Examining Automated Corrective Feedback in EFL Writing Classrooms:
A Case Study of Criterion

Thi Linh Giang Hoang
ORCID identifier of the author: 0000-0001-6890-7950

Name of degree: Doctor of Philosophy
December 2019
Faculty of Arts
School of Languages and Linguistics

Submitted in partial fulfilment for the degree of
Doctor of Philosophy
University of Melbourne
ABSTRACT

Automated writing evaluation (AWE) systems are increasingly used in classroom settings to provide formative feedback to learners. Yet, there is a scarcity of research evidence about the impact of automated feedback on accuracy development or writing/revision practices and a lack of longitudinal studies into learners’ engagement with automated feedback. This research examines the value of the automated corrective feedback (ACF) generated by ETS Criterion as a learning and assessment tool in the EFL writing classroom. Specifically, it seeks to answer the questions about (1) the nature and accuracy of Criterion ACF; (2) students’ engagement and perceptions; and (3) their changed accuracy following the use of such feedback. The interaction between individual learner factors and their response to Criterion ACF is also investigated in search of explanatory factors for selected cases’ engagement and observed accuracy development over a five-month period.

This study adopted a pre-post quasi-experimental design on a sample of 104 English majors divided into two groups: experimental and comparison. During three practice sessions, the comparison group wrote their essays on paper and submitted them to the instructor for feedback. The experimental group, however, composed their writing on Criterion and revised their drafts in response to its feedback before submitting revised drafts to the teacher for feedback. Besides the test essays, data included first and revised drafts from Criterion practice sessions, recorded think-aloud protocols conducted with 14 students as they revised essays using Criterion corrective feedback, stimulated recall interviews and end-of-term focus group interviews. Students’ changes in writing accuracy were calculated using error analysis of the test scripts of the two groups. These were triangulated with the qualitative data of how the students engaged with the feedback from Criterion and their revision practices. Further triangulation came from students’ perceptions of automated feedback in the stimulated recall and focus group interviews.

The validation of Criterion ACF as a learning and assessment tool in the EFL writing classroom reveals a mixture of support and rebuttal evidence. Criterion was able to address EFL learners’ needs for surface-level errors, but it still lacked coverage of some major issues in the students’ L2 writing. It can be praised for facilitating revising as well as self-regulatory writing strategies and triggering noticing among the students, but Criterion’s approach to feedback generation was not pedagogically based, resulting
in a lack of meaningful engagement with the feedback. Overall, despite students’ positive feelings about *Criterion* ACF, reservations about its value remain due to learners’ middling revision success rates and the absence of significant intervention or retention effects of the use of *Criterion* ACF on their accuracy gains over the studied period.

The findings extend our understanding about students’ engagement and use of the automated corrective feedback. The study’s main implications relate to formative feedback practices in the classroom, including the need to supplement *Criterion* automated feedback with teacher feedback to support L2 writing instruction and classroom-based assessment. Also, *Criterion* corrective feedback should be designed to be more adaptable to focus learners’ attention on the relevant issues for their developmental stage.
DECLARATION

This is to certify that

i. the thesis comprises only my original work toward the Degree of Doctor of Philosophy;

ii. due acknowledgement has been made in the text to all other material used;

iii. the thesis is fewer than the maximum word limit in length, exclusive of tables, bibliographies and appendices.

.................................

Thi Linh Giang Hoang

2 December 2019
PREFACE

I gratefully acknowledge The School of Languages and Linguistics and The Faculty of Arts, University of Melbourne, for generously supporting this project with the Melbourne International Research Scholarship and Melbourne International Fee Remission Scholarship throughout the years of my PhD.

This research has been made possible with Educational Testing Service’s generous permission for me to access Criterion for research purposes. I would like to thank ETS for providing free access to Criterion for more than one year from September 2016 to December 2017 which spanned the duration of the pilot study and then the main study. In addition, this project has also received ETS Small Grant for Doctoral Research in Second or Foreign Language Assessment, which provided me with a valuable extra source of fund for data analyses.

This research project has also been partially funded by The International Research Foundation through their granting me the Doctoral Dissertation Grant Award 2017, which financially supported data collection and analyses.
ACKNOWLEDGEMENTS

I am profoundly indebted to Associate Professor Neomy Storch, my primary supervisor. Without her unwavering support and expert guidance my thesis would have never come to be. Her great expertise in the field of second language writing and SLA provides me with invaluable resources all throughout the process of my PhD study. I have been particularly motivated by her great dedication as a supervisor, from whom I can always rely on for timely and insightful feedback and guidance.

I am also deeply grateful to Associate Professor Ute Knoch, my second supervisor, for her solid support and specialist advice. Ute's vast and profound knowledge about assessment and quantitative research methods has helped me resolve several challenges in the research process. She has also inspired me with sensible advice on practical matters throughout the years of my PhD.

To Associate Professor Carsten Roever, who was initially my second supervisor but handed the role over to Ute as my research underwent a change of direction, I owe numerous thanks for his support in accessing Criterion for my research, and for his advice during the first year of my PhD as well as later consultation on quantitative matters.

I am thankful to Dr. Janne Mortan for being a nice and supportive committee chair during my candidature. To Nanang Zubaidi and Apichat Khamboonruang, my PhD fellows, I am grateful to them for having enthusiastically supported me in coding the data of the study.

I also want to express my gratitude to Dr. Sandy Clarke-Errey and Associate Professor Graham Hepworth whose guidance on statistical matters was valuable in my data analyses. My special thanks also go to Ms. Diem Thai and Dr. Huy Nguyen who were willing to let me in their writing classes for the whole semester to collect data for this research. They were very supportive during the six months of my data collection.

I also want to thank all the second-year English majors in the three writing classes at Hue University, College of Foreign Languages who willingly participated in this research. They were hard-working and open-minded about a departure from the traditional writing classroom to welcome the incorporation of Criterion in their writing course, which I highly appreciated. Y Nhi, Tan Quang, Thao Nhi, and Van Anh are the
names I feel indebted to for their great contribution during the data collection process as well as their help with data transcription in the later stages of this research.

