar test requires background knowledge of the test takers and (2) whether the tasks successfully elicited the test takers’ ability to use the target grammatical features.

4.2.4.2. Individual interviews

Participants. To understand students’ perceptions concerning the use of the academic grammar test and its consequences, the investigators had individual interviews with 31 students, who took the EPT in F12 and Sp13. To recruit participants from both the Pass and the ESL groups, separate email messages were sent to those who passed the EPT in both terms and the students who were enrolled in an ESL writing class in Sp14. As a result, 15 students from the Pass group and 16 students from the ESL group participated in an interview with the investigator. Among the Pass group participants, eight were graduate students and the others were undergraduate students. Seven of the 15 participants were female students, and the remaining students were male. The demographic compositions of the ESL groups were five graduate students and 11 undergraduate students; and six male and 10 female students.

Procedures. When the participants showed interest to participate in the study as interviewees, the investigator corresponded with them by email to schedule their interview session. A private room on campus was reserved for each individual session. Each semi-structured interview lasted approximately 15 to 30 minutes and was audio-recorded. Afterwards, the recorded interviews were transcribed for a later analysis by the investigator. The interview questions are attached in Appendix D.
4.2.4.3. Writing sample collection and error analysis

Participants and procedures. The writing samples for the crosstab analysis between the EPT writing results and the academic grammar test results were collected from the same interview participants described in 4.2.4.2. The ESL group participants had written a short essay similar to the one in the EPT test on the first day of class, and these samples were collected from their ESL instructors with the students’ agreement. The Pass group participants were asked to write a short essay using the same essay prompts used in their ESL writing classes. Different prompts were used for undergraduate and graduate students (See Appendix E). As a result, 12 writing samples were collected from the ESL group participants and 15 from the Pass group participants. The hand-written essays were scanned into PDFs and also transcribed into MS Word with their errors in spelling, grammar, and mechanics preserved by turning off the automatic correction features of MS Word’s program.

Error coding. To analyze the grammaticality of each writing sample collected, a table of categories and subcategories of grammatical errors was created, based on Dulay, Burt, and Krashen’s (1982) surface strategy taxonomy of errors (omission, addition, misformation, and misordering, cited in Ellis (1994)) and Murrow’s (2010) error categories (Table 4.3). Examples of the grammatical error categories and subcategories can be found in Appendix F.

Once the error coding scheme was completed, two native English speakers with a background in Applied Linguistics were invited to independently code grammatical errors of the 27 writing samples. They were provided with the coding scheme and the examples, and after trying a couple of samples with the scheme, they checked with the investigator about their coding practices and asked questions that had arisen in the trial. To avoid any order
effect in coding grammatical errors, the investigator generated two random orders of coding and asked the raters to follow the coding order they were provided. When the raters completed their coding, the investigator collected the coding results and compared the results. The agreement rate between the two raters in the identification of grammatical errors was approximately 55%. Thus, the investigator compared errors identified by the raters, adjudicated the coding, and entered the number of grammatical errors for each writing sample in an Excel spreadsheet prepared for a follow-up analysis of grammaticality. Although identified in the coding, errors in punctuation, spelling, spacing, and letter case did not count in the number of errors because they are errors of mechanics in writing. For a similar reason, errors of content word choice and unclear meaning were excluded from the error counts, as they can be attributed to the lack of vocabulary knowledge. When the adjudication was complete, the writing samples were again coded in terms of T-Unit, clause, and sentence. Following Hunt (1965), T-units were defined as a matrix clause and all dependent clauses attached to it. When a sentence was composed of two or more clauses connected by a coordinate conjunction, each clause was considered as an independent T-unit. On the other hand, fragmented, subordinate clauses that stood alone did not count in as a T-unit (e.g., Whether among the grades in school or the things they have for birthday[F].)

**Grammaticality indices.** Grounded in Wolfe-Quintero, Inagaki, and Kim’s (1998) extensive reviews of literature in measures of L2 fluency, accuracy, and complexity, eight grammaticality indices were adopted to investigate the extent to which students’ grammatical writing ability exerted in a non-testing setting corresponds to their performance in the academic grammar test. These eight indices are:
Table 4.3 Categories and Subcategories of Grammatical Errors

<table>
<thead>
<tr>
<th>Categories</th>
<th>Subcategories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Verb</td>
<td>Tense [Will], Verbs Present Participle [Will]</td>
</tr>
<tr>
<td>Auxiliary Verb</td>
<td>Tense [Will], Verbs Present Participle [Will]</td>
</tr>
<tr>
<td>Modal Verb</td>
<td>Superfluous [Prep], Past Participle [Will]</td>
</tr>
<tr>
<td>Noun</td>
<td>Number, Word Order [Prep], Noun-Adjective [Prep], Noun [Prep]</td>
</tr>
<tr>
<td>Personalization</td>
<td>Pronoun, Gender [Prep], Number [Prep]</td>
</tr>
<tr>
<td>Infinitive Verbs</td>
<td>Missing [Prep], Supersfluous [Prep], Missing [Prep]</td>
</tr>
<tr>
<td>Incorrect Case [Cap]</td>
<td>Missing [Prep], Superfluous [Prep], Missing [Prep]</td>
</tr>
<tr>
<td>No Case between Words [Punc]</td>
<td>Missing [Prep], Superfluous [Prep], Missing [Prep]</td>
</tr>
</tbody>
</table>
(1) Total number of errors (E),
(2) Error per T-unit (E/T),
(3) Error per clause (E/C),
(4) Error per word (E/W),
(5) Total number of error-free T-units (EFT),
(6) Total number of error-free clauses (EFC),
(7) Error-free T-unit ratio (EFT/T), and
(8) Error-free clauses per sentence (EFC/S).

Among many others reviewed in Wolfe-Quintero et al., these variables demonstrated either a moderate/strong correlation with proficiency levels \( r \geq .45 \) or a significant effect for proficiency for more than one proficiency level \( p < .05 \) in two or more studies. In addition to these eight indices, three other grammaticality indices were added to the analysis to reflect more on the syntactic complexity as well. The three additional indices are:

(1) Error-free sentence (EFS),
(2) Error-free clauses per clause (EFC/C), and
(3) Error-free sentences per sentence (EFS/S).

These 11 grammaticality indices can be classified into four categories of grammatical accuracy: (1) error frequency, (2) error ratio, (3) accuracy frequency, and (4) accuracy ratio. Table 4.4 shows the classification of the eleven grammatical indices into the four categories of grammatical accuracy. The ratio indices were normalized per 100-unit (i.e., word, T-unit, clause, or sentence).
Table 4.4. Classification of Grammaticality Indices by Accuracy Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Grammaticality Index</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error frequency</td>
<td>E</td>
<td>Grammatical Error</td>
</tr>
<tr>
<td></td>
<td>E/T</td>
<td>Error per T-unit</td>
</tr>
<tr>
<td>Error ratio</td>
<td>E/C</td>
<td>Error per clause</td>
</tr>
<tr>
<td></td>
<td>E/W</td>
<td>Error per word</td>
</tr>
<tr>
<td>Accuracy frequency</td>
<td>EFT</td>
<td>Total number of error-free T-units</td>
</tr>
<tr>
<td></td>
<td>EFC</td>
<td>Total number of error-free clauses</td>
</tr>
<tr>
<td></td>
<td>EFS</td>
<td>Total number of error-free sentences</td>
</tr>
<tr>
<td>Accuracy ratio</td>
<td>EFT/T</td>
<td>Error-free T-unit ratio</td>
</tr>
<tr>
<td></td>
<td>EFC/C</td>
<td>Error-free clauses per clause</td>
</tr>
<tr>
<td></td>
<td>EFC/S</td>
<td>Error-free clauses per sentence</td>
</tr>
<tr>
<td></td>
<td>EFS/S</td>
<td>Error-free sentence per sentence</td>
</tr>
</tbody>
</table>

4.2.4.4. Individual interview with EPT coordinator

To understand the decision-maker’s perception about issues pertinent to a hypothesized use of the academic grammatical test for ESL decision-making, the EPT coordinator was contacted and met with the investigator for a single, semi-structured interview session. As an associate professor in Applied Linguistics at the time of the interview, he had been coordinating the EPT for seven years. The questions are provided in Appendix G. The interview was audio-recorded and transcribed into MS Word afterwards.
4.2.4.5. Focus group interviews with instructors of ESL writing courses

Participants. Two separate focus group sessions were held with six instructors of ESL writing at Iowa State University to listen for their perceptions about the value of the grammatical writing ability in writing instruction and potential consequences of the use of the academic grammar test for the ESL placement purpose. Four instructors met in the first session, one teaching the lower-level writing course (ENGL 101B), two teaching the upper-level undergraduate writing course (ENGL 101C), and one teaching the upper-level graduate writing course (ENGL 101D). The participants of the second focus group meeting were both teachers of ENGL 101C. All focus group participants were graduate students in Applied Linguistics. One was a native speaker of English and the other participants were advanced-level English speakers from different countries.

Procedures. Each focus group met with the investigator in a private, reserved room on campus. Each focus group interview lasted approximately for 45 minutes. Before each focus group interview began, the investigator read aloud focus group protocols (Appendix H) so that the participants were aware of the procedures. Each focus group session was audio-recorded and transcribed into MS Word afterwards.

4.3. Data Analysis Methods

Quantitative data were analyzed to address most of the research questions from the positivist perspective. The interpretation of the analysis results concerns (1) the assessment quality and generalizability of test scores across broader domains of grammatical writing tasks, and (2) their relationship to the theoretical concepts underlying the test construct—namely, productive grammatical writing ability in an academic setting.
4.3.1. Quantitative Data Analysis Methods

The data obtained from the administration of the academic grammar test were analyzed to address most of the sub-questions under Research Question 2, using the methods as follows.

4.3.1.1. Descriptive statistics (RQ2a)

First, the descriptive statistics of the test results were reviewed to explore how examinees performed on the test. If the scores were normally distributed across a wide range of possible scores, it could be assumed that the statistical characteristics of the test were suitable for norm-referenced purposes. The descriptive statistical results could also provide the grounds for further statistical investigations of the test and item characteristics (Bachman, 2004).

4.3.1.2 Two-way ANOVA (RQ 2b)

To explore if gender as a construct-irrelevant variable played a significant role in affecting students’ performances on the academic grammar test and also to see whether it interacts with students’ general English proficiency level, a two-way ANOVA was conducted with gender and general proficiency level being two factors. Both predictors are binary categorical variables: ‘male’ vs. ‘female’ for gender and ‘Pass’ vs. ‘ESL’ for general proficiency level. The proficiency levels were determined using students’ EPT results; those who had passed all sections of the test were classified into the Pass group, whereas those who had been required to take at least one of the ESL courses were classified into the ESL group.

4.3.1.3. Reliability analysis (RQ 2c)

A Cronbach’s alpha coefficient was calculated using SPSS to view the internal consistency of the test items and also the generalizability of the test over the universal
domain of test tasks. An alpha coefficient equal to or greater than .7 was considered acceptable following common rules of thumb (Kline, 2000) and also given the fact that the test items were pilot items.

4.3.1.4. Independent two-sample t-test (RQ 2f)

A statistical significance of the difference in mean score between two proficiency groups (Pass vs. ESL) was examined in terms of an independent two-sample t-test. The analysis would provide backing for the warrant of the Extrapolation Inference (W4).

4.3.1.5. ANOVA and a post hoc Scheffé test (RQ 2f & 2h)

These two methods were used to investigate the extent to which the three groups of test takers classified by their results on the EPT writing performed distinctively from one another (W4). They were also used to explore the difference of the item groups (which were classified by expected developmental stages of target grammatical features as described in 4.1.1.4) in terms of inversed item difficulty (i.e., item facility (IF)). If the mean scores of the IF indices of the item groups correspond with their target features’ expected developmental stages (and, thus, their expected difficulty levels), evidence would be obtained for the warrant that the scores of the academic grammar test are attributed to the theoretical assumptions underlying the construct (W5).

4.3.1.6. Spearman rho and Pearson correlations (RQs 2g, 2h & 2j)

In the same vein, a strong correlation between items’ difficulty levels and the expected developmental stages of the respective grammatical features could provide an additional support for Warrant 5. A Spearman rho correlation coefficient was calculated to address this question, since the developmental stage is a categorical ordinal variable, while item difficulty indices are of a continuous interval variable (Bachman, 2004).
Another correlation analysis was conducted to investigate the question whether students’ grammatical abilities demonstrate consistency across different English proficiency tests that tap into the same target construct even though the primary constructs of the tests may differ. In particular, a Pearson product-moment correlation coefficient was calculated to see the relationship of students’ performances between the Internet-based Test of English as a Foreign Language (TOEFL iBT®) and the academic grammar test, given that the scores on the two tests are on interval scales. Along a similar vein, a Spearman correlation was calculated on the relationship of scores from the EPT Writing and the grammar test, since the former is graded on a three-point ordinal scale (101B (lower-level ESL writing), 101C/D (upper-level ESL writing), and Pass).

4.3.1.7 Confirmatory factor analysis (RQ 2i)

Conceptual associations of target grammatical features with the construct of grammatical writing ability in academic setting were tested in first-order and second-order models using Mplus, a software program developed for various statistical analyses with latent variables. A dataset in which F12 and Sp13 data were combined was used for the analysis (N=736). Figures 4.10 through 4.13 illustrate the first- and second-order models to test the relationships between grammatical features and the test construct. For comparisons of these models, several fit indices will be examined, which includes the $\chi^2$ statistic, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), goodness of fit index, adjusted goodness of fit index, normed fit index, non-normed (Tucker-Lewis) fit index, and comparative fit index.
Figure 4.10. Path diagram of the first-order model 1

Figure 4.11. Path diagram of the first-order model 2
Figure 4.12. Path diagram of the second-order model 1

Figure 4.13. Path diagram of the second-order model 2
4.3.1.8. Overlay scatterplot (RQ 2g)

As a way of examining the relationship between students’ performance on the academic grammar test and their grammatical writing performance in a non-testing setting, values on some grammatical indices obtained from the interview participants’ writing samples were plotted against their scores on the academic grammar test. Since they demonstrated similar patterns to each other, the overlay scatterplot that shows the clearest distinction between the Pass and ESL groups will be chosen and presented in Chapter 5.

4.3.1.9. Identification of false positive in EPT exemptions (RQ 5)

A final analysis conducted in this phase reviewed the correspondence between grammar test scores and EPT writing results. This aimed to identify false positives in ESL writing exemptions that might be brought about when decisions were made exclusively based on examinee performance on the EPT essay writing test. For this purpose, students’ EPT writing results and their academic grammar test scores were first aligned along with information from the number raters involved in the essay rating for each student’s essay to determine an appropriate cutoff score for the grammar test. Plausible candidate cutoff scores were sought by aligning students’ grammar scores along with their EPT results and TOEFL writing scores.

When a cutoff score was determined, the number of passes and non-passes on the EPT writing was matched with the number of passes and non-passes on the academic grammar test and formatted into a contingency table (Table 4.5). This analysis, in particular, focused on the number and the percentage of false positives in EPT writing exemption (i.e., Pass on the ESL writing) that failed to pass the academic grammar test. Identification of such false positive decisions on the EPT writing test would indicate the worth of the
academic grammar test as a supplement to the EPT writing test. False negative samples were not the scope of this analysis, given that the EPT essay ratings make use of a variety of criteria, such as coherence, linguistic/discourse functions, and spelling and mechanics, in addition to grammatical accuracy. It is plausible that a student who has developed an advanced-level proficiency in English may have not developed proper discourse-level writing skills. The contingency table-based analysis was conducted with all EPT essays and controversial EPT essays that required more than two ratings in decision-making.

Table 4.5. Contingency Table of the EPT Writing and the Academic Grammar Test Results

<table>
<thead>
<tr>
<th>Academic Grammar Test</th>
<th>Non-pass (i.e., Below cutoff score)</th>
<th>Pass (i.e., cutoff score or above)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPT Writing Pass</td>
<td>False positive</td>
<td></td>
</tr>
<tr>
<td>Non-Pass</td>
<td>(False negative)</td>
<td></td>
</tr>
</tbody>
</table>

4.3.2. Qualitative Data Analysis Methods

The transcribed qualitative data were analyzed using the grounded theory method (Brott & Myers, 2002). The investigator read through the transcripts and induced common themes that arose across different samples. Those themes were coded using a qualitative data analysis software program, NVivo 10. In addition, student participants’ answers to the interview questions were tabulated in Excel to quantify the proportions of different views among the participants for each question. Among the coded qualitative data, only those relevant to the research questions will be presented in the next chapter.
4.4. Chapter Summary

This chapter described the design of the test materials and the scoring rubric, as well as the data collection procedures and the data analysis methods from both the quantitative and qualitative aspects of the research. The section on test materials described the considerations and procedures in selecting the target grammatical features for the academic grammar test and in designing three task types and test items. It also presented the rationale behind the decision on the expected developmental stages of the target grammatical features, with a summary table of the target features and their respective developmental stage. Since the quantitative data collected in Phase 1 and Phase 2 shared the analysis methods, their data collection procedures and statistical analysis methods were presented together. The descriptions about the qualitative data collection centered on the procedures of think-aloud protocols, interviews with student participants, EPT coordinator, and ESL writing instructors, and the collection of writing samples and error coding procedures. While the grammaticality analysis of the student participants’ writing samples was conducted statistically in conjunction with their scores on the academic grammar test, the procedures involved in the identification of grammatical errors had more qualitative characteristics than quantitative. Therefore, their descriptions were included in the qualitative section.
CHAPTER 5. RESULTS

In this chapter, the results of the data analysis are presented to provide answers to the research questions. Each answer targets a particular warrant in the validity argument. This chapter begins with the quantitative results and then moves to the qualitative evidence for the validity argument. Each section of the results will also start with the research question pertaining to that result to help the reader understand the relevance of the results to the overall study. Note that the Spring 2013 (Sp13) data results reported in this section are mostly obtained from the paper-based test administration unless indicated otherwise.

5.1. Quantitative Evidence for the Validity Argument

5.1.1. RQ 2a: Do the academic grammar test scores effectively discriminate among students of varying proficiency levels? (Warrant 2)

The question of whether the test differentiates among individuals within the intended population of test takers of varying proficiency levels to an acceptable extent can be answered by reviewing the distribution of test scores. The descriptive statistics and the histograms of the total scores of the academic grammar test from the F12 and Sp13 administrations were examined as the first step. Figure 5.1 shows that the test scores spread widely across the possible range of scores distributed in a more or less normal manner, regardless of the sample size, which was considerably larger for the F12 ($N = 557$) than the S13 ($N = 179$) sample. Given the large sample size for the F12 sample, the range of the total scores from F12 is slightly greater than from Sp13, as indicated in Table 5.1. It is noteworthy that, despite a high discrepancy in sample size, the mean and standard deviation
of the examinees’ total scores on the academic grammar test administered in Sp13 are very similar to those obtained in F12. Hence, the descriptive statistics provide backing for the assumption about the effective norm-referenced evaluation of examinees’ performance underlying Warrant 2 for the Evaluation Inference.

![Academic Grammar Test (F12)](image1)
![Academic Grammar Test (Sp13, Paper)](image2)

**Figure 5.1.** Histograms of total scores of the academic grammar test (F12 and Sp13)

**Table 5.1.** Descriptive Statistics of Academic Grammar Test Scores (F12 & Sp13)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>Median</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Skewness Stat</th>
<th>Skewness Std. Error</th>
<th>Kurtosis Stat</th>
<th>Kurtosis Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>F12</td>
<td>557</td>
<td>21.92</td>
<td>7.014</td>
<td>22</td>
<td>36</td>
<td>1</td>
<td>37</td>
<td>-.233</td>
<td>.104</td>
<td>-.561</td>
<td>.207</td>
</tr>
<tr>
<td>Sp13</td>
<td>179</td>
<td>21.49</td>
<td>6.235</td>
<td>21</td>
<td>31</td>
<td>6</td>
<td>37</td>
<td>.218</td>
<td>.182</td>
<td>-.359</td>
<td>.361</td>
</tr>
</tbody>
</table>
5.1.2. **RQ 2b: Does gender, as a construct-irrelevant factor, significantly affect examinees’ performances on the academic grammar test? (Warrant 2)**

If gender plays a significant role in affecting examinees’ performance in the academic grammar test, gender will be revealed as a moderator variable, creating an interaction with the proficiency level to predict performance on the grammar test, which would be undesirable. To investigate this question, both male and female students from each dataset were first divided into two general proficiency groups: (1) those who passed all sections of the English Placement Test (Pass) and (2) those who did not (Non-pass). Then, a test of Analysis of Covariance (ANCOVA) was completed between the total scores of the academic grammar test and these two proficiency groups with Gender as a covariate.

As the first step of the analysis, three important assumptions for the model (i.e., normality, homoscedasticity, and homogeneity of regression) were tested in addition to assumptions of the Analysis of Variance (i.e., additivity, linearity, and random errors) to check if the ANCOVA is an appropriate model for the moderating effect of the Gender variable on test performance (Field, 2013). First, Q-Q plots were produced to check the normality of the gender groups from each sample (Figures 5.2 and 5.3). The assumption of normality appears satisfied as most observations stay on (or close to) the line \( y = x \). Second, Levene’s test was used to check the assumption of equal variances between the gender groups (Levene, 1960). From the F12 dataset, Levene’s test of equal variances failed to reject the null hypothesis of equal error variance of the dependent variable (i.e., the total score) across groups \( F = 3.352, p = .068 \). Levene’s test on the homogeneity of variances between male and female students in the Sp13 dataset also failed to reject the null hypothesis.
of equal variance \((F = .055, p = .814)\). Therefore, the assumption of homogeneity of variances was considered met for both samples.