And as always, my deep gratitude goes to my beloved parents and family, who were thousands of miles away as I carried out this research study but without their unfailing support, my PhD would not have progressed to the point I am now. They have given me unconditional love and support to make sure I can pursue my research. Finally, I cannot say enough about the great comfort and reassurance brought to me by my husband and our two little boys who have stood by me through it all to support me in every possible way in my personal pursuits. Thanks for that, I solidly work my way up to my degree completion.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Declaration</td>
<td>iv</td>
</tr>
<tr>
<td>Preface</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vi</td>
</tr>
<tr>
<td>Table of contents</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xvii</td>
</tr>
<tr>
<td>List of Appendices</td>
<td>xviii</td>
</tr>
<tr>
<td>CHAPTER 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Significance</td>
<td>2</td>
</tr>
<tr>
<td>Briefly about <em>Criterion</em></td>
<td>3</td>
</tr>
<tr>
<td>Thesis Structure</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2: Literature Review</td>
<td>7</td>
</tr>
<tr>
<td>Theoretical Constructs</td>
<td>7</td>
</tr>
<tr>
<td>Engagement</td>
<td>7</td>
</tr>
<tr>
<td>Noticing</td>
<td>8</td>
</tr>
<tr>
<td>Written Corrective Feedback</td>
<td>11</td>
</tr>
<tr>
<td>Written corrective feedback and theoretical foundations in SLA</td>
<td>11</td>
</tr>
<tr>
<td>Empirical studies on written corrective feedback</td>
<td>12</td>
</tr>
<tr>
<td>Student Response to WCF</td>
<td>16</td>
</tr>
<tr>
<td>Engagement</td>
<td>16</td>
</tr>
<tr>
<td>Uptake</td>
<td>17</td>
</tr>
<tr>
<td>Retention</td>
<td>18</td>
</tr>
<tr>
<td>Automated Writing Evaluation</td>
<td>19</td>
</tr>
<tr>
<td>Research into AWE Systems</td>
<td>21</td>
</tr>
<tr>
<td>System-centric results</td>
<td>21</td>
</tr>
<tr>
<td><em>Research into automated essay scoring</em></td>
<td>21</td>
</tr>
<tr>
<td><em>Research into automated feedback</em></td>
<td>30</td>
</tr>
<tr>
<td><em>Automated corrective feedback</em></td>
<td>30</td>
</tr>
<tr>
<td>User-centric evaluation</td>
<td>36</td>
</tr>
</tbody>
</table>
CHAPTER 3: Methodology ................................................................. 48

Brief Overview of the Study .......................................................... 48
Context of the Study ...................................................................... 50
Criterion Feedback ....................................................................... 50
Structure of the Study .................................................................... 55
Instruments ...................................................................................... 56

Pre-post tests .................................................................................. 56
Writing tasks for practice sessions on Criterion ................................ 56
Think-aloud protocol ....................................................................... 57
Target level of verbalization ........................................................... 58
Stimulated recall interviews ........................................................... 59

The language of the interviews ......................................................... 60
Focus group interviews .................................................................... 60

Participants ..................................................................................... 61
Participant Inclusion Criteria ........................................................ 63
Participants for pre to post-test accuracy change analyses ............ 63
Participants for revision analyses .................................................... 64
Participants for think-aloud protocols ............................................. 64
Participants for stimulated recall interviews .................................. 65
Participants for focus group interviews .......................................... 65

Ethics .............................................................................................. 65

Data collection procedures ............................................................ 65
Pre-post test collection .................................................................... 68
Criterion training .......................................................................... 69

Procedures for in-class practice sessions ....................................... 70

TAP training procedures ............................................................... 70
TAP collection .................................................................................................................. 71
Stimulated recall interview collection ........................................................................... 71
Focus group interview collection ..................................................................................... 72
Data Analysis Procedures ............................................................................................... 72
Data cleaning .................................................................................................................... 72
Hand-written essays ........................................................................................................ 73
   Retyping ........................................................................................................................ 73
   Coding clauses in essays ............................................................................................... 73
   Coding learner errors ................................................................................................. 74
   Initial checking of coder reliability ............................................................................. 75
   Checking coder reliability ........................................................................................... 75
Target forms in obligatory occasion analyses ............................................................... 76
   Statistical analysis ....................................................................................................... 76
Meeting statistical assumptions ....................................................................................... 77
   Normality, linearity, range ........................................................................................... 78
   Sphericity .................................................................................................................... 78
Student first and revised drafts on Criterion .................................................................... 78
   Verification of Criterion feedback ............................................................................... 78
   Calculation of precision ............................................................................................. 79
   Coding students’ revisions from first to second/final drafts ........................................ 80
   Coding revision practices (i.e., text operations) ........................................................... 80
   Coder reliability analysis ............................................................................................ 82
   Calculation of revision success rate ............................................................................ 82
Think-aloud protocols ........................................................................................................ 83
   Transcription ............................................................................................................... 83
   Coding ........................................................................................................................ 83
   Coder reliability analysis ............................................................................................ 84
Interviews ........................................................................................................................... 84
   Transcription ............................................................................................................... 84
   Coding ........................................................................................................................ 84
   Coder reliability analysis ............................................................................................ 85
Putting Things Together .................................................................................................... 85

CHAPTER 4: Criterion Automated Corrective Feedback: Its Nature and Accuracy
........................................................................................................................................ 88
Part 3. Individual Cases’ Findings

Case 1. Trang, highly efficient reviser through limited engagement with Criterion ACF .................................................. 167
Profile ............................................................................................................. 167
Engagement episodes and revision patterns .............................................. 169
Criterion-initiated engagement episodes .................................................. 169
Self-initiated revisions ............................................................................. 175
Case 2. Nhien, extensively engaged and self-motivated reviser .............. 179
Profile ........................................................................................................................................... 179

Criterion-initiated engagement episodes ................................................................................. 181

Effective utilization of resources for correct revisions ............................................................ 183

Taking the feedback seriously ................................................................................................. 184

Failed attempts to revise ......................................................................................................... 186

Self-initiated revisions .............................................................................................................. 189

Case 3. An, confident and superficially engaged reviser ......................................................... 191

Profile ........................................................................................................................................... 191

Engagement episodes and revision patterns ............................................................................ 192

Focus on grammatical accuracy .............................................................................................. 196

Quick and successful revisions based on stored linguistic resources ................................. 197

Retrieval of learned knowledge .............................................................................................. 199

Handling unhelpful feedback: Incorrect error codes and false positives ......................... 201

Part 4. Cross-case Analyses ..................................................................................................... 205

Summary of Quantitative Findings .......................................................................................... 205

Qualitative Cross-case Analyses ........................................................................................... 207

Similarities ................................................................................................................................. 207

Differences ................................................................................................................................. 208

Different perceptions about revisions ...................................................................................... 208

Differential engagement levels with Criterion ACF ............................................................... 208

Revision patterns ...................................................................................................................... 209

Summary ........................................................................................................................................ 210

CHAPTER 7: Discussion of Findings ......................................................................................... 212

Introduction ................................................................................................................................. 212

Developers’ Claims about Criterion ......................................................................................... 212

Developing an Interpretative Argument .................................................................................. 213

Mapping Findings onto the Validity Argument ....................................................................... 216

Evaluation Inference .................................................................................................................. 217

Accuracy of Criterion ACF ......................................................................................................... 219

Clarity of Criterion ACF ............................................................................................................. 220

Specificity of Criterion ACF ...................................................................................................... 221

Summary on validity evidence for the evaluation inference .................................................... 222

Explanation Inference ............................................................................................................... 223

Range of error types addressed and second language learners’ major issues 224
Word choice: Meaning is still a missing piece in Criterion’s feedback......225
Verb tenses: Limited textual levels are targeted...................................226
Word order.....................................................................................226
Summary of validity evidence for the explanation inference..............227
Utilization Inference .......................................................................228
The comprehensibility of Criterion metalinguistic explanations.........231
Meaningful engagement with Criterion ACF ...................................232
Learner goals ..................................................................................232
The nature of Criterion corrective feedback ....................................233
The cognitive load of evaluating feedback accuracy..........................234
Use of Criterion ACF for revisions ................................................236
Revision outcomes: The error type-feedback type interaction .........237
Summary of validity evidence for the utilization inference ..................240
Ramification Inference ....................................................................240
Increased grammar awareness and mastery of linguistic features ......244
Is increased grammar awareness conducive to mastery of linguistic features?
...........................................................................................................244
The ceiling effect hypothesis ...........................................................244
Counterarguments to the ceiling effect hypothesis .............................245
Positive perceptions about Criterion ACF .......................................246
Maximized learning opportunities ..................................................247
Use of Criterion ACF facilitates self-regulatory revision strategies....248
Summary of validity evidence for the ramification inference ..............250
CHAPTER 8: Conclusions ..................................................................251
Restatement of Research Aims and Key Findings ..............................251
Implications ....................................................................................252
Theoretical implications ..................................................................252
Pedagogical implications ..................................................................254
Implications for Criterion developers .............................................255
Limitations of the Current Research ...............................................257
Recommendations for Future Research ............................................259
References .....................................................................................261
Appendices .....................................................................................273
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. Writing Prompts for Practice Sessions on Criterion</td>
<td>56</td>
</tr>
<tr>
<td>3.2. Themes and Guiding Questions in Focus Group Interviews</td>
<td>61</td>
</tr>
<tr>
<td>3.3. Participants for Different Parts of the Study</td>
<td>62</td>
</tr>
<tr>
<td>3.4. Participants’ Demographics for Pre-post Test Change Analyses</td>
<td>64</td>
</tr>
<tr>
<td>3.5. Data Collection Timeline for the Experimental Group</td>
<td>66</td>
</tr>
<tr>
<td>3.6. Five Most Frequent Error Types in Students’ Essays Tagged by Criterion</td>
<td>76</td>
</tr>
<tr>
<td>3.7. Testing Statistical Difference in Two Groups’ Pre-test Scores for Different Accuracy Measures</td>
<td>77</td>
</tr>
<tr>
<td>4.1. Criterion ACF Feedback on Grammar Errors with Cited Examples from This Study</td>
<td>91</td>
</tr>
<tr>
<td>4.2. Criterion ACF Feedback on Mechanics Errors with Cited Examples from This Study</td>
<td>94</td>
</tr>
<tr>
<td>4.3. Criterion ACF Feedback on Usage Errors with Cited Examples from This Study</td>
<td>98</td>
</tr>
<tr>
<td>4.4. Error Types with Most Frequent Occurrences in Participating Students’ Essays</td>
<td>104</td>
</tr>
<tr>
<td>4.5. Criterion ACF’s Overall Precision</td>
<td>105</td>
</tr>
<tr>
<td>4.6. Verification of Criterion ACF for Specific Error Types (N = 2787)</td>
<td>106</td>
</tr>
<tr>
<td>5.1. Student Engagement Activities through TAP Analyses</td>
<td>114</td>
</tr>
<tr>
<td>5.2. Students’ Textual Operations in Response to Criterion ACF by Error Type121</td>
<td>121</td>
</tr>
<tr>
<td>5.3. Summary of Revision Outcomes</td>
<td>126</td>
</tr>
<tr>
<td>5.4. Revision Success Rates for Most Frequently Recurring Error Types</td>
<td>127</td>
</tr>
<tr>
<td>5.5. Descriptive Statistics of Overall Accuracy</td>
<td>131</td>
</tr>
<tr>
<td>5.6. Descriptive Statistics of Article Use Accuracy</td>
<td>133</td>
</tr>
<tr>
<td>5.7. Descriptive Statistics of Third Person Singular Use Accuracy</td>
<td>135</td>
</tr>
<tr>
<td>5.8. Descriptive Statistics of Copula Use Accuracy</td>
<td>138</td>
</tr>
<tr>
<td>5.9. Descriptive Statistics of Plural s/es Accuracy</td>
<td>140</td>
</tr>
<tr>
<td>5.10. Descriptive Statistics of Comma Use Accuracy</td>
<td>142</td>
</tr>
<tr>
<td>6.1. Accuracy Levels of Three Cases</td>
<td>163</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

**Figure**  

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.</td>
<td>Strategies for providing written corrective feedback (Adapted from Ellis’ 2009 typology of WCF)</td>
<td>13</td>
</tr>
<tr>
<td>2.2.</td>
<td>Min’s 2006 taxonomy of types of revisions with explanations added in parentheses</td>
<td>17</td>
</tr>
<tr>
<td>3.1.</td>
<td>Validation of the efficacy of automated corrective feedback for EFL learners</td>
<td>49</td>
</tr>
<tr>
<td>3.2.</td>
<td>Criterion’s error types based on its scoring engine e-rater v. 11.1</td>
<td>53</td>
</tr>
<tr>
<td>3.3.</td>
<td>Example screenshot of Criterion feedback in pop-up screen</td>
<td>54</td>
</tr>
<tr>
<td>3.4.</td>
<td>Criterion’s split screen for revision activities</td>
<td>54</td>
</tr>
<tr>
<td>3.5.</td>
<td>Connecting multiple data sources with the research questions of the study</td>
<td>86</td>
</tr>
<tr>
<td>4.1.</td>
<td>Numbers of Criterion-generated error tags by individual error type in grammar</td>
<td>101</td>
</tr>
<tr>
<td>4.2.</td>
<td>Numbers of Criterion-generated error tags by individual error type in usage</td>
<td>102</td>
</tr>
<tr>
<td>4.3.</td>
<td>Numbers of Criterion-generated error tags by individual error type in mechanics</td>
<td>102</td>
</tr>
<tr>
<td>4.4.</td>
<td>Error distribution among grammar, usage, and mechanics</td>
<td>103</td>
</tr>
<tr>
<td>5.1.</td>
<td>Students’ text operations in response to Criterion ACF</td>
<td>120</td>
</tr>
<tr>
<td>5.2.</td>
<td>Overall revision outcomes in response to Criterion error tags</td>
<td>125</td>
</tr>
<tr>
<td>5.3.</td>
<td>Overall accuracy from pre-test to post-test and delayed post-test among two groups</td>
<td>132</td>
</tr>
<tr>
<td>5.4.</td>
<td>Changes in article usage from pre-test to post-test and delayed post-test among two groups</td>
<td>134</td>
</tr>
<tr>
<td>5.5.</td>
<td>Changes in third person singular usage from pre-test to post-test and delayed post-test among two groups</td>
<td>136</td>
</tr>
<tr>
<td>5.6.</td>
<td>Changes in copula usage from pre-test to post-test and delayed post-test among two groups</td>
<td>139</td>
</tr>
<tr>
<td>5.7.</td>
<td>Changes in plural usage from pre-test to post-test and delayed post-test among two groups</td>
<td>141</td>
</tr>
<tr>
<td>5.8.</td>
<td>Changes in comma usage from pre-test to post-test and delayed post-test among two groups</td>
<td>143</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