**Figure 5.2.** Q-Q plots of academic grammar test scores by Gender (F12)

**Figure 5.3.** Q-Q plots of academic grammar test scores by Gender (Sp13)
The third additional assumption of ANCOVA that the regression slopes should be the same across the covariate groups was tested by checking the slopes of the marginal total means between the two proficiency groups (i.e., Non-Pass and Pass in EPT (variable: EPT_Pass)) separated by gender. As shown in Figure 5.4, the slopes of the marginal means of the total scores between the two gender groups are similar in their directions. In particular, the slopes from the F12 dataset appear almost identical in degree. Those from the Sp13 dataset crossed each other, and yet they slightly differ in degree. Therefore, the evidence indicates that the three assumptions of normality, homoscedasticity, and the homogeneity of regression were met, thus moving to the next step of ANCOVA.

Table 5.2 presents descriptive statistics of the academic grammar test scores by gender groups from each semester. Results from the ANCOVA tests for both F12 and Sp13 are shown in Tables 5.3 and 5.4. While the main effect of gender on the total score of the academic grammar test is significant for F12, there is no significant interaction between gender and the proficiency level observed from either dataset ($F(1, 551) = .002, p = .967$ for F12; $F(1, 175) = .847, p = .359$ for Sp13). Therefore, it can be argued that gender exerts no significant moderating effect on the relationship between proficiency level and test performance. This finding adds to the backing for the assumption for Warrant 2 about the appropriate observation of test takers’ productive grammatical writing abilities.
Figure 5.4. Slopes of marginal means of academic grammar test scores between the two proficiency groups by gender

Table 5.2. Means (and Standard Deviations) of Academic Grammar Test Total Scores by Gender and Proficiency Level (F12; N = 555 with 2 missing data in gender)

<table>
<thead>
<tr>
<th>Term</th>
<th>Proficiency Group</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F12</td>
<td>EPT Pass (Higher proficiency)</td>
<td>24.02 (6.83)</td>
<td>25.60 (5.50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n = 47)</td>
<td>(n = 93)</td>
</tr>
<tr>
<td></td>
<td>EPT Non-Pass (Lower proficiency)</td>
<td>19.91 (7.35)</td>
<td>21.55 (6.73)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n = 175)</td>
<td>(n = 240)</td>
</tr>
<tr>
<td>Sp13</td>
<td>EPT Pass (Higher proficiency)</td>
<td>27.64 (4.34)</td>
<td>29.21 (4.35)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n = 11)</td>
<td>(n = 14)</td>
</tr>
<tr>
<td></td>
<td>EPT Non-Pass (Lower proficiency)</td>
<td>20.85 (6.11)</td>
<td>20.16 (5.61)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n = 41)</td>
<td>(n = 113)</td>
</tr>
</tbody>
</table>
### Table 5.3. Tests of Between-Subjects Effects (F12)

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>2210.060</td>
<td>3</td>
<td>736.687</td>
<td>16.150</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>13487.994</td>
<td>1</td>
<td>13487.994</td>
<td>295.699</td>
<td>.000</td>
</tr>
<tr>
<td>Gender</td>
<td>247.109</td>
<td>1</td>
<td>247.109</td>
<td>5.417</td>
<td>.020</td>
</tr>
<tr>
<td>EPT_Pass</td>
<td>141.446</td>
<td>1</td>
<td>141.446</td>
<td>3.101</td>
<td>.079</td>
</tr>
<tr>
<td>EPT_Pass*Gender</td>
<td>0.076</td>
<td>1</td>
<td>.76</td>
<td>.002</td>
<td>.967</td>
</tr>
<tr>
<td>Error</td>
<td>25133.291</td>
<td>551</td>
<td>45.614</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>293987.000</td>
<td>555</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27343.351</td>
<td>554</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5.4. Tests of Between-Subjects Effects (Sp13)

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>1467.558</td>
<td>3</td>
<td>489.186</td>
<td>15.699</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>4185.556</td>
<td>1</td>
<td>4185.556</td>
<td>134.321</td>
<td>.000</td>
</tr>
<tr>
<td>Gender</td>
<td>37.571</td>
<td>1</td>
<td>37.571</td>
<td>1.206</td>
<td>.274</td>
</tr>
<tr>
<td>EPT_Pass</td>
<td>3.992</td>
<td>1</td>
<td>3.992</td>
<td>.128</td>
<td>.721</td>
</tr>
<tr>
<td>EPT_Pass*Gender</td>
<td>26.400</td>
<td>1</td>
<td>26.400</td>
<td>.847</td>
<td>.359</td>
</tr>
<tr>
<td>Error</td>
<td>5453.157</td>
<td>175</td>
<td>31.161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>89556.000</td>
<td>179</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>6920.715</td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.1.3. **RQ 2c: Are the assessment results generalizable across the universe of assessment tasks to an acceptable extent? In other words, is Cronbach’s alpha coefficient equal to or greater than .7? (Warrant 3)**

The generalizability of the test results across the universe of assessment tasks was estimated in terms of internal consistency of test items; test items designed to measure the same general construct should produce similar test scores. Given that the test items were pilot items, an alpha level of .7 was set for acceptable reliability following a rule of thumb (Kline, 2000). Cronbach’s alpha coefficients for the academic grammar test were .784 and .723 for F12 and Sp13, respectively. The results thus suggest the assessment results are generalizable across the universe of assessment tasks to an acceptable extent.

5.1.4. **RQ 2d: Does the academic grammar test demonstrate similar reliability across different samples from the same population? (Warrant 3)**

While Cronbach’s alpha coefficients are greater than .7, it is also important to ensure that the alpha coefficients do not vary significantly across different samples from the same target population. In this case, the target population for the test samples should be non-native English speaking adult students entering Iowa State University. A 95% confidence interval (CI) of the difference of the two alpha coefficients was calculated (see Table 5.5). Whereas the alpha coefficient of the Sp13 dataset appears to be much smaller than that of F12, the 95% CI of their difference indicates insignificance in differential reliability, encompassing 0.

Another way to examine the reliability generalizability of the test is to investigate whether there is a significant difference in reliability between the gender groups. Hence, separate alpha coefficients for male and female examinee groups were calculated for both
terms, and 95% confidence intervals of their differences were examined. The results in Table 5.6 lead us to conclude that the test reliability is stable and, therefore, generalizable across different samples of test takers; the alpha coefficients for the two gender groups are not significantly different in either F12 or Sp13.

Table 5.5. Cronbach’s Alpha Coefficients of Academic Grammar Test (k = 20)

<table>
<thead>
<tr>
<th>Term</th>
<th>N</th>
<th>Cronbach’s alpha</th>
<th>95% Confidence Interval</th>
<th>95% Confidence Interval of ρ_{F12} − ρ_{Sp13}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower Limit</td>
<td>Upper Limit</td>
</tr>
<tr>
<td>F12</td>
<td>557</td>
<td>.784</td>
<td>.757</td>
<td>.809</td>
</tr>
<tr>
<td>Sp13</td>
<td>179</td>
<td>.723</td>
<td>.660</td>
<td>.778</td>
</tr>
</tbody>
</table>

* The 95% confidence interval of the differential reliability of a test was calculated as follows (Bonett, 2010):

\[
\text{Lower limit} = \hat{\rho}_{F12} - \hat{\rho}_{Sp13} - \sqrt{(\hat{\rho}_{F12} - L_{F12})^2 + (U_{Sp13} - \hat{\rho}_{Sp13})^2}.
\]

\[
\text{Upper limit} = \hat{\rho}_{F12} + \hat{\rho}_{Sp13} - \sqrt{(\hat{\rho}_{F12} - U_{F12})^2 + (L_{Sp13} - \hat{\rho}_{Sp13})^2}.
\]

Table 5.6. Cronbach’s Alpha Coefficients of Academic Grammar Test by Gender

<table>
<thead>
<tr>
<th>Term</th>
<th>Gender</th>
<th>N</th>
<th>Cronbach’s alpha</th>
<th>95% CI</th>
<th>95% Confidence Interval of ρ_{F} − ρ_{M}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Limit</td>
<td>Upper Limit</td>
</tr>
<tr>
<td>F12</td>
<td>F</td>
<td>222</td>
<td>.806</td>
<td>.766</td>
<td>.841</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>333</td>
<td>.761</td>
<td>.722</td>
<td>.796</td>
</tr>
<tr>
<td>Sp13</td>
<td>F</td>
<td>52</td>
<td>.738</td>
<td>.622</td>
<td>.831</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>127</td>
<td>.716</td>
<td>.639</td>
<td>.783</td>
</tr>
</tbody>
</table>
5.1.5. **RQ 2e: Are scores from the computer-based academic test comparable to those from the conventional paper-and-pencil test? (Warrant 3)**

As described in Chapter 4 (4.2.2.2), the academic grammar test was delivered via computer in Sp13 when a make-up session of the English Placement Test (EPT) was administered for late-arrival incoming students whose first language is other than English. While this group was not randomly sampled, the performance of the participants on the academic grammar test was expected to be similar to that for the sample groups, who had taken the test on paper, because it was assumed that the former group of students had been admitted to the university according to the same selection rules as the latter groups and there were no construct relevant differences in the group of late arrivals. Most had to take the make-up EPT, due to such issues as a late visa approval from the U.S. embassy or consulate, or a delayed flight schedule from their home countries. Yet, it was possible that unexpected variables could play a role to affect the performance of the students in the computerized make-up test, which could lead to a considerable difference in performance between the paper-based and computer-based testing groups. A much smaller sample size of the make-up group than the other groups might also affect the sample’s under- or un-representability of the target population of test takers. Table 5.7 shows the descriptive statistics and Cronbach’s alpha coefficient for the computer-based academic grammar test.

**Table 5.7.** Descriptive Statistics and Cronbach’s Alpha Coefficient of Computer-based Academic Grammar Test (Sp13; N = 24)

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Median</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.625</td>
<td>7.198</td>
<td>12</td>
<td>30</td>
<td>0</td>
<td>30</td>
<td>.794</td>
</tr>
</tbody>
</table>
Comparing the descriptive statistics of the paper-based academic grammar test administered in the same term (Table 5.1), it is easily shown that the statistics of the test results from these two different delivery modes are not comparable in many aspects. Only the standard deviations of the scores from the two test modes seem more or less similar to each other (i.e., 7.014 (paper, F12) and 6.235 (paper, Sp13) vs. 7.198 (computer, Sp13)). On the other hand, the Cronbach’s alpha coefficient for the computer-based test is higher than for the paper-based test administered during the same term. At first glance, these results appear to suggest a significant mode effect on test performance. However, this conclusion should not be made with haste, especially because the sample size of the computer-based test is too small for its participants to represent the target population’s test performance on the same delivery mode. The fact that only three of 24 students who took the make-up EPT passed all sections of the test also corroborates the sample’s unrepresentativeness of the population. A further study with a larger sample should follow to seek appropriate evidence to this research question.

5.1.6. RQ 2f: Do students at different proficiency levels perform distinctively from each other on the academic grammar test? (Warrant 4)

To answer this question, students’ proficiency levels were first defined in two ways: (1) in terms of whether they passed all sections of the EPT test or not (variable: EPT_Pass) and (2) by their results on the EPT writing test (variable: EPT_WRT). Whereas EPT_Pass is a binary variable, EPT_WRT consists of three levels—ENGL 101B (lower-level ESL writing), ENGL 101C/D (upper-level ESL writing), and Pass (no requirement of ESL writing). For each case, the two or three proficiency groups’ performance on the academic
grammar test were compared in terms of the difference of the marginal means. Examining the mean difference between the EPT Pass and Non-Pass groups, the group statistics, t-test statistics, and 95% CI of the mean differences in Table 5.8 all indicate that the Pass groups significantly outperformed the Non-Pass groups in both terms.

Table 5.8. Group Statistics of EPT_Pass for F12 and Sp13

<table>
<thead>
<tr>
<th>Term</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>t-test for Equality of Means (Equal variances not assumed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>t</td>
</tr>
<tr>
<td>F12</td>
<td>Pass</td>
<td>141</td>
<td>25.06</td>
<td>5.980</td>
<td>6.901</td>
</tr>
<tr>
<td></td>
<td>Non-Pass</td>
<td>146</td>
<td>20.85</td>
<td>7.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Pass</td>
<td>154</td>
<td>20.34</td>
<td>5.735</td>
<td></td>
</tr>
</tbody>
</table>

Similarly, there were significant differences in total means among the three proficiency groups defined by the test takers’ results on the EPT writing test ($F(2, 554) = 66.326, p = .000$ for F12 and $F(2, 176) = 27.549, p = .000$ for Sp13). (See Table 5.9 for the descriptive statistics of the three proficiency groups by EPT writing results from both F12 and Sp13.) Interesting are the results of the mean comparisons among the three proficiency groups determined by the students’ results on the EPT writing. Although one of the major factors that divide the lower-ESL and upper-ESL writing groups on the EPT writing test is the degree of grammatical accuracy in students’ essays, the participants falling in the upper-ESL writing group did not produce error-free essays, which often makes it difficult to draw a fine line between the two groups when defining their proficiency levels in English writing.
Despite this concern, these two ESL writing groups performed distinctively from each other in a significant manner on the academic grammar test, not to mention that the highest proficiency group (i.e., Pass on the EPT_WRT) significantly outperformed the ESL writing groups. (See Tables 5.10 and 5.11 for results of the post hoc Scheffé tests for F12 and Sp13, respectively.) This distinctive pattern is also obvious in the 95% confidence interval (CI) error bars of the groups’ mean scores in Figure 5.5. This suggests that the academic grammar test can differentiate students of a lower ESL writing proficiency group from those of a higher ESL writing proficiency group with a five percent error.

### 5.1.7. RQ 2g: Does students’ performance on the academic grammar test correspond to their performance in a non-testing setting? (Warrant 4)

Another piece of evidence for the Extrapolation Inference was sought by examining the correspondence between their performance in the academic grammar test and that in a non-testing setting using the samples collected as described in 4.2.4.3. This issue was

**Table 5.9.** Descriptive Statistics of Academic Grammar Test Scores by EPT Writing Results (F12 and Sp13)

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F12</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101B</td>
<td>87</td>
<td>17.15</td>
<td>6.22</td>
<td>.67</td>
</tr>
<tr>
<td>101C/D</td>
<td>296</td>
<td>20.89</td>
<td>6.71</td>
<td>.39</td>
</tr>
<tr>
<td>Pass</td>
<td>174</td>
<td>26.06</td>
<td>5.62</td>
<td>.43</td>
</tr>
<tr>
<td><strong>Sp13</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101B</td>
<td>29</td>
<td>16.76</td>
<td>5.19</td>
<td>.963</td>
</tr>
<tr>
<td>101C/D</td>
<td>103</td>
<td>20.77</td>
<td>5.81</td>
<td>.572</td>
</tr>
<tr>
<td>Pass</td>
<td>47</td>
<td>25.98</td>
<td>4.83</td>
<td>.705</td>
</tr>
</tbody>
</table>
Table 5.10. Results of Post Hoc Scheffé Test for Differences in the Academic Grammar Test Score among the Three Proficiency Groups by EPT Writing Results (F12)

<table>
<thead>
<tr>
<th>(I) EPT_WRT</th>
<th>(J) EPT_WRT</th>
<th>Mean Diff (I - J)</th>
<th>Std. Error</th>
<th>p</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Limit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Upper Limit</td>
</tr>
<tr>
<td>101B</td>
<td>101C/D</td>
<td>-3.736</td>
<td>.770</td>
<td>.000</td>
<td>-5.62</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>-8.914</td>
<td>.829</td>
<td>.000</td>
<td>-10.95</td>
</tr>
<tr>
<td>101C/D</td>
<td>101B</td>
<td>3.736</td>
<td>.770</td>
<td>.000</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>-5.178</td>
<td>.603</td>
<td>.000</td>
<td>-6.66</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>8.914</td>
<td>.829</td>
<td>.000</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td>101C/D</td>
<td>5.178</td>
<td>.603</td>
<td>.000</td>
<td>3.70</td>
</tr>
</tbody>
</table>

Table 5.11. Results of Post Hoc Scheffé Test for Differences in the Academic Grammar Test Score among the Three Proficiency Groups by EPT Writing Results (Sp13)

<table>
<thead>
<tr>
<th>(I) EPT_WRT</th>
<th>(J) EPT_WRT</th>
<th>Mean Diff (I - J)</th>
<th>Std. Error</th>
<th>p</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Limit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Upper Limit</td>
</tr>
<tr>
<td>101B</td>
<td>101C/D</td>
<td>-4.008</td>
<td>1.150</td>
<td>.003</td>
<td>-6.85</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>-9.220</td>
<td>1.292</td>
<td>.000</td>
<td>-12.41</td>
</tr>
<tr>
<td>101C/D</td>
<td>101B</td>
<td>4.008</td>
<td>1.150</td>
<td>.003</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>-5.212</td>
<td>.963</td>
<td>.000</td>
<td>-7.59</td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>9.220</td>
<td>1.292</td>
<td>.000</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td>101C/D</td>
<td>5.212</td>
<td>.963</td>
<td>.000</td>
<td>2.83</td>
</tr>
</tbody>
</table>
Figure 5.5. 95% CI error bars of mean scores of the three proficiency groups by EPT writing

examined in two ways, using the eleven grammaticality indices introduced in the same method section above. First, the grammaticality indices obtained from the analysis of individual participants’ writing samples were plotted against their scores on the academic grammar test. Second, an approach to the issue was a correlation analysis between the foregoing indices and the examinees’ scores on the academic grammar test.

An overlay scatterplot of the error-per-word ratios (E/W) against the grammar test scores is presented in Figure 5.6 for a clear illustration of the associational patterns between the 27 interview participants’ grammar scores and their accuracy performance in non-assessment contexts. In this figure, the horizontal axis indicates the range of academic grammar test scores, and the vertical axis indicates the range of errors per 100 words. ESL participants are labeled beginning with 101 (indicating the generic course number of ESL classes offered at ISU), whereas Pass-group participants are labeled beginning with P (which stands for ‘Pass’).
With two or three exceptions for each proficiency group, those who scored low on the academic grammar test tended to make more grammatical errors when they wrote in a non-testing context. Also, the figure clearly illustrates that the lower proficiency participants (labeled with 101) are likely to make more grammatical errors than high proficiency participants, who passed the EPT. One outstanding outlier of this pattern is Participant 101-07. This participant appears to belong to the Pass group in regard to her performance on both the academic grammar test score and the accuracy performance. However, her actual status in the development of grammatical accuracy is uncertain, given the fact that she produced a much shorter writing sample (197 words) than the other ESL participants. (Recall that the mean word counts for the writing samples by the ESL participants was 315.58 (Table 5.12).) Few grammatical errors on her writing sample can certainly be attributed to the short length...
of her written production. On the other hand, it is also possible that she was placed in an ESL writing class based on the considerable short length of her EPT essay, even if she had developed her grammatical accuracy to an advanced proficiency level, because the development of ideas is one of the primary factors considered in EPT essay rating (EPT Writing Rubric (n.d.)). If she were a slow writer, it could have considerably affected her performance on the EPT essay which was intended to assess test takers’ writing ability in both linguistic and discourse aspects.

Next, Pearson product-moment correlation coefficients were calculated to measure the strength of association between the academic grammar test scores and the eleven grammaticality indices. The error frequency (E) and ratio (E/T, E/C, and E/W) indices are expected to have a negative correlation with the participants’ scores on the academic grammar test, given the assumption that lower proficiency learners will produce more grammatical errors. With the same reason, the accuracy frequency (EFT, EFC, and EFS) and ratio (EFT/T, EFC/C, EFC/S, and EFS/S) indices will be in a positive relationship with the academic grammar test scores. Table 5.12 presents the descriptive statistics of these features produced by these two proficiency groups. The results of the correlation analysis in Table 5.13 show that all grammaticality indices, except for EFT and EFS, have a significant, moderate association with the academic grammar test scores. The error-free clause per clause (EFC/C) has the strongest association with the academic grammar test scores among the grammaticality indices ($r = .682; p = .000$). On the other hand, the results suggest that accuracy frequency indices are weak indicators of ESL learners’ writing proficiency compared to the other indices.
All results presented in this section offer evidence, which supports the assumption that students’ performance on the academic grammar test corresponds to their accuracy performance in writing that occurs in non-testing contexts. The overall results also suggest that the academic grammar test scores may serve as meaningful indicators of test takers’ grammatical writing ability.

5.1.8. RQ 2h: Does the performance of the academic grammar test items reflect the theoretical assumptions that underlie their development? (Warrant 5)

Based on the Processability Theory and the findings from previous research (e.g., Norris (2005) and Chapelle, Chung et al. (2010)), it was hypothesized that the grammatical features expected acquired at an advanced stage of L2 learning would be more difficult for test takers than those acquired at a beginning stage of the learning. The items targeting such advanced-level grammatical features should thus have lower IF values. As the first step towards the investigation of the research question above, calculations were made of the inversed item difficulty indices (i.e., item facility (IF) indices) and item discrimination (ID) indices (calculated in terms of item-total correlation) obtained from the F12 and Sp13 administrations (Table 5.14). Then, the correspondence of the items’ difficulty levels to their respective developmental stages was examined in two ways: (1) whether the items’ IF indices significantly correlate with their expected developmental stages and (2) whether the four item groups classified by expected development stage are significantly different from each other in mean IF value. Among the test items, Question 1 was excluded from these analyses, since it composed the item group of Stage 1 by itself. Conducted using the items of Stages 2, 3, and 4 (i.e., $k = 19$), the correlation and IF analyses results will be presented.
Table 5.12. Descriptive Statistics of Interview Participants’ Productivity and Grammatical Accuracy Features

<table>
<thead>
<tr>
<th></th>
<th>ESL (N = 12)</th>
<th></th>
<th>Pass (N = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>315.58</td>
<td>73.44</td>
<td>197</td>
</tr>
<tr>
<td>Clause</td>
<td>35.33</td>
<td>12.32</td>
<td>18</td>
</tr>
<tr>
<td>Sentence</td>
<td>18.75</td>
<td>8.11</td>
<td>9</td>
</tr>
<tr>
<td>T-Unit</td>
<td>22.50</td>
<td>9.58</td>
<td>10</td>
</tr>
<tr>
<td>E</td>
<td>30.58</td>
<td>14.52</td>
<td>4</td>
</tr>
<tr>
<td>E/T</td>
<td>139.85</td>
<td>67.12</td>
<td>40.00</td>
</tr>
<tr>
<td>E/C</td>
<td>88.17</td>
<td>45.15</td>
<td>22.22</td>
</tr>
<tr>
<td>E/W</td>
<td>9.29</td>
<td>3.28</td>
<td>2.03</td>
</tr>
<tr>
<td>EFT</td>
<td>8.33</td>
<td>5.03</td>
<td>0</td>
</tr>
<tr>
<td>EFC</td>
<td>16.67</td>
<td>6.67</td>
<td>3</td>
</tr>
<tr>
<td>EFS</td>
<td>5.17</td>
<td>3.64</td>
<td>0</td>
</tr>
<tr>
<td>EFT/T</td>
<td>37.28</td>
<td>21.00</td>
<td>.00</td>
</tr>
<tr>
<td>EFC/C</td>
<td>48.25</td>
<td>17.86</td>
<td>16.67</td>
</tr>
<tr>
<td>EFC/S</td>
<td>95.39</td>
<td>41.35</td>
<td>25.00</td>
</tr>
<tr>
<td>EFS/S</td>
<td>28.08</td>
<td>18.50</td>
<td>.00</td>
</tr>
<tr>
<td>Grammar Score</td>
<td>21.42</td>
<td>6.04</td>
<td>13</td>
</tr>
</tbody>
</table>
Table 5.13. Pearson Product-Moment Correlations between Academic Grammar Test Scores and Grammaticality Indices (N=27)

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>E/T</th>
<th>E/C</th>
<th>E/W</th>
<th>EFT</th>
<th>EFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GramScore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>-.606</td>
<td>-.569</td>
<td>-.584</td>
<td>-.644</td>
<td>.342</td>
<td>.409</td>
</tr>
<tr>
<td>p</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.081</td>
<td>.034</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EFS</th>
<th>EFT/T</th>
<th>EFC/C</th>
<th>EFC/S</th>
<th>EFS/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>GramScore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>.333</td>
<td>.587</td>
<td>.682</td>
<td>.645</td>
<td>.569</td>
</tr>
<tr>
<td>p</td>
<td>.089</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.002</td>
</tr>
</tbody>
</table>

First, Spearman rho coefficients were calculated between the items’ IF values and the expected developmental stages of their respective grammatical features for both terms, since the developmental stages can be considered as categorical, rank-order variables. A negative correlational relationship was expected from this analysis, given the theoretical and empirical grounds previously discussed. The results of the correlation analysis were consistent with this expectation: $r_s = -.643 (p = .003)$ for F12 and $r_s = -.665 (p = .002)$ for Sp13. While both coefficients were statistically significant, the degree of correlation was weaker than expected. This implies that the relation between the test items’ IF values and the developmental stages of the items’ target grammatical features may have been confounded by some undesirable factors, such as task type effects.

In regard to the second aspect of the investigation, visual examinations of the differences among the item groups in IF values in Figure 5.7 suggest that there appears to be a declining tendency in IF values as the level of expected developmental stages increases in both terms. (See Table 5.15 for the descriptive statistics of the item groups’ IF values.) The
results of the Analysis of Variance (ANOVA) tests among these items indicated a significant difference among the item groups for Stage 2, Stage 3, and Stage 4 (Table 5.16). A post hoc Scheffé test revealed that one pair of Stage 2 and Stage 4 items, and another pair of Stage 3 and Stage 4 items were significantly different in IF value, as shown in Table 5.17. However, the means for the IF values for Stage 2 and Stage 3 items were not significantly different from each other.

Figure 5.7. Mean IFs of item groups by target grammatical features’ expected developmental stages (Stage-1 item excluded from analysis, due to its singularity)
Table 5.14  Item Statistics and Expected Developmental Stages (F12 and Sp13)

<table>
<thead>
<tr>
<th>Question</th>
<th>Q01</th>
<th>Q02</th>
<th>Q03</th>
<th>Q04</th>
<th>Q05</th>
<th>Q06</th>
<th>Q07</th>
<th>Q08</th>
<th>Q09</th>
<th>Q10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Grammatical Feature</strong></td>
<td>Indefinite article (a)</td>
<td>Null article (Ø)</td>
<td>Preposition</td>
<td>Passive</td>
<td>Preposition (passive)</td>
<td>Past participle</td>
<td>Preposition</td>
<td>Preposition (collocation)</td>
<td>Gerund</td>
<td>Present perfect</td>
</tr>
<tr>
<td><strong>Dev. Stage</strong></td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>IF (F12)</td>
<td>1.544</td>
<td>1.065</td>
<td>1.605</td>
<td>1.354</td>
<td>0.743</td>
<td>1.478</td>
<td>1.408</td>
<td>0.890</td>
<td>1.226</td>
<td>0.765</td>
</tr>
<tr>
<td>IF (Sp13)</td>
<td>1.413</td>
<td>1.235</td>
<td>1.626</td>
<td>1.296</td>
<td>0.737</td>
<td>1.547</td>
<td>1.430</td>
<td>0.726</td>
<td>1.112</td>
<td>0.782</td>
</tr>
<tr>
<td>ID (F12)</td>
<td>.251**</td>
<td>.359**</td>
<td>.338**</td>
<td>.487**</td>
<td>.583**</td>
<td>.586**</td>
<td>.610**</td>
<td>.597**</td>
<td>.478**</td>
<td>.415**</td>
</tr>
<tr>
<td>ID (Sp13)</td>
<td>.384**</td>
<td>.217**</td>
<td>.235**</td>
<td>.403**</td>
<td>.618**</td>
<td>.463**</td>
<td>.471**</td>
<td>.540**</td>
<td>.441**</td>
<td>.409**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
<th>Q15</th>
<th>Q16</th>
<th>Q17</th>
<th>Q18</th>
<th>Q19</th>
<th>Q20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Grammatical Feature</strong></td>
<td>Present perfect (without a clear cue)</td>
<td>NEG = SV inversion</td>
<td>Present participle</td>
<td>Subjunctive</td>
<td>Subjunctive</td>
<td>NP + present participle</td>
<td>Cancel SV inversion (in an embedded question)</td>
<td>Multiple Wh-questions (embedded)</td>
<td>Relative clause (with an overt relative pronoun)</td>
<td>Relative clause (without an overt relative pronoun)</td>
</tr>
<tr>
<td><strong>Dev. Stage</strong></td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>IF (F12)</td>
<td>0.680</td>
<td>0.201</td>
<td>1.670</td>
<td>0.330</td>
<td>0.548</td>
<td>1.217</td>
<td>1.619</td>
<td>1.083</td>
<td>0.867</td>
<td>1.627</td>
</tr>
<tr>
<td>IF (Sp13)</td>
<td>0.514</td>
<td>0.134</td>
<td>1.637</td>
<td>0.307</td>
<td>0.559</td>
<td>1.223</td>
<td>1.659</td>
<td>1.073</td>
<td>0.855</td>
<td>1.620</td>
</tr>
<tr>
<td>ID (F12)</td>
<td>.479**</td>
<td>.326**</td>
<td>.384**</td>
<td>.434**</td>
<td>.447**</td>
<td>.404**</td>
<td>.296**</td>
<td>.383**</td>
<td>.509**</td>
<td>.397**</td>
</tr>
<tr>
<td>ID (Sp13)</td>
<td>.480**</td>
<td>.334**</td>
<td>.211**</td>
<td>.482**</td>
<td>.472**</td>
<td>.368**</td>
<td>.288**</td>
<td>.296**</td>
<td>.451**</td>
<td>.372**</td>
</tr>
</tbody>
</table>

** Significant at p = .00.
### 5.15. Descriptive Statistics of IF Values by Item Groups (Stages 2, 3, and 4)

<table>
<thead>
<tr>
<th>Item Group</th>
<th>K</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>F12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td>10</td>
<td>1.309</td>
<td>.331</td>
<td>.105</td>
<td>.743</td>
<td>1.670</td>
</tr>
<tr>
<td>Stage 3</td>
<td>6</td>
<td>1.035</td>
<td>.325</td>
<td>.133</td>
<td>.680</td>
<td>1.627</td>
</tr>
<tr>
<td>Stage 4</td>
<td>3</td>
<td>.360</td>
<td>.175</td>
<td>.101</td>
<td>.201</td>
<td>.548</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>1.072</td>
<td>.451</td>
<td>.104</td>
<td>.201</td>
<td>1.670</td>
</tr>
<tr>
<td>Sp13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td>10</td>
<td>1.305</td>
<td>.342</td>
<td>.108</td>
<td>.737</td>
<td>1.659</td>
</tr>
<tr>
<td>Stage 3</td>
<td>6</td>
<td>1.003</td>
<td>.394</td>
<td>.161</td>
<td>.514</td>
<td>1.620</td>
</tr>
<tr>
<td>Stage 4</td>
<td>3</td>
<td>.333</td>
<td>.214</td>
<td>.123</td>
<td>.134</td>
<td>.559</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>1.056</td>
<td>.479</td>
<td>.110</td>
<td>.134</td>
<td>1.659</td>
</tr>
</tbody>
</table>

### Table 5.16. Results of ANOVA Tests of IF values of Item Groups (Stages 2, 3, and 4)

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>F12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.090</td>
<td>2</td>
<td>1.045</td>
<td>10.604</td>
<td>.001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1.577</td>
<td>16</td>
<td>.099</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3.666</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.203</td>
<td>2</td>
<td>1.101</td>
<td>9.170</td>
<td>.002</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1.921</td>
<td>16</td>
<td>.120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4.124</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.17. Results of Post Hoc Scheffé Tests for Differences among Item Groups in Their Mean IF Values (F12 and Sp13)

<table>
<thead>
<tr>
<th>(I) Stage</th>
<th>(J) Stage</th>
<th>Mean Diff. (I – J)</th>
<th>Std. Error</th>
<th>p</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.273</td>
<td>.162</td>
<td>.271</td>
<td>-.164</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>.949</td>
<td>.207</td>
<td>.001</td>
<td>.392</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-.273</td>
<td>.162</td>
<td>.271</td>
<td>-.710</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>.676</td>
<td>.222</td>
<td>.026</td>
<td>.077</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-.949</td>
<td>.207</td>
<td>.001</td>
<td>-1.506</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-.676</td>
<td>.222</td>
<td>.026</td>
<td>-1.274</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.301</td>
<td>.179</td>
<td>.272</td>
<td>-.181</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>.972</td>
<td>.228</td>
<td>.002</td>
<td>.357</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-.301</td>
<td>.179</td>
<td>.272</td>
<td>-.784</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>.671</td>
<td>.245</td>
<td>.046</td>
<td>.010</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-.972</td>
<td>.228</td>
<td>.002</td>
<td>-1.587</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-.671</td>
<td>.245</td>
<td>.046</td>
<td>-1.331</td>
</tr>
</tbody>
</table>

5.1.9. RQ 2i: Do the selected target grammatical features have significant associations with the construct of productive grammatical writing ability? (Warrant 5)

To test the relationship between individual grammatical features implemented through test items in the academic grammar test and a construct of grammatical writing ability, several possible models were hypothesized and evaluated using confirmatory factor analysis (CFA) as described in Section 4.3.1.7. While most models failed to converge or fit poorly, a first-order model with two pairs of grammatical features correlated (i.e., one pair of
present perfect and subjunctive, and another pair of gerund and preposition), shown in Figure 5.8, converged with the best model fit among the candidate models. The model, first set among the others, was expected to demonstrate grammatical writing ability as a unidimensional concept. Each grammatical feature was expected to make a direct contribution to the make-up of the construct of grammatical writing ability instead of liaised with the grand construct by some subordinate constructs. A summary of model fit indices is presented in Table 5.18.

Figure 5.8. Best fit CFA model of the construct Grammatical Writing Ability on observed grammatical features
All fit indices presented in Table 5.17 (except for the $\chi^2$ statistic) suggest that the adopted CFA model has a good fit. For example, the models’ Comparative Fit Index (CFI) and Tucker-Lewis Non-Normed Fit Index (TLI) values are both larger than 0.95. A value is considered a good fit when it is equal to or greater than 0.95 and close to 1.0 (Hooper, Coughlan, & Mullen, 2008). A good model fit is also indicated with values smaller than 0.05 for root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR), both are observed in the summary table, too. In addition, the upper limit of the 90% confidence interval of the RMSEA value is also smaller than 0.05. The $\chi^2$ statistic rejects the null hypothesis of model fitting, but its weaknesses as a fit index are often noted according to Hooper et al. (2008). One weakness is the fact that it is likely to reject the null hypothesis due to its sensitivity to sample size. Hooper et al. (2008) also state an ironical relationship between $\chi^2$ and sample size in a sense that the statistic loses power with a small sample size. Other fit indices, such as the aforementioned, are preferred over the $\chi^2$ and taken into account altogether to judge the model’s fit. Given this, it can be safely concluded that the above first-order CFA model has a good fit.

According to the model, the construct ‘Grammatical Writing Ability’ has moderate to strong (0.303 – 0.775) factor loadings on each observable of individual grammatical features

<table>
<thead>
<tr>
<th>Table 5.18. Summary of Model Fit Indices of the CFA Model in Figure 5.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2 = 63.173$ ($df = 33$, $p = 0.0012$)</td>
</tr>
<tr>
<td>CFI = 0.974</td>
</tr>
<tr>
<td>RMSEA = 0.035</td>
</tr>
<tr>
<td>90% CI (0.022, 0.048)</td>
</tr>
</tbody>
</table>
with significance at the 0.01 level. Based on the loading results, it does not appear one type of grammatical feature (e.g., syntactic features like \(NEG + SV\) inversion or relative clauses) necessarily has stronger associations with the construct than the other type(s) of grammatical features (e.g., morphosyntactic features). This result also supports the expectation that the construct of grammatical writing ability would be a unidimensional concept. It is interesting to note that the observable of ‘Preposition’ has the strongest association with the construct. Given that the acquisition of prepositions can be considered in relation to English lexical knowledge (as in collocations like ‘engaged in’ or ‘interact with’), one might suspect that lexical knowledge should also be part of grammatical ability. Seemingly plausible, this speculation is difficult to generalize with certainty at this point because ‘Preposition’ is the only lexicon-related observable in the model. Further research is required in this regard.

5.1.10. **RQ 2j: Does examinees’ performance on the academic grammar test correlate with their performance in other writing tests that engage grammatical ability at moderate degrees? (Warrant 5)**

The academic grammar test is designed to measure test takers’ grammatical ability in academic English writing. Therefore, it focuses on quite a narrow construct compared to the other language abilities. While it may be a preliminary component of a construct of academic writing, deficiency in the grammatical ability substantially affects the reader’s impression on the writer’s overall writing ability. A positive, moderate degree of correlation between the test takers’ results on the academic grammar test and other tests that engage grammatical ability will provide evidence for the Explanation Inference. The extent to which students’ performance on the academic grammar test was correlated with (1) their
performance on the TOEFL iBT® writing and (2) the EPT writing was thus investigated. The means and standard deviations for the students’ TOEFL iBT® writing scores are presented in Table 5.19.

First, the Pearson product-moment correlation coefficients between the academic grammar test scores and the TOEFL iBT® writing scores were \( r = .491 \) (\( p = .000; \ N = 394 \)) and \( r = .613 \) (\( p = .000; \ N = 90 \)) for F12 and Sp13, respectively. The sample sizes for the correlation analysis are smaller than those for the foregoing analyses, since some students chose to report scores on another English proficiency test, International English Language Testing System (IELTS), instead of the TOEFL iBT®. A correlation analysis between the academic grammar test and the IELTS writing scores was not conducted because individual students’ subscores for each test section of IELTS were not available in the datasets the researcher obtained for the analysis.

<p>| Table 5.19. Means and Standard Deviations for TOEFL iBT® Writing Scores |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>TOEFL iBT® Writing</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F12 (N = 394)</td>
<td>22.15</td>
<td>3.19</td>
<td>21.79</td>
<td>2.81</td>
</tr>
<tr>
<td>Sp13 (N = 90)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The correlation of students’ academic grammar test scores was also examined with their results on the EPT writing graded on a three-level ordinal scale (101B, 101C/D, and Pass. The Spearman rho coefficients for the correlation between the academic grammar test scores and the ordinal EPT writing results were \( r_s = .438 \) (\( p = .000; \ N = 557 \)) and \( r_s = .491 \) (\( p = .000; \ N = 179 \)), respectively. In summary, all of these correlation coefficients met the
expectations of a positive, moderate, significant correlation of the academic grammar test scores with those of another test involving grammatical ability.

5.1.11. RQ 5: To what extent does the use of the academic grammar test help test users identify false positive decisions resulting from the decisions on ESL writing placements/exemptions made exclusively on the basis of examinee performance on the EPT essay test? (Warrant 7)

For the investigation of this research question, a cut score of the academic grammar test for a placement decision was selected first. The variable EPT_WRT_Pass (i.e., whether students passed the EPT writing or not) was selected as a criterion for the decision on the cut score of the academic grammar test because the construct of the academic grammar test is one of the important components of the construct of the EPT writing test. It was also chosen over EPT_WRT, which classified students into three different levels, because the distinction among individual students who failed the EPT writing, based on their performance on the academic grammar test, might not be as clear as ideal, although the marginal mean differences between the lower-level and the upper-level ESL writing groups were statistically significant as shown in Tables 5.10 and 5.11.

The mean grammar test scores for Pass and Fail groups of EPT_WRT_Pass were then calculated from both F12 and Sp13 samples. Table 5.20 presents the means and standard deviations of the academic grammar test scores by both Pass and Fail groups of the EPT writing test. The results in this table clearly show that the performance of the Pass and Fail groups of the EPT writing test on the academic grammar test was comparable across the terms. While the standard deviations of the academic grammar test scores were smaller in
the Sp13 sample for both Pass and Fail groups than those from the F12 sample, the mean scores were very close to each other for each corresponding group across the semesters. In particular, the mean scores for the Pass groups from both F12 and Sp13 were very close to 26 with a difference of a decimal point. Given these results, score 26 was chosen as the cut score for the investigation of the research question above.

Table 5.20. Means and Standard Deviations of Academic Grammar Test Scores by EPT_WRT_Pass

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Non-Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
</tr>
<tr>
<td>F12</td>
<td>174</td>
<td>26.06</td>
</tr>
<tr>
<td>Sp13</td>
<td>47</td>
<td>25.98</td>
</tr>
</tbody>
</table>

Once the cut score for the academic grammar test was set, in order to identify the proportion of false positives in ESL writing placement decisions, contingency tables were created in two ways in relation to the variables of EPT_WRT_Pass (i.e., whether or not students passed the EPT writing part) and Cut26 (whether or not students received 26 or higher on the academic grammar test) for both F12 and Sp13. The first contingency table was created using the entire sample set from each term, whereas the second table was pulled from a dataset consisting of students whose ESL writing placements were determined by two unanimous ratings from each term. Tables 5.21 and 5.22 present the results of cross-tabulation of the F12 samples in relation to the variables in question. As the results in these tables indicate, approximately 40% of the students who passed the EPT essay writing test in
F12 scored below 26 in the academic grammar test. Similar tendencies are also observed in the samples from the Sp13 testing as shown in Tables 5.23 and 5.24.