Appendix
A. Pre-Post Test Papers .................................................................273
B. Guidelines for Coding Clauses ......................................................276
C. Coding Errors: General Accuracy Using Weighted Clause Ratio ........279
D. Coding Errors: Local Measures Using Obligatory Occasion Analysis ....282
E. Guidelines for Coding Revision Practices ......................................287
F. Guidelines for Coding Think-aloud Protocols ................................291
G. Ethics Approval Letter ..................................................................295
Chapter 1
INTRODUCTION

The increasingly more pronounced presence of automated writing instructional programs such as Criterion, MY Access, or WritetoLearn has shifted the English as a foreign language (EFL) writing classroom paradigm from the traditional mode to a blended environment where the interplay between teacher instruction and automated scoring and feedback takes place. The purpose of developing these automated tools for the teaching and learning of writing skills, according to Burstein, Chodorow, and Leacock (2004), is allowing students to practice writing more regularly. Going into the specifics of automated feedback, these authors commented,

> It is essential that students receive accurate feedback from the system with regard to errors, comments on potentially undesirable style, and information about discourse elements and organization of the essay. If the feedback is to help students improve their writing skills, then it should be similar to what an instructor’s comments might be”. (p. 32)

These automated writing evaluation (AWE) tools are gaining ground at different educational levels and have been implemented in different contexts. Not only has the advent of these tools excited learners and writing instructors but it has also attracted the attention from researchers in the field of applied linguistics and SLA. More recently, AWE programs and their automated feedback have gradually been investigated by a small body of research that focuses on the impact of AWE implementations on classroom writing instruction and learning (Stevenson, 2016). However, as Stevenson pointed out, AWE scholarship is still more focused on the scoring functions of these systems than their classroom-based written feedback. When the focus is on classroom-based research, its predominant concern is about any improvement in students’ writing quality following the automated feedback provision.

Against this backdrop, the aim of the current research is to validate Criterion automated corrective feedback as a learning and assessment tool in the EFL writing classroom context. In pursuit of this research goal, the study includes 81 Vietnamese EFL university students who had access to the automated corrective feedback generated by ETS Criterion over a semester. The main intervention comes from compulsory in-class practice sessions when students composed and submitted essays on Criterion. Overall, the current research adopted a longitudinal and quasi-experimental design to deepen our understanding about the nature of automated feedback and EFL students’ engagement, response, and perceptions of the feedback.
Significance

This research emerges from a concern that the increasingly prominent presence of automated writing instructional programs in numerous teaching and learning contexts may not be accompanied by adequate understanding of their impact on students’ learning to write in a second language. The review of literature has revealed both a scarcity of research evidence about the impact of automated AWE feedback on accuracy development and retention (Heift & Hegelheimer, 2017) and a paucity of longitudinal studies into learners’ engagement with automated feedback and its impact on writing/revision practices. Zhang (2017) highlighted the importance of the investigation of student engagement with AWE feedback, adding that this aspect has been under-researched in the literature. This point is further reinforced in Stevenson and Phakiti’s (2014) observation based on their meta-analysis that little is known about the effects of AWE feedback on students’ independent revising. The authors also concluded that research into the efficacy of automated feedback is neither ample nor sound in research approaches.

While corporate funded researchers are in a better position to access big data to study how (well) a system scores student essays, more classroom-based research is needed to explore the automated corrective feedback function and students’ engagement with such feedback. Therefore, the current study is conducted to seek insights into learners’ actual interaction with automated form-focused feedback from original perspectives, which hopefully contributes sound implications for both service providers and end-users of automated writing instructional programs. It aims to inform pedagogical decisions about how best to maximize the benefits of AWE feedback as well as provide implications for AWE developers to enhance the value of automated feedback for end-users.

Besides adding to the body of research on the nature of AWE feedback, students’ revisions using the feedback, their accuracy changes over time and perceptions of this new mode of input in learning to write, findings from this study will provide insights into an under-researched strand in automated corrective feedback: Learners’ engagement with the feedback and the interaction of individual factors that mediate their engagement and response to the feedback. Such findings can hopefully extend understanding about automated corrective feedback from a novel perspective and add missing pieces to our knowledge about the impact of this new form of input in the field of written corrective feedback as a whole.

The findings of this study have practical implications for learners, writing instructors, and AWE developers as well. The implications highlight the strengths and weaknesses of the current feedback mechanisms existing in AWE systems like Criterion, and students’ personal experiences as well as expectations when using the feedback. By doing this, the researcher’s ambition is to create a meeting ground where developers and end-users can negotiate and yield a win-win situation. In this situation, learners can maximize their learning through the access to Criterion ACF.
and other similar tools, while AWE developers can use findings as empirical evidence about the end-users’ experiences with the feedback to improve their products.

**Briefly about Criterion**

*Criterion* is among the earliest instructional writing tools used in the classroom to provide automated feedback and scoring (Ramineni & Deane, 2017). It is also the most popular AWE program that has been used in research on automated feedback and scoring. More specifically, in a 2014 review of research in the field of AWE conducted by Stevenson and Phakiti, about one third of all the reviewed studies used *Criterion*. According to its developers’ words:

The *Criterion Online Essay Evaluation Service* combines automated essay scoring and diagnostic feedback. The feedback is specific to the student’s essay and is based on the kinds of evaluations that teachers typically provide when grading a student’s writing. *Criterion* is intended to be an aid, not a replacement, for classroom instruction. Its purpose is to ease the instructor’s load, thereby enabling the instructor to give students more practice writing essays. (Burstein, Chodorow, and Leacock, 2004, p. 27)

In actual fact, *Criterion* is composed of two complementary applications. The first application is the automated essay scoring and evaluation machine *e-rater*, while the second is a suite of programs providing feedback on errors in grammar, usage, mechanics, style, organization, and development in learner writing called *Critique* (Burstein, Chodorow, & Leacock, 2003). Both of these applications were developed based on natural language processing (NLP) methods and statistical machine learning techniques. While details on *Criterion* feedback, especially its form-focused feedback including grammar, usage, and mechanics, are provided in Chapters 2, 3, and 4, for now, Ramineni and Deane’s (2017) description gives us a brief guide as to how *Criterion* provides scoring and feedback to student essays.