**Table 5.21.** Contingency Table I on EPT_WRT_Pass and Cut26 (F12 – All)

<table>
<thead>
<tr>
<th>Cut26</th>
<th>EPT_WRT_Pass</th>
<th>Pass</th>
<th>NonPass</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>103</td>
<td>85</td>
<td>188</td>
</tr>
<tr>
<td>26 or Above</td>
<td>% within EPT_WRT_Pass</td>
<td>59.2%</td>
<td>22.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>18.5%</td>
<td>15.3%</td>
<td></td>
</tr>
<tr>
<td>Below 26</td>
<td>Count</td>
<td>71</td>
<td>298</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td>% within EPT_WRT_Pass</td>
<td>40.8%</td>
<td>77.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>12.7%</td>
<td>53.5%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>174</td>
<td>383</td>
<td>557</td>
</tr>
</tbody>
</table>

**Table 5.22.** Contingency Table II on EPT_WRT_Pass and Cut26 (F12 – Unanimous Ratings)

<table>
<thead>
<tr>
<th>Cut26</th>
<th>EPT_WRT_Pass</th>
<th>Pass</th>
<th>NonPass</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>66</td>
<td>49</td>
<td>115</td>
</tr>
<tr>
<td>26 or Above</td>
<td>% within EPT_WRT_Pass</td>
<td>58.9%</td>
<td>21.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>19.2%</td>
<td>14.2%</td>
<td></td>
</tr>
<tr>
<td>Below 26</td>
<td>Count</td>
<td>46</td>
<td>183</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>% within EPT_WRT_Pass</td>
<td>41.1%</td>
<td>78.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>13.4%</td>
<td>53.2%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>112</td>
<td>232</td>
<td>344</td>
</tr>
</tbody>
</table>
Table 5.23. Contingency Table I on EPT_WRT_Pass and Cut26 (Sp13 (Paper-based) – All)

<table>
<thead>
<tr>
<th>Cut26</th>
<th>EPT_WRT_Pass</th>
<th>Pass</th>
<th>NonPass</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>26 or Above</td>
<td>Count</td>
<td>22</td>
<td>24</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>% within</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPT_WRT_Pass</td>
<td>46.8%</td>
<td>18.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>12.3%</td>
<td>13.4%</td>
<td></td>
</tr>
<tr>
<td>Below 26</td>
<td>Count</td>
<td>25</td>
<td>108</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>% within</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPT_WRT_Pass</td>
<td>53.2%</td>
<td>81.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>14.0%</td>
<td>60.3%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>47</td>
<td>132</td>
<td>179</td>
</tr>
</tbody>
</table>

Table 5.24. Contingency Table II on EPT_WRT_Pass and Cut26 (Sp13 (Paper-based) – Unanimous Ratings)

<table>
<thead>
<tr>
<th>Cut26</th>
<th>EPT_WRT_Pass</th>
<th>Pass</th>
<th>NonPass</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>26 or Above</td>
<td>Count</td>
<td>15</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>% within</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPT_WRT_Pass</td>
<td>53.6%</td>
<td>14.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>14.4%</td>
<td>10.6%</td>
<td></td>
</tr>
<tr>
<td>Below 26</td>
<td>Count</td>
<td>13</td>
<td>65</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>% within</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPT_WRT_Pass</td>
<td>46.4%</td>
<td>85.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>12.5%</td>
<td>62.5%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>28</td>
<td>76</td>
<td>104</td>
</tr>
</tbody>
</table>

These results suggest that test takers’ ability in grammatical writing ability may often be neglected or considered less important in EPT essay rating. Another explanation of these tendencies could be that students tend to rely on grammatical features they are comfortable in
using to maintain grammatical accuracy in their writing, particularly in a test setting. In either case, use of the academic grammar test reveals an uncomfortable fact that a considerable number of students who have been exempt from the requirement of ESL writing courses in the past may have actually needed further assistance to improve their productive grammatical writing abilities.

5.2. Qualitative Evidence for the Validity Argument

As presented earlier, most pieces of evidence for the assumptions behind the warrants for the inferences pertaining to score interpretations (i.e., Evaluation, Generalization, Extrapolation, and Explanation) could be sought through quantitative analyses of data collected in the two-time EPT administrations. Meanwhile, qualitative investigation methods were also needed to pursue backing for the warrants for the Utilization and Ramification Inferences concerning the use of scores. Students’ performance on the test also had to be qualitatively examined to discover evidence for the first assumption underlying the Evaluation Inference that examinee performance on the academic grammar test should not be affected by field-specific content knowledge. Evidence presented in the following is excerpts from transcripts of students’ performance on the test collected using think-aloud protocols or those of comments made by participants in the interviews with the investigator.

A few notation schemes of the transcripts should be introduced before proceeding to the main discussions about the yet-to-be answered research questions. The transcript excerpts below are in italics, and participants’ comments, which are illustrative or highly relevant to the topic of the discussion, are marked in bold. Words or phrases were underlined when the participants placed emphasis on them by changing prosodic features like tone and
amplitude. Unclear words or phrases were inserted within parentheses (e.g., *(It sell)*). Lowercase X’s surrounded by parentheses (e.g., *(xxx)*) are used to mark inaudible or incomprehensible words or phrases, and *(…)*i*icates omission from the transcription irrelevant or redundant remarks made by the participants. Extra information needed for the clarification of participants’ remarks or output is provided using square brackets ([ ]).

Finally, truncated words are marked with a short dash (-).

5.2.1. **RQ1: Do the test tasks on the academic grammar test require any field-specific knowledge of test takers? Does the test properly elicit evidence of examinees’ grammatical ability? (Warrant 2)**

Three individual sessions of think-aloud protocols were held with two Korean-speaking students and one Chinese-speaking student in order to investigate this question by examining ESL students’ cognitive processes that take place during test-taking. Observations of the participants’ think-aloud performances and examinations of their transcripts did not reveal any concrete counterevidence that the academic grammar test items required technical background knowledge to answer them.

One participant (P3) identified the topic or genre of a given text, but this recognition did not lead to further consideration of the topic in question. Rather, he instantly directed his attention to grammatical aspects of the given text and restricted his reasoning to them regardless of the topic of the text at hand. Examples below show occasional incidences in which one of the participants of the think-aloud protocols (P3) identified the topic or genre of a given sentence (as a test item).
This question is **about the study of plants**. “plants make their own food through a chemical process.” This question is about the usage of “know.” Is it “knowing” or “knowned?” (In response to a gap-filling question using a given key word: Plants make their own food through a chemical process ________ photo-synthesis. [know])

I think Question 7 is easy. **It is about the rise of energy cost.** It should be in simple present tense. (In response to a gap-filling question using a given key word: Energy costs _______ in real money terms since the early 1970s. [rise])

This, “all geological have proceeded,” **Geology.** This may be very slow. In this sentence, the subject is “it” and the verb is “run.” (In response to a gap-filling question using given key words: __________ more slowly, all geologic activity would have proceeded at a slower pace. [it, be, run])

This sentence sounds a little **academic.** “Tell” is to tell something. I am sure a verb should be there. (In response to a jumbled word order question: Talking about communication does not __________. [why, tell, happening, us, is, the communication])

Another potential source in the test item which may demand recalling technical knowledge to solve grammar questions is lexical items. For instance, the word ‘photosynthesis’ in Question 5 on the academic grammar test could be considered a relatively scientific term, although students who have had a K-12 education would know what it refers
to in their native language. A lack of vocabulary knowledge that led to an incorrect answer was witnessed in Participant 3’s think-aloud performance as the following:

“plants make their own food through” this process. It may be “know(n) by.” Wait. In that case, it means the plant makes food through some chemical process. **There is a “photo-synthesis” in the sentence – according to a synthesized photo. It should be the passive form of a verb. It should be “know(n).” That is, through photo theory, we can see that plants make food for themselves through a chemical process.**

He apparently did not know the exact meaning of the term ‘photosynthesis.’ He might have interpreted the stem ‘photo-’ in the word in terms of a ‘photograph’ or ‘picture’ instead of its actual meaning ‘light,’ given his interpretation “according to a synthesized photo.” In the end he understood the term as a type of theory and entered ‘... know by (photosynthesis)’ as his answer to mean “through photo theory.” (The intended correct answer to this question is ... known as (photosynthesis).)

On the other hand, Participant 1, who also struggled with the term ‘photosynthesis’ in handling the same question, though not as much as did Participant 3, ended up with the correct answer:

‘Plant make their own food through a chemical process, photosynthesis’ and what is given-, non- [apparently incorrect pronunciation of ‘know(n)]. At a first glimpse, I think a past participle should be used here, so I wrote the past participle form of ‘know(n),’ because plants make our food through chemical processes and those chemical processes are light-, that photo-, I’m not sure if it means ‘photosynthesis [in Korean],’ anyways, known as something that is combined, made by it, so what follows should modify ‘chemical process.’ So if the verb that should be entered here
becomes an adjective modifying (something). pa-, ah, it becomes an adjective if it is transformed into a past participle, and it should modify what precedes it (...).

Although she was uncertain about the exact meaning of ‘photosynthesis,’ she regarded the term as the appositive to a noun phrase ‘chemical process’ and, thus, was able to correctly complete the given sentence by entering ‘known as’ in the blank of the sentence.

Revisiting Participant 3’s response to the same question, considering it along a similar vein, his conclusion to the answer appears to be guided by ‘know(n) by’ that struck him during the reasoning. Given the fact that he even did not consider a different possible collocation ‘known as,’ he might not have learned the grammatical point that, depending on the intended meaning, different prepositions like ‘as,’ ‘to,’ or ‘for’ as well as ‘by’ can be associated with ‘known.’ In addition, he apparently did not fully understand the meaning of the question sentence, “Plants make their own food through a chemical process,” because if he had, he would have realized that it did not make sense when connected to a phrase meaning ‘through photo theory.’ His reasoning to his answer was possible as he unconsciously added “we can see that” at the beginning of the given sentence. In other words, his output ‘know by (photosynthesis)’ meaning ‘through photo theory’ functioned as an adjunct of a verb ‘see’ in the new superordinate clause he created, instead of serving as an appositive to ‘chemical process’ as intended in the original question sentence. Therefore, participant 3’s poor performance on this question could be attributed to the lack of knowledge in academic vocabulary. (In the target question sentence, the two key words, ‘chemical’ and ‘process,’ are the only academic words and the remaining words (i.e., ‘plants,’ ‘make,’ ‘their,’ ‘own,’ ‘food,’ ‘through,’ and ‘a’) are all K1 words (Cobb (n.d.); Heatley, Nation, & Coxhead (2002)). Performance on this question may thus be affected by
the test taker’s depth of vocabulary knowledge. Retrieval of particular technical background knowledge was not witnessed in any participant’s problem-solving procedures for the other test items, either. The evidence so far presented supports two assumptions (1) that the tasks on the academic grammar test do not require field-specific knowledge for their solutions, and (2) that the test tasks target and properly elicit students’ grammatical ability in academic writing.

5.2.2. RQ3: Does the decision-maker appreciate the value of grammatical ability in academic writing when making decisions on ESL placements/exemptions? (Warrant 6)

An interview with the coordinator of the English Placement Test revealed that he, in fact, weighed grammatical ability more than other aspects of academic writing, such as cohesion and development of ideas. His belief on the importance of the educational value of grammar instruction and students’ grammatical writing abilities was grounded in the literature of the ESL field, as well as personal experiences with those who had little contact with non-native English speaking students:

*I may be an outlier, but I attribute a great amount to grammatical knowledge. Being able to write well, using grammatically correct sentence structures and things like that, I attribute a very high value to-, (...) higher than organization or coherence, development of ideas. Because I firmly believe that folks get hung up on grammatical problems. It’s difficult for them to see past the grammar mistakes. And I look back at the error severity studies R. Vann and others have done, and some are just judged very severe. So if you make a mistake outside an ESL setting or somebody who deals*
with ESL students, you have a very difficult time getting a good grade or meeting the
demands if it's not grammatically correct.

The EPT coordinator’s reaction to the relatively high numbers of False Positive
students (refer to Tables 5.15 through 5.18) was also consistent with his belief in the value of
grammatical ability in academic writing. Presented with a simple version of the results in
Tables 5.15 and 5.17, he seemed to be frustrated to see the high numbers of passes in EPT
writing, whose performance in the academic grammar test had not reached the average level
of the performance of the students who had passed the writing part of the EPT exam:

"I think it's a little high because it would be an additional 71 people that we currently
have as a Pass that would not pass."

where the first ‘it’ refers to the number of False Positive (i.e., 71) in Table 5.21. Therefore,
the second assumption underlying Warrant 6 for the Utilization Inference is supported by
witnessing the EPT coordinator’s strong appreciation of the importance of grammatical
ability in academic writing.

5.2.3. RQ4: Could academic grammar test scores serve as a tangible ground for
decisions on individual students’ ESL placements/exemptions? (Warrant 6)

Under the current ESL writing placement system where decisions are made merely on
the basis of students’ performance in a single-essay writing test, it is often challenging to
make decisions about whether to place a student in an ESL writing classroom or not,
especially when he/she is a borderline case. The use of the academic grammar test as a
supplementary measure to the essay test could certainly offer a tangible ground for the
decisions on ESL placements/exemptions because it would supply the decision-maker with additional information on test takers’ grammatical writing abilities. Moreover, the grammar test scores could serve as palpable evidence for appropriate placement decisions as the EPT coordinator hoped:

> It would tremendously add to the value of the placement test. (…) We would be able to better convince learners that they need to take a class as opposed to having two raters and say, “okay, you’re in 101B.” It’s like “show me why.” You know what I mean? We don’t get a lot of complaints by and large. But we do get some. And for those it would be used, “hey look, this is where you’re at. When they come and ask about their listening or reading, I tell them the score; they’re happy. They are not happy, but they accept it. They have a hard time, some of them, (…) accepting placement decisions that are made holistically.

5.2.4. **RQ6(a): If the scores on the academic grammar test were used to make decisions on ESL placement, how would students perceive its impact on them with respect to advantages and disadvantages of additional ESL instruction? (Warrant 7)**

To understand students’ perspectives on the use of academic grammar test scores as one party of stakeholders of the test use, the investigator had individual interviews with a total of 31 students (16 from the ESL group and 15 from the EPT Pass group). Among them, twenty-three participants were asked a question about possible benefits of grammatical instruction for students that may be caused by taking into account the academic grammar test scores, in addition to the EPT essay test results, for ESL placement/exemption decisions. Fourteen of the 23 respondents appreciated its benefits to varying degrees (e.g., ‘50%
helpful’ to ‘always helpful’). When asked about possible negative impact of using the academic grammar test scores in addition to the EPT writing for making ESL placement decisions, they addressed the issue in terms of the following aspects:

- Time issue (‘waste of time’ or time constraint),
- More tuition to pay,
- Extra work to do,
- Taking more credits that do not count towards graduation,
- Delayed graduation, and
- Disappointment in themselves.

As a follow-up, 11 of these 14 respondents were asked for their opinions about the degrees of impact of the two aspects of the test use on students. Seven of these 11 respondents (63.6%) believed that benefits of extra grammatical instruction brought about by the use of the academic grammar test scores on non-native English speaking students would outweigh potential disadvantages, which students may face in their academic paths. Below are some of their comments in response to the questions about the positive impact of the test use and their differences in weight:

_The benefits I would say, ‘cause like you know, you just need to know how were you are in a English, you know, how were can I just here, so if you get good score, that means your transition here will be easier, but if your English is not very good, then you just need to fix that one (xxxx) for you, taking other classes and, you know. So, I think it’s like more benefits than negative impacts._ [P-U07]
The benefit is to enforce you to learn some-, the more grammars in English, and I think this is really helpful for your future no matter spoken English or written English, especially if you want to have a job in academics. And their adverse part for you is a waste of time, but I don’t think that is that big waste of time. So I think compared to the benefits you can get (them in) for your rest of life, I think the adverse influence on your life is not very big. [P-G05]

It’s definitely beneficial, because (…) if a student is grammatically not correct, it will be reflected in his essay also. (…) If he does pass in essay and still fails in grammatical test, I think he’s, he still should take the 101 class if that teaches grammar. Because broken English or grammatical wrong English doesn’t sound professional, so for graduate students, you need to present yourself at conferences, and there are people from all of the world listening to you. (…) as well for writing. (…) Negative impact? (…) it’s their mistake. They are not proficient in English, so it’s their extra work they have to do. So it shouldn’t be counted as a loss or (truly) counted as a disadvantage. [P-G07]

I think it’s may better for people who fail this[grammar] part to take 101B. I think it’s necessary, and it’s useful. Somewhat. Because you will write many articles, assignments, in your future study in ISU, and if you have a poor grammar, the instructors, professors may misunderstood some of your assignment homework, and I think it’s better to take 101B. [101-05]
Interestingly, those participants who had not passed the EPT writing test tended to perceive taking an extra class for additional instruction on grammatical writing ability in a more negative way than those who had passed the placement test. This phenomenon may be explained in that high proficiency students have developed positive attitudes toward ESL instruction through their learning and improvement in English as a target language. Some of the lower proficiency students’ doubts on grammatical instruction were, in fact, based on their experience in previous grammar classes such as ENGL 101B or a grammar class offered by the Intensive English Orientation Program (IEOP) at Iowa State University. Even some positive remarks on the benefits of taking a grammar course were lukewarm at best:

> If 101B really helped me improve my grammar ability, I willing to take it, but if like the class I take last semester [in IEOP], I will feel useless. (I: Can you describe the class for me? The grammar class you took in IEOP?) It’s like, you got a grammar for, like grammar law, and there are a lot of sentence for you, and you write down like the similar way, and you try to memorize how it is correct way to write the grammar. But you didn’t, I think, when you want to say something about, or you want to write something about, you won’t really use it. [101-02; placed in ENGL 101C]

I think those things in 101B is pretty simple. Just repeat those things I learned probably in high school. (...) I just want to take class like 101C to learn some academic writing skill. As for grammar, I have learned a long way just those things you need to repeat again and again, so I need little time to use some in my daily life, in my writing. Probably I can correct them. Don’t need to spend as a time to learn
them again. **I know them. I just don't know how to use them.** [101-10; placed in ENGL 101B]

*I think it*[taking 101B]* helps. Although it just recall what I have learned, because in my preparation class, I only took at once level of writing, and here I need to go over the 101C, but then it just like recall what I learn before because those English, which I can’t transfer here, so I need to take another ones on writing here, so it just makes me proper in writing, so it helps although it's not as much as I hope, but it also help.* [101-09; placed in ENGL 101B]

In regard to this, nine among the ten interview participants who had been placed in the lower-level ESL writing class (i.e., ENGL 101B) were asked about their experience in the course. As a result, only three expressed strong, positive opinions about the course. Five of the remaining students thought that they had not learned much new information unlike their expectations, although the course was ‘okay’ or ‘beneficial’ in a sense that it ‘was easy to follow’ or that the class reviewed what they had already learned before entering the university.

Comments made by Student 101-02, in particular, suggested that some instructional methods still grounded in the old-fashioned, drill-based approach to language learning may have affected their perceptions about grammar instruction in a negative way. Another important issue pertaining to students’ negative impressions on grammar instruction is the opinion students expressed that their grammatical ‘knowledge’ in English, as the target language, has failed to transform to their grammatical ability as performance (as witnessed in
the remark by Student 101-10). Given the respondents’ comments, pedagogical approaches to grammar instruction might need to be revised in a way to facilitate students’ applications of target grammatical features to actual uses in order for adult ESL learners situated in academic contexts to benefit from the use of academic grammar test scores for decision-making in students’ placements of ESL writing.

In summary, while students perceived both positive and negative impacts of usage of the academic grammar test as part of the criteria for ESL decision-making, a majority of these students believed that the positive impact would outweigh the negative impact on students’ academic lives. Those who had a higher proficiency in English were more likely to appreciate the importance of instruction on grammatical ability than those at a lower proficiency level. These findings provide backing for part of the first assumption identified for Warrant 7 of the *Ramification Inference* in Table 3.1.

### 5.2.5. RQ6(b): *If the scores on the academic grammar test were used to make decisions on ESL placement, how would students perceive its impact on them with respect to the relationship between grammatical writing ability and academic writing? (Warrant 7)*

To address this research question, the investigator asked the following questions of the student interview participants:

> What do you think about grammatical ability in academic writing? How much do you think it is important in academic writing?

In response to this question, 27 of the 30 respondents who were asked the question believed that grammatical ability was ‘important’ or ‘very important’ in academic writing. Following
are some comments from participants, who expressed a strong opinion about the importance of grammatical ability in academic writing:

*Very important. (…) For each language, the grammar is the rule, how the language you can use, and if you didn’t follow the rule, you have a very poor grammar skills and your-, no matter what’s you write or say, (there) may have many mistakes and faults, and which may misunderstood by others and sometimes may effect your efficiency of working and, yeah. [101-05]*

*Undoubtedly it’s pretty much important because it’s-, it might change the whole meaning of the sentence. [101-07]*

*I do think that’s very important, because without correct grammar, then your essay, your writing will be sometimes will be very misleading or vague the ideas. [101-08]*

*Eight [out of 10]. (…) The right grammar can reflects our real thoughts when we write about the essays or papers or journals, so, and I use my native language to write about a story or essay, the others the readers can easily understand what do you thought, what do you want to showed. But if I write something with incorrect grammar, maybe the reader will confused, maybe will affect my score. [101-13]*

*I think it’s important (…) like 8 [out of 10] (…) because sometimes it may confuse others when they reading your writing. Then, maybe, you are not suppose, like, want*
to say something like that, but they just thought that’s your thoughts. Then, that’s just a problem. [101-16]

I actually think it’s really important. Like I said, if you don’t dress your essay well, a lot of times the teacher just can’t stand reading it. Honestly, I’ve been a tutor and a teacher of the English language myself, and when I see essays of younger students, I just can’t really stand it because it’s grammatically wrong, and even though they have some good ideas there, but a lot of times if they address wrong mistakes everywhere, you just tend to think the whole thing’s a mistake. [P-U06]

As shown in these quotes, students perceived the importance of grammatical ability in writing in relation to the effectiveness of communication. They were concerned that their message would not be clearly or effectively delivered to the audience if their writings were grammatically flawed. Some graduate participants stressed the importance of grammatical writing ability, particularly with respect to the publication of papers in academic journals:

*It is important, and every time, weekly we send our report to a professor, he reads through it, and like makes all the grammatical mistakes or ask us to structure the thing or how can I explain the engineering, so, but for publishing a paper, it’s really important, and there are no grammatical mistakes and the structure is really good.*

[P-G03]

*At least in my case I have to read a lot of papers and you can tell when there some papers are published in Europe, for example, like Spain or Germany, where their*
grammar is not completely accurate, and sometimes it’s hard to understand what they are saying. So I think if you are to be publishing in English, then you should really have-, you should really know what you’re doing at least with the grammar part. [P-G04]

Meanwhile, four students believed that their insufficient grammatical writing ability could be compensated for by help of online or offline resources available around them:

*I think it’s pretty important. It doesn’t look professional if you make grammar mistakes. Well, but there’re resources and people that are able to help you. If I needed help, I would obviously go to them and make sure that everything is okay.* [P-U03]

*But we do have auto-correction for writing. Because nowadays I don’t think, like, most of our transactions are online, and they are (viewed xxxx of) the world probably. So as far as these softwares are concerned, they correct you grammatically even if you are wrong. And, so you can just right-click and just see what is correct and just interact. You don’t need to write correctly, even the spelling mistakes they are recognized. So, for writing, I guess you can always use some online softwares which will not, like, which will help you (your) grammatical errors, so that does not mean that you should not stress on writing grammatically correct things, but the emphasis is not much as it will be on speaking because, in speaking, once you say it, you cannot correct it again. (It sell.) People will judge you based on what you are speaking, but for writing, before sending a document to someone, you can always see that it is*
grammatically correct. It’s taken care of by the software itself, so. For today’s world, I think, in the electronic world, you don’t need to bother about being grammatically correct in writing stuffs. [P-G07]

For grammar, there are some software application that can help us to, like the Criterion, they will have parts to reduce the mistakes about the grammar. [101-01]

It’s very important, but nowadays [Microsoft] Word corrects grammar. I know it’s not perfect, but it’s kind of work. And for me, I feel them in problem not in, in the grammar, like shaping this sentence or how to explain your idea with simply way. [101-11]

In a few other cases, students did not believe that grammatical writing ability should be important to them, since their academic disciplines, such as chemistry and mathematics, did not require much writing:

In my math, the only part’s like application question, we can clearly understand what the problem state and what the question want to explain the situation or (xxx), and in chemistry, I think it’s like a lot of (specific) noun and you should memorize it and after that you will be fine, I think. So I think grammar doesn’t matter. [101-02]

I read some math papers, and I think in those papers we really do not need so much grammar, you know. We just need to know the basic grammar, and we just need to say “because”, “and”, something like that, and just arrange them. But I think in the
abstract we need some, some grammar there, but not too much grammars. So we
don’t need too many long sentence and how to arrange them. So I don’t think we
need a lot of grammar like expected. (...) We have a lot of symbols, so we have a lot
of math symbols, we just write those symbols, we don’t need, we don’t involve any
words (...). [P-G02]

Students’ perceptions about the importance of grammatical ability thus appear to be
closely associated with the extent to which writing is required of them in their field of study.
However, some participants differentiated the value of grammatical ability in a general
domain of academic writing from that in their own academic discipline:

In our major[architecture] maybe not a lot, because you have to explain the idea and
they get it somehow. But in general, I think it is very important because the only
matter you have to explain is your words and they have to be accurate. I think
grammar is very important. [101-14]

It’s very important, but nowadays Word corrects grammar. (...) For me, I don’t. For
the grammar, I can correct it, I can review my paper, I can search for every word and
see sentence by sentence, its subject, verb, and, what tense, etc. But for me, it’s more
than how I can explain my idea by English very easy, especially with-. Our
department[Regional Planning], all-, it’s not like science, it’s all writing about new
idea, new research problem. [101-11]
One way to explore the relation between students’ grammatical writing ability and their academic success at a U.S. college is to examine whether students’ performances on course assignments or exams are penalized due to deficiency in grammatical writing ability. While a majority of the interview participants perceived grammatical writing ability to be (very) important in academic writing, they mostly reported that their performances in course assignments or exams had not been necessarily affected in an adverse way by their deficiency in grammatical writing ability. Many believed their instructors or professors tended to ignore grammatical errors and only focus on the contents of their writing:

*My instructor just take a look and grade it by the content, but not by grammar, so I think it doesn’t matter my grade.* [101-02]

*In this semester, I took the sociology class, it’s okay, because I don’t know, we just think about, but when I writing some, the, papers, and the writing, I think it’s okay.* [101-13]

*So I need to write a lot of reports, but, um, I think I’m, um, a score of my reports are no bad. I don't know why. (...) Maybe for science major, like reports, maybe my teaching assistant, he just didn’t care about our writing skill.* [101-15]

On the other hand, there were a couple of cases reported in which students had experienced that their poor grammatical ability adversely affected their performance in academic success. While some instructors provided them an opportunity to revise their assignments for grammatical improvement or a warning of a potential downgrade in advance,
a few others were apparently intolerant of non-native English speaking students’ lack of grammatical writing ability:

> I took speech communication last semester, I wrote one paper, and the professor, she point out my mistakes during the- in the paper, and there were some grammatical mistakes, she will have- she would have me to (point it) and let me change it. [101-01]

> I haven’t done with the, my paper alre-, but I talked to my professor. Because we have to write on, in a blog every week. It’s short, it’s like a paragraph or something. I talked to my professor, and she said “it’s okay because I know you are new here. But for your paper, that you have to give it in at the end of semester, you need to be better than what you write right now.” So, I think, yes, it’s my grammar is going to, that’s going to affect my grade. [101-14]

> Sometime I have a comment, you can have-, I can have comment my paper said, “I can’t understand, I understand your writing, you should work with writing, you have some grammatical issue.” I have-. Not from all the professor, but there is one always gives this comment, I don’t know why, (xxxx) maybe other understand we are not first language, and they just grade the idea or assignment. Yeah, but some, it’s (happen) with some professor. [101-11]

> Last semester, I got, uh, Design 183, it’s a design culture class, and, uh, one of our major assignment is to write your ideas about how to design a monument. And, our group, because we are all international students, we only got B. And when we talked
to the professor, he said, the grammar errors he on-, he also like, the- (for that), so that’s the reason why we got B. [101-16]

In summary, a majority (90%) of the students believed that grammatical ability is important for academic writing in a general sense, although its importance to individual students varied depending on the degree and characteristics of writing required by their own academic discipline, which also supports the second assumption of Warrant 7 of the Ramification Inference. The extent to which university instructors have prior experience with non-native English speaking students and their leniency on these students also seem to be potential variables that determine whether students’ grammatical writing ability affects their academic performance in a U.S. college. Although only a few cases (four respondents) were witnessed, the fact that students’ performances in some courses has actually been adversely affected because of grammatical mistakes they made in written assignments implies that the grammatical aspects of academic writing should not be neglected in making decisions on ESL placements/exemptions for students’ benefits in the long term.

5.2.6. RQ7a: How much do ESL instructors appreciate the value of grammatical ability in academic writing? (Warrant 7)

Two separate focus group meetings were held with six ESL writing instructors at ISU to investigate ESL instructors’ perspectives on the values of grammatical writing ability and pedagogy in this ability. Four instructors participated in the first focus group (Instructors A through D), and two in the other group (Instructors E and F). These meetings revealed that the curricula of these courses weighed students’ grammatical writing ability differently from
each other. For example, the lower-level ESL writing course (i.e., ENGL 101B) weighed 30% of grammatical ability for the final grade as the primary intention of the course was to improve students’ grammatical writing ability, whereas the instructor of the upper-level ESL writing course targeting graduate students (i.e., ENGL 101D) considered grammar in grading only for 10%. The upper-level ESL writing course for undergraduate students (i.e., ENGL 101C) attributed 25% of the students’ final grades to their grammatical writing ability (See Q2 in Appendix D). On the other hand, most instructors personally considered more intensified instruction on grammatical ability to be important or necessary, especially when they were teaching undergraduate students:

For me, I wish there was a grammar test and for those students who failed the test can go through a grammar course. I really think it is necessary, and I think 101B is not enough, because 101B is too compact, it’s too intense. A lot of grammar points compacted in just one semester (...). I would want my students to have more chances. [Instructor C]

Some students are still struggling with how to write a complete sentences, and then logically to develop the essay. (...) I really want to sit down with my students individually to talk about their grammar thing rather-, maybe also content as well, but grammar, I mean, I need to talk about that as well. [Instructor B]

As an instructor, for about 2 years, I found that 101B students’ level kind of-, level’s getting worse. Yeah, compared to the previous years, I don’t know why, so that’s why these days I always focus on the grammar one by one. So, for example, as during the
individual conferences, as long as students fill the basic requirement in terms of content, our conversation is always focused on the grammar, because there are very, very serious grammar mistakes all around students’ essays. So that’s what happens. [Instructor A]

I think it[instruction on grammatical knowledge] is necessary. (…) There are two types of students coming to class. Those who know grammar, they know the name of the concepts, like compound sentences, relative clauses, or adjective clauses. But they don’t know how to use in writing when they produce. And other group is smaller but they don’t know this even. So we have two types of students deal with, I think, in 101C classes, and you cannot ignore this as an instructor. And the same time we don’t want to change the class to grammar class. There’s another. I’m trying to create that balance. [Instructor E]

These remarks of the instructors suggest that they often struggle to balance between students’ grammatical writing abilities and other writing skills from a discourse perspective (e.g., organization skills and development of thoughts). Yet, many of them are concerned about their students’ limited grammatical abilities and wish to offer them extra help to improve grammatical writing ability, if needed. Therefore, the participating instructors of ESL writing classes appreciated the value of grammatical ability in academic writing, although to varying degrees, which provides additional backing for Warrant 7.
5.2.7. RQ7b: To what extent could ESL instructors benefit from feedback on their students’ grammatical abilities provided by the use of the academic grammar test? (Warrant 7)

The current ESL placement system at ISU lacks instant feedback on students’ grammatical writing abilities. Instructors of ESL writing courses should wait until they collect diagnostic writing samples from their students on the first day of class to assess students’ needs in terms of grammatical writing ability. Such evaluations are usually impressionistic, as acknowledged by a focus-group participant:

I guess I don’t feel confident, maybe you guys do, that I’m necessarily prioritizing the grammar points that are most important. I don’t know. I’m prioritizing the ones that are most salient to me, or better at the beginning of the paper, or the pattern of a particular paper, but I’m not always sure that these really are the ones that are the biggest problems. [Instructor D]

As such, most instructors participating in the focus-group interviews indicated that they would appreciate feedback before the beginning of a semester, which classifies the types of grammatical features students have difficulty in using or producing and presents the percentage of errors students make for each grammatical feature, preferably with visual aids:

If it were categorized like I said and I had this actual student mistakes, it would help me. (...) If it were formatted in a way where it was clear exactly what students need help with, it would be awesome. [Instructor D]
If you give us like the percentage and the type of feedback, the type of errors the students make then, we can think of the grammar instruction first. [Instructor B]

I would like to know the most common kind of error that they make and they type of it. (...) I would like to see the, you know, kind of chart, maybe like results where you have the type of error and how much they succeeded on that area. [Instructor E]

From such classified feedback on students’ performance on the academic grammar test generated by an automated scoring system, the instructors expected benefits of timely diagnosis of students’ weaknesses in grammatical writing ability and early adaptation of their course plans to meet students’ needs in the target skills, which would be systematically identified by an automated scoring system. In other words, feedback generated by the system could facilitate instructors’ needs analysis for their classes and also allow them to provide personalized instruction corresponding to their students’ needs to some extent:

I can also change the teaching schedule based on the information from your automatic, EPT grammar test. (...) I think your EPT test can provide instructor most specific errors and most specific information about students’ grammar performance, that will definitely help teachers to organize and then prepare related exercises. [Instructor A]

I think it would be nice for us as an instructor to see, okay, the need. It’s kind of need analysis to see how much emphasis we should put on grammar or not. (....) That’s going to give me the nice information about where to put focus, where to start even,
because it might not fit to the order of the book, I mean, what they need. And then we shouldn’t assume that order of the book is exactly what they need, you know. That’s going to be the nice thing for us to personalize the instruction, I think. [Instructor E]

Another important benefit that the feedback of an automated scoring system of the academic grammar test could provide for instructors is that it could help them identify students’ weaknesses in grammatical abilities, which may not surface when the essay test is the only measurement of students’ grammatical writing abilities:

Just as I mentioned, they avoid using some difficult one. But in this one[grammar test], they’re focused to do that to show their understanding and their mastery of this structure. So I think it’s a good way for us to make a decision. [Instructor F]

In fact, this could be a convincing explanation for quite a few discrepancies in students’ performances between the EPT essay test and the academic grammar test, as witnessed in the results for RQ 5 in the previous section of quantitative evidence for the validity argument of the score interpretation and use of the academic grammar test.

To summarize, the foregoing results of interviews with different stakeholder groups pertaining to the English Placement Test at ISU suggest that they could benefit from use of the academic grammar test scores for the purpose of decision-making in non-native English speaking students’ placement in ESL writing classes. For example, the EPT coordinator could make more informed decisions regarding students’ needs for additional instruction in ESL academic writing. Instructors of ESL writing courses could identify students’ needs in terms of grammatical writing abilities and adapt their course design to the students’ needs in a timely manner. Non-native English speaking students might have another opportunity to
improve their grammatical writing ability. Some potential, major adverse effects of the use of the academic grammar test scores on ESL students as stakeholders include (1) delay in their program of study towards graduation, (2) increase in workload, and (3) increase in educational expenses. However, positive impacts on the use of these test scores could be greater on students than the negative impacts in the long term, because students could use extra systematic help to improve their grammatical writing ability, which is often considered a necessity for academic success in a context where English is the only language of instruction.

5.3. Chapter Summary

This chapter presented the results of the analyses of both quantitative and qualitative data collected in F12 and Sp13 to investigate the research questions arising from the assumptions underlying the warrants of the inferences identified in the interpretive argument for the development and evaluation of the academic grammar test (Chapter 3). Most of the quantitative results centered on questions concerning the observation of test performance and interpretation of scores (Warrants 2 through 5). The qualitative results addressed issues pertinent to the use of test scores for the intended purpose and its potential impacts on ESL students and instructors of ESL writing courses at Iowa State University as the immediate stakeholders of test use (Warrants 6 and 7), except for the data from the think-aloud protocols collected for Warrant 1 for the inference of Domain Description.
CHAPTER 6. DISCUSSION AND CONCLUSION

In the previous chapter, quantitative and qualitative analyses of the data collected in F12 and Sp13 were conducted to seek backing evidence for each assumption underlying the warrants for seven inferences identified in the interpretive argument constructed for the interpretation and use of the academic grammar test scores for ESL placement decisions (see Table 3.1). This chapter summarizes the findings of these investigations in relation to each inference and establishes a validity argument for the score interpretation and use of the academic grammar test for its intended purpose—supplying relevant, reliable, and palpable evidence for decisions made on students’ placements in, or their exemptions from, ESL writing courses at Iowa State University.

6.1. Building a Validity Argument

The validation framework chosen for this study included aspects of Bachman and Palmer’s (2010) Assessment Use Argument (AUA) model, while the main skeleton of the framework was based on Chapelle et al.’s (2008, 2010) argument model which was also derived from Kane’s (2006) interpretive/validity argument model. By taking Bachman and Palmer’s (2010) AUA approach, an interpretive argument was constructed beginning with the Ramification Inference before embarking the development of the academic grammar test in order to ensure the intended consequences of the test to be borne in mind throughout test development. Following the same AUA approach, the validity argument for the evaluation of the score interpretation and use of the academic grammar test will now begin with the inference of Domain Description, since the argument pertains to evaluations of test development, implementation, and use for the intended purpose.
6.1.1. Domain Description Inference

Warrant 1: Test tasks contain the language of the target domain of academic English.

The underlying assumption of this warrant was that the linguistic nature of the tasks in the academic grammar test represents those for academic English. To ensure that this assumption is satisfied, target sentences of the test items were adapted from actual introductory-level college textbooks in academic fields, such as biology, sociology, and communication. Also, as described in the chapter of Methods, Biber et al.’s (1999) *Longman Grammar of Spoken and Written English* (LGSWE), a collection of corpus-based extensive descriptions of English grammar, was referenced to ensure the target grammatical features of the test reflected the linguistic characteristics of academic prose. As a result, the feature ‘modal + present perfect’ (as in expressions like ‘should have apologized’ or ‘could have chosen’, usually used to express regrets or an alternative possibility about a past event), was not chosen for candidate features of the test items among those grammatical features used in Chapelle, Chung et al.’s (2010) study, because its use was determined relatively rare in academic prose compared to other discourse genres according to Biber et al. (1999). Given these attempts, it can be safely argued that the test tasks properly reflect the characteristics of the language utilized in the target domain of academic writing in English.

6.1.2. Evaluation Inference

Warrant 2: Observations of examinee performance are evaluated to provide observed scores informative of examinees’ task performance.

Two assumptions were identified in regard to this warrant: (1) no requirement of field-specific content knowledge in the examinee performance on the test and (2) proper
evaluation of examinees’ performance. To address the first assumption, three non-native English speaking students from South Korea and China were invited to take the test using think-aloud protocols. An analysis of their performance indicated that examinees could solve the questions without relying on particular content knowledge relevant to an academic field. One vocabulary item ‘photosynthesis’ of Question 5 (*Plants make their own food through a chemical process [know] photosynthesis*) appeared to be somewhat challenging to the participants. The uncertainty about its meaning did however not prevent them from providing an answer to the question as long as they were able to analyze the structural characteristics of the target sentence. Whether or not they answered the question correctly was determined by their grammatical knowledge on the usage of past participle and associated prepositions. The observations of their think-aloud performances also revealed that the test items effectively elicited their ability to use English grammatical features; they continuously attempted to retrieve knowledge from previous ELS learning relevant to target grammatical features.

The second assumption about proper evaluation of students’ performance on the test was examined in terms of descriptive statistics, especially dispersion of the test scores, and investigation about a possibility of the test interacting with gender as a construct-irrelevant variable. Examination of the descriptive statistics of the data collected from F12 and Sp13 showed that the observed scores spread across the range of possible scores, forming a normal distribution. The means and medians of the scores were similar across these two samples, well reflecting the fact that the test takers were from a homogeneous target population although not randomly sampled. An Analysis of Covariance (ANCOVA) also revealed that...
gender did not function as a moderating factor between the grammar test score and the proficiency level (measured in terms of whether or not examinees passed the EPT).

A design of the test’s scoring rubric also contributed to the appropriate observation of students’ performance in the academic grammar test. Instead of adopting binary (0 and 1) scoring, the rubric adopted a three-level polytomous scale (0, 1, and 2) so that examinees’ developing ability in productive grammatical writing ability could be reflected in their scores. High agreements between the two raters on scoring of the test as well as individual responses ($r = .99$ and $r = .97$, respective; $p = .00$ for both correlations) were also achieved. All evidence presented supports the two aforementioned assumptions for Warrant 2. Therefore, the validity argument of interest may continue to the next step in the chain of inferences, which is the Generalization Inference.

6.1.3. Generalization Inference

Warrant 3: Observed scores are stable estimates of expected scores in the universal domain of test tasks.

Once the observation of examinees’ performance on the test is verified, it is necessary to check if their performance would be consistent across similar tasks developed for the test. Two assumptions were identified for this warrant: (1) that the test scores have an acceptable level of reliability and (2) that test takers’ performance on the academic grammar test is consistent, regardless of the test delivery format. These two assumptions were checked in terms of the internal consistency realized as a Cronbach’s alpha coefficient and a comparison of examinees’ performance on the paper-and-pencil mode with that on a computerized test. For the satisfactory level of internal consistency, in particular, .7 was set as the criterion
because the test items were pilot items and some of them would need a revision after item evaluations.

Cronbach’s alpha coefficients obtained from the F12 and Sp13 datasets were both above .7 (.784 and .723, respectively), which met the expectation of internal consistency. On the other hand, the 95% confidence interval for the difference between the two coefficients did not confirm with certainty that the two coefficients were not statistically significant, as the lower bound of the confidence interval resided almost on zero. As a next step, the Cronbach’s alpha coefficients were calculated separately for the male and female examinee groups of both academic terms of data collection. The alpha coefficients for both groups were all higher than .7 as shown in Table 5.6. In addition, the difference of the alpha coefficients between these two gender groups was not statistically significant in both F12 and Sp13.

The investigation on the second aspect of consistency in examinee performance was attempted in Sp13 by administering the computer-based version of the test to a group of students who took the make-up session of the English Placement Test offered to late-arrival students. This attempt drew upon the assumption that, while the examinees of the computer-based academic grammar test were not randomly sampled from the target population, they would form a homogeneous group with those who had already taken the test on the paper because they were selected by the university programs based on the same criteria applied to the other groups of students. Unfortunately, it turned out to be inappropriate to compare their performance on the computer-based academic grammar test with that of the paper test group of the same term. The sample size of the computer group was too small to represent the target population (N=24). While the Cronbach’s alpha coefficient for this group was above
.7, the descriptive statistics of this group’s test performance were substantially different from those of the other groups. Particularly, the mean and median scores of the computer group (13.63 and 12, respectively) were much lower than those of the paper groups (around 21 and 22 for both means and medians).

It is, however, too early to determine that students’ performance on the test across different delivery modes is not consistent. Above all, the sample size of the computer group was too small for the group to represent the target population. In addition, some unknown factors related to the test delivery mode or the test administration setting might have unfavorably affected the test takers’ performance. The relatively high internal consistency (.794) of the examinee performance on the computerized test suggests a bright side of the implication as well.

The foregoing findings provide partial support for the warrant, which calls for further investigations on this issue. Especially, the examinee performance on the different delivery modes should be compared with larger samples, ideally with equal number of students. Since the assumption of an acceptable level of reliability was satisfied, the discussion could continue to examine the consistence in students’ performance of productive grammatical writing ability between testing and non-testing contexts.

6.1.4. Extrapolation Inference

Warrant 4: Examinees’ expected scores estimated by the observation of the academic grammar test represent their target scores manifested in the target language use (TLU) domain.
Two assumptions were identified in relation to this warrant. First, it was assumed that examinees’ performance on the academic grammar test would correspond to their proficiency level in English writing. For comparison analysis, the test takers were classified into two groups by their results of EPT (Pass vs. Non-Pass) and also classified into three groups by their performance on the writing portion of the EPT exam (i.e., ENGL 101B, ENGL 101C, and Pass). Mean comparisons sustained the first assumption. The three examinee groups classified by their results in EPT writing performed distinctively from each other on the academic grammar test, with 95% confidence intervals of their means not overlapping with each other. Those who passed the writing portion of the EPT, in particular, performed far better on the academic grammar test than the other groups placed in one or more ESL writing classes.