The holistic scoring and feedback provided by *Criterion* is supported by the *e-rater* automated scoring engine. Built-in scoring models are provided, based on a representative sample of student responses written under high-stakes testing conditions and scored by trained human raters. Each model has been evaluated on an independent testing sample to confirm that there is close agreement between the automated and human scores…. Underlying all of *e-rater*’s features and advisories are NLP algorithms trained on human annotated data. These algorithms are evaluated on how well they perform to identify errors in an annotated corpus of student essays. (pp. 167-169)
Despite being the most researched NLP system among all other AWE programs, there is still a lack of insights into students’ interaction with Criterion or its general performance (Lavolette, Polio, & Kahng, 2015). Against this backdrop, the current study narrows its focus on the automated corrective feedback generated by Criterion. This includes the feedback on grammar, usage, and mechanics in student essays, while Criterion feedback on style, organization and development was suppressed throughout the research process. This research focus is informed by previous research on AWE feedback which has clearly pointed out a major shortcoming of automated feedback: its inability to comprehend meanings in essays the way a human reader does. Regarding specific AWE systems, automated writing instructional programs generally have great difficulty reading the meaning in student essays, as shown in previous studies on WritetoLearn (Liu & Kunnan, 2016), MY Access (Hoang & Kunnan, 2016), or Criterion (Dikli, 2006). This probably explains one of the common findings in previous research which shows teachers and learners’ generally more positive perceptions about AWE error feedback than the automated feedback on other aspects or automated scoring (Stevenson, 2016).

Therefore, at this stage in the developmental history of automated essay evaluation, it would be too much to expect of these systems to “read” essays at the discourse level as human readers do. Instead, recent improvements in these systems at diagnosing problems at the sentential, phrasal, and lexical levels in student essays (i.e. their grammatical or corrective feedback) are worth exploring to see how such feedback helps students self-edit and revise their writing. More specifically, the conclusion that automated feedback on essay organization, content, and development tends to be generic, repetitive, and thus unhelpful for revisions and learning has been echoed in a few studies (e.g., Dikli, 2010; Li, Link, & Hegelheimer, 2015; Warschauer & Grimes, 2008; Yang, 2004). For example, teachers participating in Li et al.’s (2015) research commented that the feedback on content was not meaningful as it discouraged creativity in writing. Added to this, in examining MY Access, Mohsen and Alshahram (2019) reported that their participants had difficulty understanding the system’s content and organization feedback in their essays but experienced no issue with its feedback on their writing accuracy.

The decision to focus on the automated corrective feedback (ACF) generated by Criterion is also empirically founded on learners’ actual use of the feedback. For example, Warschauer and Grimes (2008) found that few of the middle school students participating in their study spent time reading the content and organization feedback. While Warschauer and Grimes’ study was conducted with middle school students whose mother tongue is English, the situation experienced by many English language learners is that their interlanguage development has not progressed to the point where they can be free from linguistic accuracy concerns to focus exclusively on content and fluency in writing essays in the target language. This is probably why Chen, Chiu, and Liao (2009)
proposed that “for many intermediate-level ESL learners, immediate written grammar feedback is the key feature why they want to subscribe to these AWE programs because they still have great difficulty writing error-free sentences” (p. 6). Therefore, for a study that seeks insights into how the automated feedback has impacted students’ learning to write in the target language as the current research, it seems reasonable and practical to focus attention on an aspect of the automated feedback that has been of actual use among learners.

The next section elaborates on the different chapters making up the whole thesis.

**Thesis Structure**

The current research is comprised of eight chapters, including this introductory chapter, the review of the literature, methodology, three chapters of findings, discussion, and conclusions. Below are the specifics of each of the eight chapters.

**Chapter 1: Introduction.** This chapter provides an overview of the thesis by familiarizing readers with the background of the topic, a brief introduction about Criterion, the significance as well as rationale for the study focus on automated corrective feedback, and the overall thesis structure.

**Chapter 2: Literature Review.** This chapter reviews scholarship on teacher feedback, followed by an extensive literature on automated feedback and scoring by different automated essay evaluation systems. The central part of the literature review is on previous research on automated corrective feedback. The five research questions of the study are presented towards the end of this chapter.

**Chapter 3: Methodology.** In this chapter, detailed information about how the study was conducted is given. This includes the explanation of the procedures of the research, introduction of the theoretical framework used to structure the discussion of findings, participant recruitment, research instruments, and procedures of data collection and analyses.

**Chapter 4: Criterion Automated Corrective Feedback: Its Nature and Accuracy.** This chapter presents findings related to the first two research questions of this study. Both qualitative and quantitative analyses of the automated error tags and accompanying metalinguistic explanations generated by Criterion were conducted to provide detailed information regarding the nature and accuracy of the feedback.

**Chapter 5: Learner-centric Perspectives on Criterion ACF: Learners’ Engagement, Revisions, and Perceptions.** Following the introduction about the characteristics of Criterion automated corrective feedback in Chapter 4, students’ engagement – cognitive, behavioral as well as affective response to such feedback is presented. This is triangulated with their observed changes
in writing accuracy after five months accessing Criterion in their writing course. This chapter, therefore, provides the answer to Research Questions 3, 4, and 5 of the current study.

Chapter 6: The Multiple-case Study. Unlike the previous chapters, Chapter 6 can stand alone as a mini study itself with its own rationale, participant selection procedures, data analyses, and findings. This chapter elaborates on the embedded multiple-case study that extends understanding about individual learners’ engagement, perceptions, and subsequent response to the feedback, as well as their developmental trajectories over the studied period. Thus, this chapter seeks to reveal explanatory factors through making the connections between each individual case’s engagement with the feedback with their observed changes over time. In order to do this, it presents three narratives about three different learners carefully selected from the pool of participants for the current research.

Chapter 7: Discussion of Findings. This chapter brings all the pieces of evidence together in a coherent discussion which adopts an argument-based approach to validating the value of Criterion automated corrective feedback as a learning and assessment tool in the EFL writing classroom context. By the time readers reach this chapter, they have transitioned from learning about Criterion feedback in Chapter 4 to knowing details about learners’ engagement and response to the feedback in Chapter 5, to gaining insights into individual learners’ stories in Chapter 6. The discussion is a synthesis of all these pieces to guide conclusions about Criterion ACF.

Chapter 8: Conclusions. This final chapter wraps up the thesis by presenting the empirically based conclusions before offering relevant theoretical and pedagogical implications for the use of Criterion ACF in an EFL context. Limitations of the current research are then presented, and the recommendations for future research to address the limitations and fill major gaps revealed in the literature review bring the thesis to an end.
Chapter 2

LITERATURE REVIEW

The overall assumption underlying this research is that high quality corrective feedback (CF) helps students improve their linguistic accuracy as it is hoped that corrective feedback draws learners’ attention to their errors and this helps them produce correct forms subsequently (Storch & Wigglesworth, 2010). In his reflection on the potentials of written corrective feedback, Bitchener (2012) also made the observation that available research supports the view that written corrective feedback is beneficial for L2 learning and development, although “it is not clear how extensive its role might be” (p. 360). This review of the literature presents relevant theoretical concepts regularly used in research on written feedback in second language writing, followed by a more extensive discussion about teacher feedback which provides methodological frameworks and approaches for investigating the feedback generated by automated writing evaluation (AWE) programs. The chapter concludes with reviewing the limitations of previous research on AWE as well as recommendations offered about future research on automated corrective feedback which informed the focus of the current study. Finally, the research questions for this research are presented to bring the chapter to an end.

Theoretical Constructs

The two most relevant theoretical constructs are going to be presented in this section: engagement and noticing. While both of these constructs can be applied more broadly to all classroom activities, only their application to teacher and automated feedback, especially corrective feedback, in second language writing will be the focus.

Engagement. Frequently mentioned in the discussion about written corrective feedback is the notion of engagement. Broadly speaking, engagement refers to learners’ interest and participation in an activity (Philp & Duchesne, 2016). Zhang and Hyland (2018) further defined engagement as the extent to which students are devoted to their learning, which is demonstrated in their responses to texts and attitudes to writing. According to these authors, engagement “is an umbrella term which brings together students’ degree of attention, curiosity, interest and willingness to employ their language proficiency and a repertoire of learning skills to make progress” (p. 91).

More specifically addressing oral and written corrective feedback research, Ellis (2010) used engagement to indicate learners’ response to the feedback they received and proposed a three-perspective approach to investigating learners’ engagement with the written feedback, including cognitive, behavioural, and affective perspectives. Following Ellis’ (2010) three-perspective
framework, Zhang and Hyland (2018) considered engagement with feedback as composed of three dimensions and operationalised each of them as:

(i) *Cognitive engagement* which indicates students’ paying attention to the feedback through the utilization of cognitive and metacognitive strategies.

(ii) *Behavioural engagement* which refers to students’ reaction to the feedback including revision operations.

(iii) *Affective engagement* which relates to students’ emotional and attitudinal responses to the feedback.

Also building on Ellis’ (2010) framework for engagement with corrective feedback, Han and Hyland (2015) conceptualize the three dimensions of engagement with written corrective feedback, stressing the importance of noticing for learners to engage with the feedback. In their conceptualizations, cognitive engagement is characterised by learners’ employment of cognitive strategies such as memorization or visualization, while meta-cognitive strategies also form a critical aspect of the cognitive engagement process with regulatory strategies like feedback evaluation. Behavioural engagement deals with not only revisions (or uptake of the written corrective feedback) but also utilization of learning/revising strategies that facilitate the processing of written corrective feedback. Regarding the affective dimension, the authors noted the importance of the dynamic nature of students’ affective engagement with WCF.

These three dimensions of cognitive, behavioural, and affective engagement combine to help students effectively respond to the feedback they receive. Previous research has also provided empirical evidence which suggests that “student engagement with written corrective feedback facilitates language acquisition and writing development” (Zhang & Hyland, 2018, p. 91).

**Noticing.** While learner engagement is open to multiple interpretations and is comprised of several sub-constructs (Han & Hyland, 2015), probably the most thoroughly researched sub-construct in the literature is the depth of processing which relates to the cognitive dimension in Ellis’ (2010) multi-dimensional framework. The cognitive dimension of the three-perspective approach to investigating written corrective feedback is concerned with the question about how deeply WCF is processed, which has previously been researched in several studies (e.g. Kim & Bowles, 2019; Qi & Lapkin, 2001; Sachs & Polio, 2007; Storch & Wigglesworth, 2010). Following Han and Hyland (2015), this includes both cognitive operations students conduct to process the feedback and the meta-cognitive strategies. While cognitive operations relate to learners’ mental efforts to consider how and to what extent their texts can be revised, the meta-cognitive level of engagement is manifest in students’ monitoring and regulation of these mental efforts to process WCF. Additionally, what indicates the depth of feedback processing is learners’ awareness of the
WCF which is demonstrated at two levels of noticing or understanding of the feedback. The construct of noticing which emphasizes the cognitive dimension of engagement originates from Schmidt’s (1993, 1995) noticing hypothesis. Noticing, which Schmidt associated with paying attention to or detection, is when the learners notice some mismatch between their current interlanguage output and the target input, while the higher level of attention is awareness and understanding which implies “recognition of a general principle, rule, or pattern” (Schmidt, 1995, p. 29).

Of the two levels of awareness, Schmidt (1993, 1995) hypothesized that paying attention or noticing is a condition for the learner to convert L2 input to intake. Extending on Schmidt’s hypothesis, Ellis (2010) and Zheng and Yu (2018) made the point that a prerequisite for corrective feedback to have any effect on a learner’s interlanguage development is his/her paying attention to it. When students deliberately pay attention to some language input as in the form of corrective feedback, what matters is noticing. It is noticing that becomes the prerequisite for any further learning stages to take place.

In second language writing, noticing has been investigated as it takes place either during the production/composing stage or when learners process feedback and conduct revisions. According to Swain and Lapkin (1995), “noticing is also an important cognitive process in L2 composing” (p. 278). That is, while composing in the target language, students ‘notice’ a problem, which encourages them to conduct an analysis to arrive at modified output. In the absence of external feedback, learners are still able to go through simple or complex thinking while generating alternatives, assessing them, and producing revised output. Swain and Lapkin (1995) proposed that “what goes between the first output and the second … is part of the process of second language learning” (p. 386). Turning to the revising stage with the provision of external feedback as in the form of teacher written feedback, noticing plays a significant role in facilitating interlanguage development by drawing learners’ attention to a gap in their language production. Students process the feedback they receive on a linguistic problem they have and modify their output accordingly to arrive at the revised output. In short, according to Swain and Lapkin (1995), the first type of noticing is noticing mismatches between what they have written and the target language forms, while the second type relates to learners’ noticing the gaps or ‘noticing the holes’ when composing which primes them to notice the corrective feedback they subsequently receive on their writing.

The role of corrective feedback as a noticing enhancer has been extensively researched in second language writing using different approaches and in light of various hypotheses in second language acquisition. In particular, interactionist perspectives view corrective feedback as a source of external feedback which triggers learners’ conscious awareness about a gap in their interlanguage development. With noticing being among the central tenets of the interactionist perspective when
discussing the effects of corrective feedback, this research takes the theoretical stances offered by Schmidt’s noticing hypothesis (1990, 1993) as a guide for the investigation into students’ engagement with automated corrective feedback (ACF).

Schmidt’s construct of noticing has been open to different operationalizations. For example, Qi and Lapkin (2001) defined noticing as “awareness of a stimulus via short-term memory” (p. 279). In order to capture students’ depth of processing the feedback when investigating the role of noticing in promoting L2 development, Qi and Lapkin (2001) operationalized the quality of noticing as “perfunctory (i.e., noticing only and without giving reasons) or substantive (i.e., noticing and providing reasons)” (p. 291). Their findings suggest the critical role of the quality of noticing for the improvement of students’ L2 writing through feedback provision in the form of reformulated versions of learners’ written texts. The results also pointed out that higher proficiency students were more likely to substantively notice the gaps in their output and accept more reformulated items or structures than their lower proficiency counterparts, suggesting a correlation between the quality of noticing and enhanced likelihood of linguistic acquisition.