The second assumption was that students’ performance on the academic grammar test would be consistent with their performance in writing that occurs in a non-testing context. For this investigation, most of the interview participants’ writing samples (produced either at the site of interview with the investigator or on the first day of their ESL writing class) were collected and analyzed in terms of four categories of grammaticality indices on error frequency (E), error ratio (E/T, E/C, and E/W), accuracy frequency (EFT, EFC, and EFS), and accuracy ratio (EFT/T, EFC/C, EFC/S, and EFS/S). The error frequency and ratio indices were expected to be negatively correlated with students’ academic grammar test scores, given the assumption that those who scored high on the test would produce fewer grammatical errors in their writing. The same assumption led to another expectation that accuracy-related indices would positively correlate with the academic grammar test scores. The actual correlation analysis of the interview participants’ scores on the academic grammar
test with their output on those grammaticality confirmed that these expectations were correct.

Eight of 11 grammaticality indices had a moderate correlation with the test scores, with the Pearson Product-Moment correlation coefficients ranging from .41 to .68 in their absolute values. Nine of the 11 correlation coefficients were statistically significant. The numbers of error-free T-units (EFT) and error-free sentences (EFS) were the only two indices which were not significantly correlated with the interview participants’ scores on the academic grammar test.

This finding concerning the second assumption implies that what makes high proficiency students’ writing ability distinctive from that of lower proficiency students resides not only in grammatical accuracy but also in the complexity of sentence structures they produce in writing. In other words, higher proficiency students would produce more complex sentences containing one or more embedded clauses than lower proficiency students. As shown in Table 5.12, productivity—in regard to word counts and the numbers of clauses, sentences, and T-units—does not differ much between the higher proficiency and lower proficiency groups among college students when they are asked to produce a written sample within a limited time. A post hoc t-test on the productivity variables confirms this observation: \( t(25) = 1.721 \) \( (p = .098) \) for word counts; \( t(25) = .418 \) \( (p = .680) \) for the number of clauses; \( t(25) = .505 \) \( (p = .618) \) for the number of sentences; and \( t(25) = .084 \) \( (p = .934) \) for the number of T-units.

In the end, both the comparison of students’ performance on the academic grammar test and their proficiency level in English writing and the error analysis of the student interview participants’ writing samples provide backing for Warrant 4. The Extrapolation Inference is accordingly established so that the target scores (i.e., students’ performance of
grammatical writing in the target domain of language use) can serve as a ground for the next claim that the theoretical construct of grammatical English writing ability governs the examinees’ grammatical writing ability estimated and expected to be exerted in the target context.

6.1.5. Explanation Inference

Warrant 5: Target scores estimated by the grammar test are attributed to the theoretical construct definition underlying the construct of productive grammatical writing ability in English.

Now that evidence has been found to support that students’ grammatical writing ability in academic English measured by the academic grammar test is consistent across the universal domain of test tasks and the general target domain of language use, it is necessary to investigate the question of whether their performance is ascribed to the construct of interest to ensure that the score interpretations are properly made in light of the target construct. Three assumptions identified for this warrant are: (1) that examinee performance on test tasks reflects the cognitive difficulty levels of the target grammatical features in their acquisition; (2) that observables of sub-construct target grammatical features are meaningfully associated with the construct of the productive grammatical writing ability in English; and (3) the construct of productive grammatical writing ability in the academic setting is consistently exerted in other criterion measures engaging the same construct.

Investigation of the above research question was conducted in three different ways to address each of these assumptions. First, it was examined whether the items of advanced grammatical features were more difficult for test takers than those of lower-level
grammatical features, as expected by the Processability Theory. Second, possible relationships between the grammatical features employed in the test and the target construct of grammatical writing ability were modeled and tested using confirmatory factory analysis. Third, students’ performance on the academic grammar test was compared with their performance on the writing portion of the TOEFL iBT® in order to investigate whether students’ ability of grammatical writing ability is sustained, even when it may be affected by other factors playing in the measurement of writing ability. Since one of the writing tasks adopted in the TOEFL iBT® is an integrated task involving examinees’ reading and listening skills as well as writing ability (Educational Testing Service, 2014), a moderate correlation between the academic grammar test scores and the TOEFL iBT® writing scores could provide evidence that grammatical writing ability is consistently demonstrated regardless of the language use context. In a similar vein, a moderate correlation was also expected between the students’ academic grammar test scores and their results on the EPT essay writing test, which takes into account students’ skills of developing and organizing ideas, functionality, and vocabulary, as well as grammatical accuracy and complexity in language.

The findings of these investigations all provide backing evidence that the target performance of students’ grammatical writing abilities is attributed to the theoretical definition of the construct. The IF values of the test items grouped by the expected developmental stages of their target grammatical features showed a declining pattern, on average, as the target grammatical features are expected to be acquired at a higher stage of the development of English as a second language. The Spearman rho correlation analysis also showed a negative, moderate association between the inversed item difficulty and the expected developmental stages. Looking into the differences among the test items in IF
value, the mean of the IF values of Stage 4 items was significantly different from those of Stage 2 and Stage 3 items. The non-significant difference between Stage 2 and Stage 3 items in IF value suggests that some items (such as Q20, which attracted the correct answer from most test takers due to the simplicity in task design) need a careful revision. In addition, more research should be conducted in regard to the actual developmental stages of the grammatical features which are perceived to be difficult and thus are expected to be acquired in a later stage of ESL acquisition.

According to the confirmatory factor analysis results, the unidimensional model (Figure 5.6) had the best model fit among the candidate models hypothesized. The grammatical features in the model had varying degrees of association with the target construct, all of which were statistically significant. The correlation analyses of the academic grammar test scores with the TOEFL iBT® scores on one hand and with the EPT essay writing results on the other hand also demonstrated moderate degrees of correlations at a significant level. Therefore, one can safely argue that students’ productive grammatical ability in academic writing is ascribed to the theoretical definition of the construct of grammatical writing ability.

The inferences whose backing has been summarized so far concern those pertaining to the validity in test development, observation, and score interpretation. Most inferences (except the Generalization Inference, which needs further research for more solid backing) are sustained by strong backing for each assumption identified for the respective warrant. Now that the inferences on score interpretation are established, the validation effort should continue to the domains of score use for the intended purpose and its consequences. Before proceeding, however, it should be noted that the remaining inferences are made on a
hypothetical scenario where the scores of the academic grammar test would be actually taken into account in making decisions on students’ placements in—or exemptions from—ESL writing classes, in addition to the EPT essay test results. Most of the evidence for the following inferences is drawn upon the participants’ perceptions about the importance of grammatical writing ability in academic writing and/or their needs in relation to the instruction of grammatical ability in ESL writing courses.

6.1.6. Utilization Inference

Warrant 6: Equitable decisions are made with respect to examinee students’ placement in ESL writing by using the test results along with the results of the writing test.

Three assumptions underlying this warrant were identified as follows:

(1) decisions made on ESL writing placement/waiver is fair and impartial for every individual student;

(2) scores on the academic grammar test are used to make ESL placement decisions in conjunction with EPT writing results; and

(3) educational values concerning the instruction of grammar in academic writing are carefully considered in making decisions on ESL placements or exemptions.

When academic grammar test scores are used for ESL placement decisions, it is crucial that the weight of the scores should not be neglected by the decision-maker and should serve as a fair, objective criterion for ESL decisions made. An interview with the EPT coordinator as the decision-maker revealed, indeed, that he considered the grammatical writing ability as important as any other factors of academic writing. His strong belief on the value of the grammatical writing ability was grounded both empirically and heuristically. He
also expressed concerns about the students who had missed the opportunity to obtain more guided assistance to develop their grammatical writing ability, when he was presented the high numbers of students who had passed the EPT writing but performed poorly on the academic grammar test in F12 and Sp13, as discussed in 5.1.11.

In addition, the EPT coordinator wished that the academic grammar test scores could serve as tangible and undeniable evidence of the decisions made when students have a doubt on their results in the writing portion of EPT. The academic grammar test scores could also serve as an objective criterion for determining whether students need an extra help in ESL instruction, especially when their essay test results are on a borderline between two adjacent levels. Given the inferences on score interpretations justified above, the EPT essay raters and the decision-maker will be able to make more informed decisions on such cases using the scores obtained from the academic grammar test.

6.1.7. Ramification Inference

Warrant 7: Intended consequences of the use of the academic grammar test for making decisions on students’ placements in or exemptions from ESL writing courses are beneficial to groups of stakeholders.

When test scores are used for an intended purpose, it is necessary to ensure that the consequences brought about by their use should be beneficial to varying parties of stakeholders with adverse effects as minimal as possible. In this investigation, students as test takers and instructors of ESL writing courses were chosen as the stakeholders, who would be immediately affected by the use of the academic grammar test scores. Three primary investigation approaches were taken to pursue the backing for this general
assumption underlying Warrant 7. The first method employed was a crosstab analysis between the EPT essay results and possible results from the use of the academic grammar test scores with a tentative cut score of 26, the approximate mean score of the grammar test among those who passed the EPT essay test in F12 and Sp13. The second investigation method was individual interviews with student participants. The final investigation was two separate focus group interviews with ESL writing instructors at Iowa State University.

The crosstab analysis showed a striking result that, in both terms of data collection, around 40-50 percent of the test takers who passed the EPT writing did not perform better than the average score earned by the Pass students on the EPT writing. This finding also held true among the students whose Passing results were unanimously agreed by two EPT raters. If the academic grammar test had been used for the ESL decision-making along with the EPT results, these students could have received additional explicit ESL instruction to raise awareness in their own grammatical weaknesses in academic writing or even to improve their grammatical writing abilities. Of course, taking an extra ESL course might be disadvantageous to international students for various reasons, including time and tuition issues among others, as identified in 5.2.4. However, as many interview participants indicated, potential benefits from taking an extra course in ESL writing could be greater than its negative impact on the students in the long run, especially if they sought a career in academia or in an English-speaking country like the U.S. Most of the interview participants also perceived that grammatical writing ability is important in academic writing in a general sense.

ESL instructors’ perceptions about the value of the grammatical writing ability slightly varied, depending on the level of the writing class they taught and their students’
academic status (i.e., whether they were undergraduate or graduate). Among the focus group participants, the ESL instructors who mostly dealt with undergraduate students were more concerned about their students’ grammatical writing ability than the one who taught the upper-level graduate writing course. They also agreed on the importance of instruction on grammatical writing ability for the students to not be disadvantaged in their academic activities due to the limitedness in grammaticality of their writing. One of the instructors explicitly stated that he tried to offer extra help to some students who had more difficulty in grammatical writing than others, using his individual conference hours with his students. Overall, the instructors appeared to be alert about their students’ needs in grammatical writing ability and tried not to be negligent in their writing instruction.

Balancing between grammatical ability and other discourse-related aspects of writing is certainly a challenge for ESL writing instructors. The use of academic grammar test scores can assist them with this issue, especially if the test is accompanied by an instant automated scoring system that also summarizes and generates feedback on individual students’ performance on the test. The instructors’ demand for such a system was quite specific; they wanted a system that could identify students’ strengths and weaknesses both individually and as a whole class. They wished to receive feedback before the beginning of a semester. The instructors believed that such instant feedback could help them tailor course contents to their students’ needs regarding grammatical writing ability in a timely manner, from which both the instructors and their students would benefit throughout the course. Instructors would be able to design or modify the course syllabus to meet their students’ needs in grammatical writing abilities before the course begins and also prepare extra materials for individual students who needed extra guidance to develop their grammatical writing ability. As a
consequence, students could receive instruction better tailored to their needs, which would lead to effective development in the target ability.

This section reviewed the backing evidence found for the assumptions of various warrants identified to make inferences on test development, evaluation, score interpretation, and test use, in order to establish a validity argument for the development and use of the academic grammar test for an ESL placement purpose. Most of the inferences were well sustained, although the Generalization Inference requires further research for more robust backing. Table 6.1 presents a summary of the validity argument for the development, score interpretation, and use of the academic grammar test, providing backing evidence found for each warrant.

6.2. Further Validation Issues

As witnessed previously, a few results found in the pursuit of the research questions failed to provide solid backing for some assumptions underlying the warrants for the Generalization and Explanation Inferences. Evidence for validity of the use of the academic grammar test scores is also tentative, since stakeholders’ perceptions about the use of the test and its impact on them were based on a hypothetical situation of the test use. This section will discuss three important issues that arose in the evaluation of the backing evidence for the validity argument of interest.

6.2.1. Task Design (Evaluation, Generalization, and Explanation)

The test items used in the academic grammar test were pilot items, which need revision based on the results of item evaluation presented in Section 5.1.8. One issue that
### Table 6.1. Summary of Validity Argument for Academic Grammar Test

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Assumptions</th>
<th>Backing Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain Description</strong></td>
<td>W1: Test tasks contain the language of the target domain of academic English.</td>
<td>• Linguistic nature of the tasks in academic grammar test represents that for academic English.</td>
<td>• Texts used in test tasks adapted from authentic academic texts, as in biology, sociology, and communication. • Biber et al.’s (1999) <em>Longman Grammar Spoken and Written English</em> was referred to when determining final target grammatical features commonly found in genre of academic prose.</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td>W2: Observations of examinee performance are evaluated to provide observed scores informative of examinees’ task performance.</td>
<td>• Examinee performances on the test are not affected by students’ background knowledge relevant to the content of the texts of test tasks or lack thereof. • Examinee’s performances on test are properly evaluated to generate observed scores reflective of their relevant abilities to produce grammatical sentences in academic writing.</td>
<td>• Completion of test tasks did not require examinees to have content knowledge. • Think-aloud protocols revealed participants continuously relied on their knowledge and understanding relevant to the target grammatical features throughout the test taking. • Scoring rubric created to identify different degrees of language acquisition (e.g., emergence vs. mastery) of target grammatical features. • Scores of individual students on academic grammar test spread across the possible score range of the test and are normally distributed. • Gender did not affect test takers’ performance on academic grammar test as a moderating factor in relation to their proficiency level.</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>W3: Observed scores are stable estimates of expected scores in the universal domain of test tasks.</td>
<td>• The test scores have an acceptable level of reliability.</td>
<td>• Cronbach’s alpha coefficients are .784 and .723 for F12 and Sp13, respectively. • The 95% CI for the difference between the two alpha coefficients from the F12 and Sp13 datasets include 0, although its lower limit is almost on the point of 0. No significant difference in the alpha coefficient was found between gender groups in both terms. • Scores from the computer-based and paper-based academic grammar test are not comparable due to the extremely small sample size of the computer group. Further research is needed with an approximately equal number of test takers on each test delivery mode.</td>
</tr>
<tr>
<td>Inference</td>
<td>Warrant</td>
<td>Assumptions</td>
<td>Backing Evidence</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Extrapolation| W4:     | Examinees’ expected scores estimated by the observation of the academic grammar test to represent their target scores manifested in the target language use (TLU) domain. | Examinee performance at different ESL proficiency levels is significantly distinctive from each other. The 95% CI of the mean scores of the three groups did not overlap with each other.  
Examinees’ performances on the academic grammar test are in accordance with their grammatical writing ability exerted in their writing in a non-testing context, with a few exceptions. |
|              |         | • Examinees performances on test are consistent with general proficiency levels determined by another criterion.  
Examinees’ estimated expected scores are reflective of their performances in other spontaneous writing situations within the academic context. |                                                                                                                                                                                                              |
| Explanation  | W5:     | Target scores estimated by the grammar test are attributed to the theoretical construct definition underlying the construct of productive grammatical writing ability in English. | Item facility indices have general tendencies of inverse correspondence with expected developmental stages of their target grammatical features.  
Mean scores of the test items tapping into the same grammatical features have significant factor loadings on test construct of grammatical writing ability in moderate to strong degrees ($d = .31$ to $.78$).  
Examinees’ performances on the academic grammar test correlate with their performances on other criterion measures (e.g., EPT writing or TOEFL writing) at moderate degrees ($r = .491$ to $.613$ with TOEFL iBT$^{18}$ and $r_s = .438$ to $.491$ with EPT). |
|              |         | • Examinee performance on test tasks reflects cognitive difficulty levels of target grammatical features in their acquisition.  
• Observables of sub-construct target grammatical features are meaningfully associated with the construct of the productive grammatical writing ability in English.  
• The construct of productive grammatical writing ability in the academic setting is consistently exerted in other criterion measures engaging the same construct, even though learner performances may be affected by other factors in these measures. |                                                                                                                                                                                                              |
| Utilization  | W6:     | Equitable decisions are made with respect to examinee students’ placement in ESL writing by using the test results along with the results of the writing test. | Use of academic grammar test scores could provide a tangible ground for the decisions on individual students’ ESL placements/exemptions.  
The decision-maker would refer to the grammar scores to make decisions on ESL placements/exemptions to improve reliability of the test.  
Decision-maker highly appreciates value of grammatical ability in academic writing. |
|              |         | • Decision made on ESL writing placement/waiver is fair and impartial for every individual student.  
• Scores on academic grammar test are used to make ESL placement decisions in conjunction with EPT writing results.  
• Educational values concerning instruction of grammar in academic writing are carefully considered in making decisions on ESL placements or exemptions. |                                                                                                                                                                                                              |
Table 6.1. Summary of Validity Argument for Academic Grammar Test (Continued)

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant</th>
<th>Assumptions</th>
<th>Backing Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramification</td>
<td>W7: Intended consequences of the use of the academic grammar test for making decisions on students’ placements in or exemptions from ESL writing courses are beneficial to groups of stakeholders.</td>
<td>• ESL instructors benefit from the decisions made on ESL placements through the use of scores from academic grammar test.</td>
<td>• ESL writing instructors appreciate value of grammatical writing ability in academic writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Examinee students benefit from the decisions made on ESL placements by use of academic grammar test.</td>
<td>• ESL writing instructors could benefit from use of the academic grammar test, if it is accompanied by an automated scoring and feedback generation system.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• False positive decisions in exemptions from ESL writing placement are avoided or diminished.</td>
<td>• While many students appreciate benefits of instruction on grammatical writing ability, revision in pedagogical approaches to grammar instruction is inevitable to increase their satisfaction level on instruction.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Students appreciate importance of grammatical ability in academic writing.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Students believe benefit of instruction on grammatical writing ability outweighs its several potential negative impacts.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Use of test scores on academic grammar test effectively identifies the number of false positive exemptions from ESL writing placement.</td>
</tr>
</tbody>
</table>

needs particular attention in relation to this is the task design. In the academic grammar test, three different task types (i.e., two gap-filling task types and a jumbled word order task type) were employed to elicit the target grammatical features effectively. However, it appears that evaluation of the test takers’ ability to use grammatical features was sometimes confounded by the simplicity of the task design. Test takers’ perceptions on the difficulty of the three task types shed light on this issue. At the individual interview sessions with student participants, the investigator asked them a question about the easiest and the most difficult task types among the three included in the academic grammar test. Table 6.2 shows the results of their responses.
One notable finding from this table is that a majority of lower-proficiency students (i.e., ESL group participants) perceived the jumbled word order tasks the easiest among the three task types. These tasks were intended to elicit test takers’ ability to construct a sentence using the target grammatical feature. Due to their cognitive complexity, the Processability Theory usually considers the syntactic features of the English grammar more difficult than morphosyntactic features. Yet, most test takers perceived the third task type centering on syntactic features the easiest among the three task types, while perceiving the second one, targeting morphosyntactic features, the most difficult.

Table 6.2. Students’ Perceptions on the Difficulty of the Task Types

<table>
<thead>
<tr>
<th></th>
<th>Pass (N = 15)</th>
<th>ESL Group (N = 16)</th>
<th>TOTAL (N = 31)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easiest</td>
<td>Most Difficult</td>
<td>Easiest</td>
</tr>
<tr>
<td><strong>Part 1:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap filling without a key word provided (Q01-Q03)</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td><strong>Part 2:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap filling with key words provided (Q04-Q15)</td>
<td>0</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td><strong>Part 3:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jumbled word order (Q16-20)</td>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

One possible explanation for this discrepancy between the test takers’ perceptions and the expected difficulty levels of the task types may lie in the simplicity of the jumbled word order items. When developing these items, some words of the target sentence were grouped together and provided as a phrase in order to avoid deviant answers that might fail to attract
the raters’ attention for the possibility of a partial credit in scoring. One of the most confounded items is Question 20, shown in Figure 6.1.

Figure 6.1. Question 20 of the academic grammar test

The target grammatical feature of this question is a relative clause with a covert relative pronoun (Stage 3). In this item, test takers only needed to place five given units into a correct order. The expression, *part of the situation*, would have been easy for them to construct due to its extremely high frequency as can be found in the number of hits—more than billions—when the expression is searched for in a search engine. The test takers could also have had little difficulty in constructing the relative clause, *he is observing*. The Processability Theory would expect the acquisition of such a clause, which is in the canonical word order (Subject-Verb) of English with the Progressive Aspect (*is observing*), to occur at an intermediate-level developmental stage. The combination of the two parts—*part of the situation* and *he is observing*—could also have been easy, because most of the test takers’ English proficiency levels would be upper-intermediate or higher as they had already been screened through the use of TOEFL iBT® or IELTS in their admission to the university. The presentation of ‘*of the situation*’ as a single phrase would thus have served as a conspicuous conduit to the key. The item facility (IF) value of this test item also indicates that it was one of the easiest items (*IF* = 1.627 and 1.620 for F12 and Sp13, respectively).
Another possible reason for the discrepancy is that, although the acquisition of many morphological (e.g., Tense and Aspect) or morphosyntactic features (e.g., articles and Subject-Verb agreement) may occur during the early stage of interlanguage development, their mastery may be achieved at a very advanced stage of language learning, if ever. Similar tendencies were witnessed in Lardiere’s (2007) case study with a subject named Patty, who had been in the U.S. for more than two decades. In her study, Lardiere found that Patty had developed almost native-like knowledge of English in the syntactic features, whereas she still made some mistakes in morphological features like tense marking and SV agreement. While this finding was truer for her oral English ability, similar patterns were also found in her written English, although with lower error rates. This implies that both the acquisitional stage and the mastery stage of a target grammatical feature should be considered when determining its expected developmental stage.