Partially replicating Qi and Lapkin’s (2001) analysis of noticing, Sachs and Polio (2007) developed a more extensive list of categories related to the quality of noticing and aligned these categories with Schmidt’s (1991, 1993) and Robinson’s (1995) two levels of awareness: awareness at the level of noticing/detection versus understanding. Two categories grouped under the ‘awareness with understanding’ level, use of metalanguage and provision of a reason, are representative of Qi and Lapkin’s (2001) substantive kind of noticing. Sachs and Polio (2007) further coded the changes (or the lack thereof) associated with these two categories and their findings corroborate Qi and Lapkin’s (2001) result that substantive noticing was highly conducive to a revised form in learners’ texts.

Similar to Qi and Lapkin (2001) and Sachs and Polio (2005), the quality of noticing was operationalized as extensive or limited engagement in Storch and Wigglesworth’s (2010) investigation of the nature of student engagement with two forms of written feedback: editing symbols (indirect feedback) and reformulations. Yet, departing from the use of individual learners’ recorded think-aloud protocols when processing written feedback, Storch and Wigglesworth (2010) collected audio recordings of pair talks when pairs of learners collaborated in their writing tasks. In this study, language-related episodes detected in pair talks were coded for extensive engagement (i.e., “learners offered suggestions and counter-suggestions, explanations, or any comments that showed evidence of meta-awareness of the feedback received” (p. 311)) or limited engagement (i.e., “one member of the pair simply read the feedback and the other merely acknowledged and repeated it” (p. 311)). Findings in previous research are reinforced by Storch and Wigglesworth’s (2010) results that among other factors, students’ levels of engagement with the feedback has some
correlation with the recorded uptake levels, and that students’ depth of engagement may vary with the types of error. Specifically, extensive engagement with the feedback tended to lead to uptake, while limited engagement was less conducive to uptake, and thus less helpful for error corrections.

Depth of feedback processing was also investigated by Kim and Bowles (2019) with a sample of 22 adult ESL learners who attended an academic writing course in the US. Not directly using the term noticing, these researchers operationalized the students’ processing of the two forms of written feedback, direct correction versus reformulation, as high depth of processing and low depth of processing. Each level of processing was clarified by a list of criteria, but generally these two levels are characterized by similar features to Qi and Lapkin’s (2001) substantive/perfunctory noticing categories or Storch and Wigglesworth’s (2010) extensive/limited engagement. The study found interactions between feedback type and students’ depth of processing as well as between error types and depth of processing. Specifically, reformulations were associated with more reports of high depth of processing episodes than direct corrections. Also, surface-level errors were likely to be processed at a low depth of processing.

Although noticing has taken various forms, there seems to be some common features shared by all the reviewed studies. Firstly, noticing has been sub-divided into two levels representing either cursory or deep processing of the written corrective feedback. Secondly, these studies have focused on the cognitive perspective in Ellis’ (2010) three-dimensional framework of engagement as a starting point to connect cognitive processing with behavioral or affective dimensions in an effort to explain observed changes. In the current research, noticing also serves as a foundational construct upon which students’ engagement with automated corrective feedback is operationalized.

**Written Corrective Feedback**

Written corrective feedback is a well-researched area informed by both theoretical foundations and empirical evidence. The following sections present theoretical perspectives most relevant to the discussion of written CF, followed by empirical studies that provide evidence regarding the effectiveness of different types of written corrective feedback.

**Written corrective feedback and theoretical foundations in SLA.** The use of written corrective feedback is based on several theoretical hypotheses. Worth mentioning is Swain’s (1985, 1995) Output Hypothesis which stresses the importance of output in helping learners notice gaps in their production. Most relevant to the discussion regarding corrective feedback is Swain’s (1993) acknowledgement about the limitations of output alone, pointing to the supplementary role of corrective feedback in aiding learners go through mental processes to modify their output more effectively. Similarly, Schmidt’s (1990, 2001) Noticing Hypothesis suggests that acquisition takes
place when learners pay conscious attention to the input they receive. Schmidt (2001) emphasizes the importance of paying attention to form as a way to help learners notice the gaps in their interlanguage, which paves the way for input to become intake. Despite the absence of any direct reference to corrective feedback in Schmidt’s (2001) Noticing Hypothesis, corrective feedback as a pedagogical tool to draw learners’ attention to the gaps in their output can serve to trigger the process of internalizing the target language forms. This finds support in Long’s (2009) focus on form perspective which supposed that implicit learning alone is not sufficient and that focus on form is necessary in language teaching and learning (cited in Pérez-Núñez, 2015). Focus-on-form interventions as in the shape of corrective feedback can thus steer leaners’ attention to gaps in their interlanguage development.

Besides theoretical hypotheses that provide support for written corrective feedback, there are arguments against its use. One of the most compelling arguments against WCF is a debate initiated by Truscott (1996) who points out problems that invalidate the use of corrective feedback, including the lack of systematic and consistent approaches to delivering error correction among teachers as well as the ability and willingness to pay attention to the feedback among learners. Time-consuming and tedious as it is, according to Truscott who based his arguments on the analysis of a series of relevant empirical studies, the practice of error correction does not benefit learners and thus should be abandoned. Truscott’s (1996, 2004) arguments find partial support in Krashen’s five hypotheses in second language acquisition. Firstly, Krashen’s (1982) learning-acquisition distinction “predicts that error correction has little or no effect on acquisition, although it might be useful for conscious learning” (Pérez-Núñez, 2015, p. 16). This lack of impact on learning is further reinforced by Krashen’s (1982) Natural Order Hypothesis which claims that language rules are acquired in a predictable order. Following this hypothesis, pedagogical interventions if not properly delivered are not able to contribute to language acquisition. The only way for learners to acquire a language is understanding messages through receiving comprehensible input, which is codified as $i + 1$ in Krashen’s (1982) Input Hypothesis. Krashen’s hypotheses did not make any direct reference to error correction or grammar instruction in general, but there is strong indication of the lack of value for these pedagogical tools in second language acquisition, as learners acquire a language by progressing along the natural order when they receive comprehensible input.