Given the discussion so far, a careful item revision could result in the improvement in the quality of backing for the Evaluation, Generalization, and Explanation Inferences. On the other hand, improvement of backing for the Explanation Inference also requires more research in regard to the acquisitional order of advanced grammatical features of English. Without more solid empirical evidence, the predictions of performance of test items in relation to their respective target grammatical features would remain groundless or only result in weak backing of the inference.

6.2.2. Developmental Stages of Grammatical Features (Explanation)

As noted in Section 4.1.1.4, the expected developmental stages of the target grammatical features of the academic grammar test items were tentatively determined, due to
insufficient or inconclusive theoretical and empirical grounds in regard to the acquisition of
some grammatical features. The expected developmental stage of the null article (Ø) is one
element. According to Lang’s (2010) review, English learners whose first languages have
the article system acquired the English article in the order of the definite article (the), the
indefinite article (a/an), and the null article (Ø) (Gorokhova, 1990; Master, 1988). The
acquisition of English as the first language also demonstrated a similar pattern, where
children acquired the English article system in the order of the indefinite article, the definite
article, and the null article (Zehler & Brewer, 1982). On the other hand, Master (1988) found
that English language learners whose first languages lacked the article system began with the
acquisition of the null article. The lack of an article should, however, be distinguished from
the use of the null article for these language learners, because the covert article is often
claimed to have specific semantic functions, such as marking the generality or a high
specificity in reference (Celce-Murcia & Larsen-Freeman, 1999; Master, 1997). As such,
Master (1997) differentiated the null article (the covert form of the article) from the zero
article (i.e., the lack of article). Celce-Murcia and Larsen-Freeman (1999) discuss the order
of English articles acquisition as follows (p. 280):

The zero article is the most indefinite of English articles, whereas the null article is
the most definite (the following continuum is adapted from Chesterman 1991):

(least definite) zero → some → a/an → the → null (most definite)

Empirical findings on the acquisition of the English article system in the literature
thus left room for debate concerning the natural order of acquisition of the English articles.
Furthermore, Lang (2010) demonstrated that the interpretation in the order of the English
article acquisition could vary, depending on what coding scheme the investigator choose to use to determine the mastery level of the grammatical feature. For example, when Brown’s (1973) scheme of supplied in obligatory context (SOC)\(^2\) was used, Lang’s subject appeared to acquire the English articles in the order of zero article, a, the, and an. On the other hand, when Pica’s (1985) Target Language Use (TLU) scheme\(^3\) was applied, acquisition of the English articles turned out to be in the order of a, the, zero article, and finally an. The results could also differ, depending on whether a and an are treated as separate features or just allomorphs of a single indefinite article. Likewise, whether treating a covert article as the same or separating it into two types of covert articles (i.e., zero article and null article) following Chesterman (1991) would result in different implications in the acquisition of the English article system.

Another example of the difficulty in determining the acquisition order of a target feature pertains to the status of prepositions as a grammatical feature in English. In English, prepositions are important function words structurally required to indicate meanings of location, time, instrument, reason, etc. From a perspective of cognitive linguistics, the meaning of a preposition expands its semantic domains from a concrete, spatial concept to those of abstract concepts (Lakoff, 1987). Semantic aspects of the preposition are thus important in their acquisition, and their expected developmental stages may differ depending upon the target meaning of the preposition. For example, as illustrated by Celce-Murcia and Larsen-Freeman (1999), language learners will first acquire the prototypical meaning of a

\(\text{SOC} = \frac{n \text{ of the correct use of the target grammatical feature}}{n \text{ of the obligatory contexts of the target feature}} \times 100.\)

\(\text{TLU} = \frac{n \text{ of correct use of in obligatory contexts}}{(n \text{ of obligatory contexts} + n \text{ of suppliance in non-obligatory contexts})} \times 100.\)
preposition at (i.e., a point as place) and then later acquire its expanded usages to indicate meanings of time point, state/area, manner, and circumstance/cause in this sequence (p. 409).

The usage of a preposition as part of collocations (e.g., consist of, compensate for, engage in, be content with, be married to, etc.) is another issue related to the acquisition of prepositions. Such usages of English prepositions should be acquired in terms of the target collocations. These collocations are expected to be acquired at a later stage of second language (L2) development, given the facts that they are composed of two or more words and that the meaning of a collocation may differ from the literal combination of the meanings of the composing words. However, to the author’s knowledge, little research has been done conducted in regard to the natural order of vocabulary items. While many collocations are believed to be advanced vocabulary items, it is unclear when their acquisition starts to occur. Moreover, acquisition of different collocations may occur at varying stages of L2 development, if second language acquisition is affected by the frequency of target features (Ellis, 2002).

Determination of the expected developmental stages of these grammatical features should be preceded by further research in their acquisition from a second language acquisition (SLA) perspective. The results from such research could provide stronger grounds in determining expected developmental stages of grammatical features than simply relying on expert judgment. Backing for the Explanation Inference in the foregoing validity argument could then be improved as well.
6.2.3. Curriculum Revision and Pedagogical Improvement (Ramification)

In the current ESL program offered by the Department of English at Iowa State University, students placed in ENGL 101B, which focuses on the grammatical writing ability in academic English, are also required to take the upper-level ESL writing course (ENGL 101C or 101D, depending on their academic status), which centers on the improvement of discourse-related aspects in academic writing. If the academic grammar test is used as a supplement for the essay writing test of the English Placement Test (EPT) under the current ESL program, those who fail the academic grammar test (by scoring lower than 26) would have to take both ENGL 101B and 101C/D, even if they passed the essay portion of the test (e.g., P-U02 and P-G08 in Figure 5.6). The current ESL program could thus work unfavorably for such students if the academic grammar test were used in making decisions on students’ placements in (or exemptions from) ESL writing courses, because they would have to take an additional course (i.e., ENGL 101C/D) that might not be useful for them. This might then bring about a rather serious negative impact on test takers, burdening them with additional coursework and possibly making them postpone their graduation, due to the increased demand in coursework. Therefore, it is necessary to make a dramatic revision in the curriculum of ESL writing, by detaching the ENGL 101C/D requirement from placement in ENGL 101B. In other words, students who are placed in ENGL 101B should be able to get waived from ENGL 101C/D requirement if they had developed a sufficient level of discourse competence in written academic English. Such evaluation could be implemented at the end of the course by the instructor through various types of assessment such as portfolio assessment or an in-class final essay test.
Another issue relevant to the consequence of the test use lies in the pedagogical issues in the instruction of grammatical writing ability in the ESL writing classroom. As witnessed in Section 5.2.4, many of the students who had taken ENGL 101B or a similar grammar-focused writing class thought that the course had not been of much use in terms of the improvement of their grammatical writing ability. One possible reason of this impression could be the drill-type exercises students were given in class. Alternatively, the pedagogical practice in the instruction on grammatical writing ability could have remained at a rule-based level. If the pedagogy of grammatical ability were not improved in ENGL 101B, many students who passed the essay part of the EPT but failed the grammar test might think that the course is useless for them. In order to avoid this unintended negative consequence of test use, ENGL 101B instructors could adopt a different pedagogical approach to grammatical ability. For example, they could develop teaching materials, based on cognitive linguistics, to help students improve their understanding of subtle differences between similar grammatical features (e.g., gerund vs. to-infinitive, or different prepositions of similar functions like in, on, and at) from a semantic or cognitive perspective of language development and use. Focus on the function of grammatical features could also lead to effective improvements of grammatical writing ability, since students would develop a better understanding of when a particular grammatical feature is appropriate to use and when it is not. Students’ perception and appreciation about course contents and their effectiveness in the development of the target ability would therefore improve, if the contents of the ESL writing course were devised by adopting a pedagogical approach that highlights semantic, cognitive, and functional aspects of grammatical features.
6.3. Two Additional Issues on Test Implementation

In addition to the three issues addressed previously, two other conditions should be met to actually implement the academic grammar test in EPT administration. First, development of an automated scoring system for the test is inevitable, and the quality of the assessment made using the automated system should be carefully examined before its actual use. Different parties of the stakeholders of the test (i.e., the EPT coordinator as a decision maker, ESL writing instructors, and students as test takers) could benefit from use of an instant automated scoring and feedback generation system. This system would expedite the entire process of scoring and the decision-making process, and the instructors of the ESL writing courses could identify their students’ strengths and weaknesses in academic grammatical ability before the course begins, so that they could tailor course materials to their students’ needs effectively and efficiently. By taking such a class, students would also receive an opportunity to improve their grammatical writing ability in a more effective way than when trying to accomplish it on their own.

In order for an automated scoring system to be implemented, both the academic grammar test and the EPT should be administered via computer as well. Some logistical issues would arise in this regard, such as finding a proper platform for test delivery, test security, timing of test administration (e.g., when these tests should be delivered, and whether they should be used together or could be utilized separately on different schedules), and securing a server capacity to accommodate delivery of the test to hundreds of students at one time. These apparently challenging issues may be resolved by administering the academic grammar test ahead of the EPT either on campus or via the Internet. Test security
issues would also need careful consideration. A large pool of test items should also be available to prevent cheating among test takers.

6.4. Conclusion

6.4.1. Summary of the Dissertation Study

This study was initiated by a motivation to supplement the decisions made about non-native English speaking students’ placements in ESL writing courses on the basis of their performance on a single essay test item in the English Placement Test (EPT) at Iowa State University. A test of productive grammatical writing ability in academic English (i.e., the academic grammar test) was thus developed to measure test takers’ productive grammatical ability in the academic setting. The test development project began by outlining an interpretive argument for the validation of the score interpretation and use of the academic grammar test. The validation framework chosen for this project was an adapted version of Chapelle et al.’s (2008, 2010) validity framework, which was developed from Kane’s (2006) interpretive/validity argument model. The validation framework of the present study also adopted aspects of Bachman and Palmer’s (2010) Assessment Use Argument (AUA) model, with intent for the consequences of the test use to be borne in mind throughout the entire process of test development and evaluation. Seven inferences (Domain Description, Evaluation, Generalization, Extrapolation, Explanation, Utilization, and Ramification) were set for the interpretive argument, and their warrants and underlying assumptions were also identified. These were then accompanied by the types of backing evidence which would support the assumptions and, in turn, sustain the warrants so that the validity argument for the score interpretation and use of the academic grammar test could be established.
Test development began with the selection of target grammatical features. This process was grounded in theoretical discussions and empirical findings of SLA research on the natural developmental order of English as a second language as well as the findings of the language assessment research grounded in the same theoretical background. The contents of the test were also chosen from authentic introductory-level college textbooks from different academic disciplines in order for the test items to reflect the characteristics of the language used in the real-world academic setting. After administering the test to entering non-native English speaking students along with the EPT in Fall 2012 and Spring 2013, attempts were made to validate the score interpretation, and use of the test for the intended purpose using a mixed methods research design.

Evaluation of the evidence as the result of a series of investigations revealed that the validity of the score interpretation and use of the academic grammar test could be well sustained, although some evidence for the Generalization and Explanation Inferences was tentative or relatively weak, calling for further research in these areas. The evidence for the Utilization and Ramification Inferences was also tentative, because the data were collected under a hypothetical situation of the test use for the ESL placement purpose. In the second section of the Discussion chapter, three additional issues were addressed as a way to improve the quality of backing evidence for the Evaluation, Generalization, Explanation, and Ramification Inferences. These three issues concerned (1) task design, (2) additional research on the expected developmental stages of grammatical features, and (3) a revision of the current ESL program and pedagogical approaches to grammar instruction in the ESL writing classroom.
6.4.2. Limitations

This study is limited in four ways. First, the test items utilized in the academic grammar test should have been piloted and revised for better measurement of the construct. While most of the target grammatical features were adopted from earlier studies (Chapelle, Chung et al. (2010) and Chung (2012)), new test items were created and used in this study to better reflect linguistic characteristics of academic prose. The test with carefully revised items, based on item evaluation revision and students’ perceptions concerning the difficulty level of task types, might result in more robust evidence for the warrants of the Generalization and Explanation Inferences, as discussed in Section 6.2.1. Second, the sample size for the computerized academic grammar test was too small to examine the comparability of scores between the paper-based and computer-based academic grammar tests. An approximately equal number of examinees per test format should present a better picture about the score comparability across the two delivery modes. Potential effects of various test delivery platform (e.g., Blackboard Learn vs. Moodle) on examinees’ performance on the academic grammar test should also precede the comparability study to minimize the effect of construct-irrelevant factors on test performance and to prevent score interpretations from being misguided.

Another limitation in the methodology is that investigations about the validity of test use relied on a hypothetical situation in which the academic grammar test would be actually incorporated into the EPT and referred to in making decisions on whether the test takers should take one or more ESL writing courses or not. Stakeholders’ perceptions about potential effects of the test use on themselves might change if the test were actually
administered as an official part of the EPT, in part because other variables like the use of an automated scoring system could result in different consequences.

Finally, the study is limited in that field research had not been conducted prior to test design to explore the extent to which grammatical writing ability would be demanded in a variety of academic disciplines and professional fields. This study relied merely on students’ perceptions and opinions about the importance of grammatical writing ability. To investigate the multidimensional aspects of validity of the test use, triangulation of the data from different sources is necessary. Interviews or surveys with instructors and professors in various academic disciplines as well as those with professionals in many different industrial areas could allow us to understand a long-term impact of the use of the academic grammar test. Examinations of grammatical characteristics of professional writing samples from various academic disciplines and professions would also enrich the backing evidence.

6.4.3. Suggestions for Future Research

While the aforementioned limitations are all possible topics for future research, one of the immediate follow-up research studies should be on the development of an automated scoring and feedback generation system. Performance of the automated scoring/feedback system should be evaluated in comparison to human ratings. It should also be investigated whether or not the scores generated by the automated scoring system would result in score interpretations comparable to those achieved by human ratings. This study should also be accompanied by a close investigation on the effect of different task types on the examinees performance on the academic grammar test. Finally, more SLA research should be
conducted in regard to the natural order of acquisition of advanced-level grammatical features so that the backing for the *Explanation Inference* can be reinforced.
APPENDIX A: TEST SPECIFICATION

Specification of the Academic Grammar Test

1. General Description (GD)

1.1 Background

This test is designed to supplement the essay writing test of the English Placement Test at Iowa State University in making decisions on students’ placements in ESL writing (ENGL 101B/C/D) courses. Currently, judgments on students’ placements in these courses are made on the basis of students’ performances on a single-item essay test. The validity as well as the reliability of the EPT writing test is thus not acceptable from the assessment perspective. Referring to an indicator of students’ grammatical ability in spontaneous academic writing setting will help increase the validity of the score interpretation and use of the EPT writing test. The test should last no more than 15 minutes. It should be designed to elicit test takers’ spontaneous writing responses.

1.2 Target grammatical features

Grounded in the Processability Theory (Pienemman, 1999; Pienemann & Keßler (2011)), Norris (2005), and Chapelle et al. (2010) as well as findings of SLA studies (e.g., Anderson, 1978; Bardovi-Harlig, 1999; Bailey, Madden, & Krashen, 1974; Bayley, 1999; Dulay & Burt, 1973, 1974; Ladiere 2007), the items of the academic grammar test will target morphosyntactic, lexico-syntactic, syntactic, and functional features of the grammar of the American English language. The grammatical features of interest are as follows:

- Morphosyntactic features
  - Articles (indefinite, definite, null)
  - Tense (past, present)
  - Aspect (simple, progressive, and perfect)
  - Passive
  - Conditionals & Subjunctives (True & untrue conditionals)

- Lexico-syntactic features
  - Preposition
    - Preposition of a collocation
    - Preposition required in passive
    - Other semantically related uses of prepositions (in relation to the argument structure of the associated verb or noun)
  - Sentential adverbials

- Syntactic features
  - Cancellation of SV inversion in embedded question (nominal clause)
  - SV inversion with sentence-initial negation
  - Relative clauses
  - Present/past participle phrases (adjectival phrases as a short form of relative clauses)
Subjacency or multiple wh-questions in an embedded clause

- Functional features
  - Past tense in relation to the intrinsic semantic attributes of English verbs (i.e., +/- Punctual, +/- Telic, +/- Activity)
  - Aspect
  - Modal + verb / Modal + present perfect

1.3 Considerations of characteristics of academic prose

Selection of target grammatical features should also take into account the candidate features’ actual distributions and uses in academic prose whose findings are reported by corpus-based grammatical analysis (Biber, Johansson, Leech, Conrad, & Finegan, 1999). For example, a sequence of modal + have p.p. is more commonly used in conversations than in academic prose. This grammatical feature will thus not be appropriate for candidate target features of the academic grammar test, since it does not properly reflect the characteristics of academic prose. Target constructions or sentences used in the test should be excerpted or adapted from college textbooks at introductory level. Task performance should not require test takers to have background knowledge in particular subject areas.

2. Prompt Attribute (PA)

2.1 General descriptions: Prompts and general instructions

Each test prompt generally consists of (a) a target sentence, (b) blank, and (c) a list of key words which test takers are supposed to use to complete the given sentence. The test prompt will ask test takers to fill in the blank of the respective sentence using the key words given in the list. The prompt will also ask them to add one or two words and/or change the forms of given words if necessary to make the sentence complete. Test items of the same task type will be grouped together under a general instruction. A few key words may be underlined in the instructions in order to help test takers’ clear understanding of the directions. In the instructions, a few important aspects may be underlined as shown below:

Complete the sentences below using all the given words or phrases. You may change word forms if necessary. Also, add minimally required words if necessary.

Reorder the given words or phrases to make complete sentences. Use all words given in the list. Do NOT add more words or change word forms.

2.2 Task types

Four constructed-response task types will be employed in this test, which are:

(a) To fill in the blank by adding a word if necessary;
(b) To fill in the blank using all the given words;
   a. Change word forms if necessary
   b. Add minimally required words
c. Rearrange the given words or phrases into a right order.

Task types of (a) and (b-a) will be exploited to target morphosyntactic and lexico-syntactic (especially prepositions) features. Task types of (b-b) and (b-c) are useful for targeting lexico-syntactic and syntactic features of the grammar. Tasks of (b-a) and (b-b) may be combined in the test instructions. In tasks of (b-a) and (b-b), words will be provided in a random order such that test takers need to use those words to construct the target structure of the question.

3. Response Attribute (RA)

Test takers will read the general directions of each section of the grammar test and read the target sentence of each question. They will also read expressions provided in the list and cognitively work to construct the target structure using those expressions. They will put the given key expressions into a correct order first and then decide whether to add a new word or make word forms unless they are asked not to (for jumbled word order tasks).

4. Sample Items

Fill in each of the blanks with a necessary word, if any, to complete the sentences below. You may leave them blank if no additional word is required.

In 2007, a new mathematical model was developed for evaluation of alternative natural gas policies.

Task type: Add a new word (Task (a))

Target grammatical feature: Article (Morphosyntactic)

Complete the sentences below using all the given words or phrases. You may change word forms if necessary. Also, add minimally required words if necessary.

The greenhouse effect has been known since 1824. [know]

Task type: Use the given word(s), add new words, change word forms (Tasks (b-a) & (b-b))

Target grammatical features: Passive (primary) & present Perfect (secondary) (Morphosyntactic)
Reorder the given words or phrases to make complete sentences. Use all the expressions given in the list.

[ proposed, which, who, theory ]

The author put together a brief history of learning theories and summarized who proposed which theory at the end of the chapter.

**Task type:** Rearrange given expressions (Task (b-c))

**Target grammatical feature:** Subjacency (multiple wh-words within an embedded clause)

### References


APPENDIX B: ACADEMIC GRAMMAR TEST

Read and follow the directions for each section carefully. Add minimally required words or change word forms only when the directions say so.

I. Fill in each of the blanks with a necessary word, if any, to complete the sentences below. You may leave them empty if no additional word is necessary. (1-3)

1. Society is defined as _______ network of social relations.

2. _______ Human behavior is always according to some norms.

3. Plants release waste gases _______ the air through their stomata.

II. Complete the sentences below using all the given words or phrases. You may change word forms if necessary. Also, add minimally required words if necessary. (4-11)

4. The hypothesis ___________________________ the help of facts. [ is, verify ]

5. Plants make their own food through a chemical process _______________________ photo-synthesis. [ know ]

6. We are actively engaged ______________________ _ sense of TV programs. [ make ]

7. Energy costs ____________________________ in real money terms since the early 1970s. [ rise ]

8. Linguists ________________________________ that spoken languages say a lot about peoples’ perspectives of the world. [ long, argue ]

9. Not only ________________________________ by 37%, many were held worldwide
   ________________________________ India, China, Pakistan, Taiwan, Israel, France, and Germany.
   [ lectures, increase ]
   [ include ]

10. _____________________________ more slowly, all geologic activity would have proceeded at a
    slower pace. [ it, be, run ]

11. Without her support and input, this book simply _____________________________ possible. [ will, not, be ]
III. Reorder the given words or phrases to make complete sentences. Use all expressions given in the list. Do NOT add more words nor change word forms. (12-16)

12. various stages with beginning observations

The social research method has ____________________________.

13. why tell happening us is the communication

Talking about communication does not ____________________________

__________________________________________________________________________.

14. what how they they have communicate common in and

We are going to discuss the examples of mass communication in Chapter 5 and talk about

__________________________________________________________________________ with us.

15. control or govern relationships that the norms among

In every society there are social ____________________________

__________________________________________________________________________ the members.

16. of the situation part he observing is

When an observer is a ____________________________

it is called participation observation.
## APPENDIX C: SCORING RUBRIC

<table>
<thead>
<tr>
<th>Surface Q No</th>
<th>Question No</th>
<th>Target Grammatical Features</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Indefinite article</td>
<td>Article</td>
<td>is, that, this, some, the, each</td>
<td>every, most, such (modern, studying, it's said, the facts said)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Null article</td>
<td>Generality</td>
<td>before, after, as</td>
<td>in; other prepositions (that do not have a function of subordinate conjunctions)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Preposition</td>
<td></td>
<td>is verification; verify; is verify; had verified; are verifying; was verifying, were verifying, will verify</td>
<td>is to verify; is verified; were verified; are verified; is used to verify; was used to verify; is verifying; is a way to verify; verified; is verification; has been verified; is used to verify a theory</td>
</tr>
<tr>
<td>4.1</td>
<td>4</td>
<td>Passive</td>
<td>(Structural grammaticality)</td>
<td>is to verify; is verified; were verified; are verified; is used to verify; was used to verify; is verifying; is a way to verify; verified; is verification; has been verified; is used to verify a theory</td>
<td>is verified; was verified; could not be verified; can be verified; has been verified; should be verified; is being verified; is truly verified; is to be verified</td>
</tr>
<tr>
<td>4.2</td>
<td>5</td>
<td>Preposition (associated with Passive)</td>
<td>(Concept and usage of prepositions)</td>
<td></td>
<td>other prepositions; due to</td>
</tr>
<tr>
<td>5.1</td>
<td>6</td>
<td>Past participle (ed)</td>
<td>Passive</td>
<td>known; is known</td>
<td>as known; known; is known; called; which/that known; which was known; which are known; is known; has been known; was known; are known; which known (the), were known</td>
</tr>
<tr>
<td>5.2</td>
<td>7</td>
<td>Preposition (associated with Passive)</td>
<td>(Semantic differences of known as vs. known for vs. known by)</td>
<td></td>
<td>by; for, (other prepositions)</td>
</tr>
<tr>
<td>6.1</td>
<td>8</td>
<td>Collocation (involved in -ing/ing)</td>
<td>Preposition</td>
<td>to (infinitive)</td>
<td>other prepositions [e.g., with, of, by, into (when followed by 'making')];</td>
</tr>
<tr>
<td>6.2</td>
<td>9</td>
<td>Gerund</td>
<td>Collocation ('making sense')</td>
<td>make</td>
<td>making the a</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>Present perfect (Tense &amp; Aspect)</td>
<td>Perfect</td>
<td>rose; raised; risen; have risen; have been risen; has risen; have been risen; had risen; had been risen; have been rising; had been rising; had risen had been raised; have been raised; has been raised; has been raised; had been raised; had been rising; had been rising; had risen; has risen; have been rising; have risen; have been rising</td>
<td>have risen; have been rising</td>
</tr>
<tr>
<td>Surface_Q_No</td>
<td>Question No</td>
<td>Target Grammatical Features</td>
<td>Primary</td>
<td>Secondary [+ grammatical]</td>
<td>0</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>----------------------------</td>
<td>---------</td>
<td>--------------------------</td>
<td>---</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>Present perfect</td>
<td>Place of an adverb 'long' (in relation to have+p.p.)</td>
<td>have long argument; long argued about; argue longly; argued a long; long argues; long argue,</td>
<td>have argued for long time; have argued for a long; long argued; argued long; has long argued; argued for long; have argued longly; argued in length; has been long argued; have been arguing since long; have a long argument; are longing to argue; professors have long argued; have longly argued; longly argued; have taken the long argument; take a long argument; had argued for a long time; had long argued; have long been argued; has long been argued; has been argued long; had a long argument; has been arguing as long as; have been arguing long; made a long argument; make a long argument; argued for a long time; long have argued; who have argued longly; have been long argued; have such a long argument; has argued long; had argued long; argued for a long time; spent a long time arguing; for a long time argued; have a long argued; no longer argued; spend a long time arguing; argued long time; argued at length; argue for a long time; argue for a long time;</td>
</tr>
<tr>
<td>9.1</td>
<td>12</td>
<td>NEG + SV inversion</td>
<td>Past tense</td>
<td>that lectures (have) increased; lectures increased; the number of lectures increased; were lectures increase;</td>
<td>have lectures increased; has lectures increased; do lectures increase; did (the) lectures increased; were/are/was (the) lectures increased; had lectures increased; did lectures be increased; had the lectures been increased; do lectures increased; does (the) lectures increase</td>
</tr>
<tr>
<td>9.2</td>
<td>13</td>
<td>Present participle (adjectival)</td>
<td>Small relative clause</td>
<td>which including; included; include</td>
<td>including in; which/that includes; which/that included; but also including; also including; including from;</td>
</tr>
<tr>
<td>Surface_Q No</td>
<td>Question No</td>
<td>Target Grammatical Features</td>
<td>Primary</td>
<td>Secondary (+ grammatical)</td>
<td>0</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>--------------------------</td>
<td>---</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Subjunctive (Entire conditional) - Allow both Past &amp; Past perfect</td>
<td>Understanding the intransitivity of the verb 'run'; also the correct form of its past participle (run) // Past tense or Past Perfect in the if-clause</td>
<td>if it has been running; was it run; was it running; was it ran;</td>
<td>had it been run; if it had been run; if it had been to run; if it was/were run; had it ran; if it were/was to run; if it were run; if it had run; had it ran; if it ran; if it was ran; if it had been ran; if it had ran; had it be run; were it to ran; if it would be ran; would it be running; should it be ran; should it be run; should it be running; if it could be run; if it would run; if it would be ran; if it would be running; if it would run;</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>Subjunctive (Entire conditional) in main clause</td>
<td>Tense &amp; Aspect (past perfect)</td>
<td>would not be; will not be</td>
<td>wouldn't be (made); wouldn't be (made); will not have been; won't have been; would haven't been; wouldn't have been;</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>NP + present participle phrase (a short form of relative clause): 'beginning with'</td>
<td>NP + adjectival/prepositional phrase; Semantic meaning</td>
<td>beginning, observations with various stages; observations with beginning various stages</td>
<td>various stages with beginning observations; observations beginning with various stages; observations beginning with various observations; been beginning with various stages of observations;</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>Cancellation of SV inversion in embedded clause</td>
<td>Indirect question (embedded clause)</td>
<td>tell us why is the communication happening; tell why the communication is happening us; us tell why the communication is happening</td>
<td>tell us why is the communication happening; tell us why the communication happening; tell why the communication is happening to us; tell us, why the communication is happening; tell us why the communication is happening</td>
</tr>
<tr>
<td>Question No</td>
<td>Target Grammatical Features</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Multiple Wh-clauses</td>
<td>how they communicate and what they have in common; how they have communicate and what they have in common; how they communicate in and what they have common; how they communicate and in what they have common; what they have common and how they communicate in;</td>
<td>what they have common in and how they communicate; how they communicate and what they have in common, what they have common and how they communicate, how they communicate and what they have common; how they communicate and what they have common in; what they have and how they communicate in common; how they communicate in common and what they have; how they communicate and what in common they have; what they have in common and why they communicate; what they have in common and what they communicate; what they have and how they communicate in common</td>
<td>what they have in common and how they communicate</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Relative clause</td>
<td>relationships that control or govern the norms among; relationships among the norms that control or govern; norms among the relationships that control or govern; norms that control or govern relationships among; relationships that the norms control or govern among; norms that the relationships control or govern among; norms that control and govern the relationships among; relationships among the norms that control and govern</td>
<td>norms that control or govern the relationships among</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Relative clause</td>
<td>part of the situation, he is; observing part of the situation; he is; observing he is part of the situation; part of the situation he is</td>
<td>part of the situation, he is observing / part of the situation; he is observing / part of the situation he is observing / part of the situation in which he is observing</td>
<td>part of the situation he is observing</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D: SAMPLE INTERVIEW QUESTIONS FOR STUDENTS

1. You took a grammar test in the EPT. What do you think about the test? Was it easy or difficult?

2. What do you think were the easiest and most challenging questions (or question types)?

3. Do you think your grammatical ability was satisfactorily represented through the grammar test score?

4. You got a score of (XX) on the grammar test. If this score had been taken into account in the decision of your ESL writing placement, you may have been placed [101B/C/D or Pass], which is [the same as/different from] the current placement result. What do you think about that?

5. How much do you think the use of the grammar test score in your placement decision could have been beneficial for you?

6. Also, what adverse impact would it have had on your study at Iowa State? How severely do you think it would have affected your college/graduate curriculum?

7. Have you ever felt or experienced that your school work in the U.S. has been seriously affected by your grammatical ability? If so, please describe the situation(s) in detail.

8. What do you think about grammatical ability in academic writing? How much do you think it is important in your writing required in the U.S. college education (for example, homework assignments, lab reports, exams, and thesis)?

9. Would you be willing to take an additional course to improve your grammatical ability? Why or why not?

10. What (else) would you do to improve your grammatical writing ability? To what extent do you think it will be effective?

11. How much do you think grammatical ability in writing will impact your career if you plan to get a job in the U.S.?
APPENDIX E: ESSAY WRITING PROMPTS

For Undergraduate Students

Read the prompt carefully, plan your response, and write your draft on this sheet. Be sure to leave time to review and edit your draft at the end. You have 30 minutes to complete this task.

Peer Pressure

Peer pressure – the influence that friends and classmates have on one another – is almost always described as a negative force that leads to undesirable behavior, but it can also encourage positive behavior. Under what circumstances can peer pressure have positive effects? Write an essay explaining the potentially positive effects of peer pressure. Support your position with reasons and examples from your own experience, observations or reading.

Write your answer here. Include any pre-writing planning or outlining.

For Graduate students

Read the prompt carefully, plan your response, and write your draft on this sheet. Be sure to leave time to review and edit your draft at the end. You have 30 minutes to complete this task.

Time Management

In 2009, new first-year students at Iowa State University were asked to indicate the amount of time spent on various activities during their final year of high school (secondary school). In the chart below, you can find the percentage of students who said they spent more than six hours per week on these activities:

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet or Television</td>
<td>79%</td>
</tr>
<tr>
<td>Working for pay</td>
<td>72%</td>
</tr>
<tr>
<td>Socializing with friends</td>
<td>52%</td>
</tr>
<tr>
<td>Studying or doing homework</td>
<td>31%</td>
</tr>
<tr>
<td>Partying</td>
<td>28%</td>
</tr>
<tr>
<td>Exercising or sports</td>
<td>23%</td>
</tr>
<tr>
<td>Participating in clubs and organizations</td>
<td>12%</td>
</tr>
</tbody>
</table>

Briefly summarize the results of this survey. Then explain how these percentages compare to the amount of time you spent in various activities when you were in your last year of high school. You do not need to comment on all the activities and you can add some different categories if you wish. Can you make any generalizations about the differences or similarities you find between your experience and the experiences of the average ISU student? Do you think these differences are a result of personal differences, cultural expectations, educational expectations or some other reason?

Write your answer here. Include any pre-writing planning or outlining.
## APPENDIX F: EXAMPLES OF GRAMMATICAL ERRORS

<table>
<thead>
<tr>
<th>Code</th>
<th>Example</th>
<th>Coding Example</th>
</tr>
</thead>
</table>
| [V(0)] | I play(V(0)) soccer with my friends when I was in high school. | He play(V(a)) video games everyday. It believed(V(p)) that the Earth was flat. You believed(V(p)) that the Earth was flat. I believed(V(p)) that the Earth was flat. John took(K(p)) a shower when the doorbell rang. I am(V(a)) watching TV. I am(V(a)) watching TV. I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) watching TV. Can you(V(a)) watch TV? I am(V(a)) w
<table>
<thead>
<tr>
<th>Categories</th>
<th>Subcategories</th>
<th>Code</th>
<th>Example</th>
<th>Coding Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td></td>
<td>[MdM]</td>
<td>Jane came in the second place. She ___ be the first one if she ran a bit faster.</td>
<td>Jane came in the second place. She [MdM] be the first one if she ran[V(c)] a bit faster.</td>
</tr>
<tr>
<td>Superfluous</td>
<td></td>
<td>[MdS]</td>
<td>He may can send some money to his family.</td>
<td>He may can[MdS] send some money to his family.</td>
</tr>
<tr>
<td>Noun</td>
<td>Superfluous</td>
<td>[NS]</td>
<td>… about the reason or cause why we lose.</td>
<td>about the reason or cause[NS] why we lose</td>
</tr>
</tbody>
</table>
| Number           | Inappropriate (N/Det) | [NN]  | • I bought two apple yesterday.  
• … much reasons …                                                                                                                                                  | • I bought two apple[NN] yesterday.  
• … much[NN] reasons …                                                                                 |
| Article/Determiner | Inappropriate | [Art] | There is the pond in the park.                                                                                                                                                                         | There is the[Art] pond in the park.                                                            |
|                  | Missing       | [ArtM] | There is ___ pond in the park.                                                                                                                                                                         | There is [ArtM] pond in the park.                                                              |
|                  | Superfluous   | [ArtS] | Sandy lent the her book to me.                                                                                                                                                                         | Sandy lent the her[ArtS] book to me.                                                           |
|                  | Missing       | [PrepM] | The office notified me ___ the result of the contest.                                                                                                                                                    | The office noticed me [PrepM] the result of the content.                                        |
| Pronoun          | Gender        | [Pro(g)] | I have a sister. He is 8 years old.                                                                                                                                                                   | I have a sister. He[Pro(g)] is 8 years old.                                                    |
|                  | Case          | [Pro(c)] | • Me is a student.  
• Most people love them families.                                                                                                                                                   | • Me[Pro(c)] is[Aux(a)] a student.  
• Most people love them[Pro(c)] families.                                                      |
<p>|                  | Number        | [Pro(n)] | I love puppies. It is really cute.                                                                                                                                                                     | I love puppies. It[Pro(n)] is really cute.                                                      |
|                  | Person        | [Pro(p)] | She said to me nobody will laugh at me                                                                                                                                                               | She said to me nobody will laugh at me[Pro(p)]                                                |
|                  | Superfluous   | [ProS]  | Children they should go to bed before 9pm.                                                                                                                                                             | Children they[ProS] should go to bed before 9pm.                                                |
| Relative Pronoun | Inappropriate | [Rel]  | I know a lady what has a beautiful mansion.                                                                                                                                                           | I know a lady what[Rel] has a beautiful mansion.                                               |
|                  | Missing       | [RelM]  | I know a lady ___ has a beautiful mansion.                                                                                                                                                            | I know a lady [RelM] has a beautiful mansion.                                                  |</p>
<table>
<thead>
<tr>
<th>Categories</th>
<th>Subcategories</th>
<th>Code</th>
<th>Example</th>
<th>Coding Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjunction</td>
<td>Inappropriate</td>
<td>[C]</td>
<td>• I love him despite he does not love me.</td>
<td>• I love him despite[C] he does not love me.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• I didn’t study although I passed the exam.</td>
<td>• I didn’t study although[C], I passed the exam.</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>[CM]</td>
<td>Sandy kicked a ball, she left the place.</td>
<td>Sandy kicked a ball, [CM] she left the place.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(or [CM] Sandy kicked a ball, she left the place.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Displaced</td>
<td>[CP]</td>
<td>Mary loves me, so.</td>
<td>Mary loves me, so[CP].</td>
</tr>
<tr>
<td></td>
<td>Superfluous</td>
<td>[CS]</td>
<td>Even though we didn’t get a good ranking compare to the others, but I really enjoyed this experience.</td>
<td>Even though we didn’t get a good ranking compare to the others, but[CS] I really enjoyed this experience.</td>
</tr>
<tr>
<td>Parts-of-Speech</td>
<td>Noun-Verb</td>
<td>[N-V]</td>
<td>Dave believes that his little sister came from the heaven.</td>
<td>Dave believes[N-V] that his little sister came from the heaven.</td>
</tr>
<tr>
<td></td>
<td>Noun-Adjective</td>
<td>[N-A]</td>
<td>I love pretty.</td>
<td>I love pretty[N-A].</td>
</tr>
<tr>
<td></td>
<td>Run-on Sentence</td>
<td>[RO]</td>
<td>She said to me nobody will laugh at me, at the same time, it’s a good way to performance myself to speak clearly, loudly in front the people.</td>
<td>She said to me nobody will laugh at me, at the same time, it’s a good way to performance myself to speak clearly, loudly in front the people[RO].</td>
</tr>
<tr>
<td></td>
<td>Fragmented</td>
<td>[F]</td>
<td>Because Terry hit me on my face.</td>
<td>Because Terry hit me on my face[F].</td>
</tr>
<tr>
<td></td>
<td>Missing Subject</td>
<td>[MSb]</td>
<td>____ Runs really fast.</td>
<td>[MSb] Runs really fast.</td>
</tr>
<tr>
<td></td>
<td>Missing Object</td>
<td>[MOb]</td>
<td>• The child told me ____.</td>
<td>• The child told me [MOb].</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Sarah thinks ___ fair that the rich pay more taxes than the poor.</td>
<td>• Sarah thinks [MOb] fair that the rich pay more taxes than the poor.</td>
</tr>
<tr>
<td></td>
<td>Missing Subject Complement</td>
<td>[MSC]</td>
<td>Terry seems ____.</td>
<td>Terry seems [MSC].</td>
</tr>
<tr>
<td>Categories</td>
<td>Subcategories</td>
<td>Code</td>
<td>Example</td>
<td>Coding Example</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------</td>
<td>-------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Missing Object Complement</td>
<td></td>
<td>[MOC]</td>
<td>They named their son ____</td>
<td>They named their son [MOC].</td>
</tr>
<tr>
<td>Semantic/Lexical</td>
<td>Word Choice (content word or collocations)</td>
<td>[WC]</td>
<td>• Joyce is qualified into three identity groups.</td>
<td>• Joyce is qualified[WC] into three identity groups.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• The landscape was made advantage of by the soldiers.</td>
<td>• The landscape was made[WC] advantage of by the soldiers.</td>
</tr>
<tr>
<td></td>
<td>Unclear meaning at phrase level</td>
<td>[UP]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unclear meaning at clause level</td>
<td>[UC]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unclear meaning at sentence level</td>
<td>[US]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td>Punctuation</td>
<td>[Punc]</td>
<td>Since China has a great number of population, and the number of students who can enter college is limited, we have to stand out from cruel competition if we want to become a college students, this is why we have to spend most time on study, because it's the best chance to chase our dreams.</td>
<td>Since China has a great number of population,[Punc] and the number of students who can enter college is limited, we have to stand out from cruel competition if we want to become a college students,[Punc] this is why we have to spend most time on study, because it's the best chance to chase our dreams.</td>
</tr>
<tr>
<td></td>
<td>Spelling</td>
<td>[S]</td>
<td>licence</td>
<td>licence[S]</td>
</tr>
<tr>
<td></td>
<td>No space between words</td>
<td>[NoS]</td>
<td>none</td>
<td>none[NoS]</td>
</tr>
</tbody>
</table>
APPENDIX G: INTERVIEW QUESTIONS FOR EPT COORDINATOR

1. What are the current decision making procedures of the English Placement Test?

2. If automated scoring for the academic grammar test were available, would you like to use the grammar test scores in making decisions on ESL writing placements/exemptions?

3. If so, how would you like to use the grammar test scores?

4. If you used both the grammar test scores and the EPT essay results together to make ESL placement decisions, what would be appropriate proportions of the two results contributing to the final decisions?

5. What would be advantages and disadvantages of the use of the academic grammar test in ESL decision making?

6. What do you think about the value of grammatical knowledge in academic writing?

7. To what extent do you believe it is important in comparison to other writing constructs like coherence, development of ideas, mechanics and spelling, functional language use, etc.?

8. Do you refer to examinees’ demographic or educational background in making decisions? Why or why not?
APPENDIX H: INTERVIEW QUESTIONS FOR ESL INSTRUCTORS

1. What aspects of the writing construct do you consider in grading students’ essays?

2. How much do you weigh each of those aspects in grading? Among those aspects, what do you consider most and least when grading essay assignments?

3. How much do you take into account grammatical accuracy (or lack thereof) in grading essay assignments?

4. What do you think about the importance of instruction on grammatical knowledge in ESL writing class?

5. How much do you focus on grammatical accuracy (and/or complexity) in teaching ESL writing?

6. What types of feedback do you give on students’ writing assignments?

7. What strategies do you use to draw students’ attentions to common grammatical errors that they make?

8. To what extent do you think students will benefit from the instruction on grammatical form and meaning in academic writing classrooms?

9. If an automated scoring were implemented on the academic grammar test and thus it was officially included in the EPT, you might be able to get students’ scores on the grammar test at the beginning of the semester (or before the class begins upon your request). How informative would this be to have information on students’ grammatical ability?

10. What kind of feedback would you like on students’ grammatical ability if an automated scoring were available?

11. If so, to what extent would you benefit from it in terms of teaching ESL writing? In what ways, would you be able to utilize such feedback in designing your lesson plans or in instructing students?
REFERENCES