**Empirical studies on written corrective feedback.** The bulk of research on teacher feedback has focused on written corrective feedback (WCF), which is teacher response to linguistic errors found in student essays. Ellis (2009) presented a relatively extensive typology of WCF (see Figure 2.1). In this Figure, Ellis divided written corrective feedback into three different types, including direct corrective feedback which is when the writing instructor provides direct error correction in
student writing. The feedback provides the replacement of the erroneous forms with the correct forms. In contrast to this, the second type of corrective feedback is indirect. Indirect feedback can either be a general indication in the margin that an error has taken place in students’ writing or the error being underlined but not accompanied by a correct form. The third type is metalinguistic explanations, which can be error codes written in the page margin to tell students the types of error they have committed. Another way to provide metalinguistic feedback is that the teacher can number the errors in the texts, which is followed by grammatical explanations for each numbered error at the bottom of the page. The second major column of the figure provides two different scopes of WCF: focused versus unfocused. Focused CF includes feedback on only a limited number of targeted error types, while unfocused or comprehensive CF covers all or most error types. Two other types of CF mentioned in Figure 2.1 are electronic feedback and reformulation. The former refers to teachers’ practice of providing hyperlinks for specific errors they have indicated in student texts to which students can refer as a resource for error correction, while the latter provides native or near-native speaker rewritten texts based on learners’ original writing to improve linguistic expressions yet still keeping the meaning.

<table>
<thead>
<tr>
<th>Type of CF</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Direct CF</td>
<td>The teacher provides the student with the correct form</td>
</tr>
<tr>
<td>2 Indirect CF</td>
<td>The teacher indicates that an error exists but does not provide the correction.</td>
</tr>
<tr>
<td>a. Indicating + locating the error</td>
<td>This takes the form of underlining and use of cursors to show omissions in students’ texts.</td>
</tr>
<tr>
<td>b. Indication only</td>
<td>This takes the form of an indication in the margin that an error or errors that have taken place in a line of text.</td>
</tr>
<tr>
<td>3 Metalinguistic CF</td>
<td>The teacher provides some kind of metalinguistic clue as to the nature of the error.</td>
</tr>
<tr>
<td>a. Use of an error code</td>
<td>Teacher writes codes in the margin (e.g. ww = wrong word; art = article).</td>
</tr>
<tr>
<td>b. Brief grammatical descriptions</td>
<td>Teacher numbers errors in text and writes a grammatical description for each numbered error at the bottom of the text.</td>
</tr>
<tr>
<td>4 The focus of the feedback</td>
<td>This concerns whether the teacher attempts to correct all (or most) of the student’s errors or selects one or two specific types of errors to correct. This distinction can be applied to each of the above options.</td>
</tr>
</tbody>
</table>
a. Unfocused CF  
Unfocused CF is extensive

b. Focused CF  
Focused CF is intensive

<table>
<thead>
<tr>
<th></th>
<th>5 Electronic feedback</th>
<th>The teacher indicates an error and provides a hyperlink to a concordance file that provides examples of correct usage.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 Reformulation</td>
<td>This consists of a native speaker’s reworking of the student’s entire text to make the language seem as native-like as possible while keeping the content of the original intact.</td>
</tr>
</tbody>
</table>

*Figure 2.1. Strategies for providing written corrective feedback (Adapted from Ellis’ 2009 typology of WCF).*

Second language writing and SLA research on WCF has consistently pursued the question whether the feedback is effective for students to improve their grammatical accuracy. More specifically, previous research has investigated the efficacy of different types of WCF and can be sub-grouped into those that examined the type of feedback (i.e., direct vs. indirect WCF) and the scope of feedback (i.e., focused vs. unfocused/comprehensive WCF). Other factors that have been investigated as mediating the efficacy of WCF include learner proficiency, the study setting, the number of treatments (the time duration of the treatment), or the genres of the writing tasks.

*Research into different types of WCF.* There is a lot of empirical research into the impact of direct and indirect corrective feedback (e.g. Benson & DeKeyser, 2018; Chandler, 2003; Shintani & Ellis, 2013; Van Beuningen, de Jong, & Kuiken, 2008; Van Beuningen, de Jong, & Kuiken, 2012). Most of these studies have compared the effectiveness of direct error correction with indirect feedback or metalinguistic explanations. Certain studies even focused on the more nuanced comparison between different forms of direct WCF as in the case of a study by Bitchener and Knoch (2009). In a meta-analysis conducted by Kang and Han (2015), a total of 21 studies on WCF were examined to provide the result that previous research tended to show that WCF is helpful for the improvement in L2 learners’ written accuracy. In addition, their meta-analysis did not show any clear difference in the effects of the two types of WCF, direct versus indirect. In trying to explain this lack of differential effects, the authors hypothesized that in more context-specific cases, the type of WCF is likely to interact with other factors such as learner proficiency to produce a certain impact. A more recent review of the literature by Harrasi (2019) also suggests inconclusive empirical evidence regarding the comparative effects of direct versus indirect corrective feedback. Harrasi’s (2019) conclusions accord with early research studies by Semke (1984) or Robb, Ross, &
Shortreed (1986) which found no significant difference in the effects of these two types of corrective feedback. Similarly, Mubarak’s (2013) study of the effectiveness of direct and indirect WCF on Bahrain learners’ measures of accuracy and complexity found no effect of either feedback strategy on student writing. Unlike Mubarak’s (2013) choice of a comprehensive feedback approach, Hosseiny (2014) targeted only definite and indefinite article use in his study which compares the effects of direct and indirect CF. The study revealed that the two treatment groups outperformed the control group, but no statistically significant difference between the direct and indirect CF groups was found.

There is also empirical evidence in favour of indirect over direct written CF. Laland (1982), for example, found that the intermediate learners of German who received error codes (indirect feedback) experienced more gains in grammatical accuracy compared to their direct correction counterparts. In a similar vein, Eslami’s (2014) study of low-intermediate Iranian EFL learners revealed that the indirect CF group experienced more sustained improvement in simple past tense usage compared to the direct CF students. As opposed to such studies, Chandler (2003) reported superior effects of direct corrections in helping learners during the revision stage. Van Beuningan et al.’s studies (2008, 2012) also showed positive effects for both direct and indirect WCF, although only the direct CF group sustained their improved accuracy in the new composition in the 2008 study.

*Research into different scopes of WCF.* Turning to the differential effectiveness of focused and unfocused written CF (e.g. Ellis, Sheen, Murakami, & Takashima, 2008; Farrokhi & Sattarpour, 2012; Sheen, Wright, & Moldawa, 2009; Sun, 2013), some studies have provided empirical evidence in favor of focused feedback over unfocused feedback (Ellis, 2009; Farrokhi & Sattarpour, 2012; Harrasi, 2019). This is attributed to the value of focused feedback in its ability to help learners more easily recognize the gaps between their output and the target forms (Ellis, 2009), which is reinforced by theoretical perspectives such as Schmidt’s (1990, 2001) Noticing Hypothesis or Krashen’s (1982) Natural Order Hypothesis. Focused WCF allows learners’ attention to be directed to specific linguistic features that are most relevant to their developmental stage while their cognitive load is also substantially reduced. Ellis et al. (2008) researched the comparative effects of focused written CF (article errors) and comprehensive feedback on Japanese intermediate students. No statistically significant difference between the two feedback approaches was found, but sustained improvement in accuracy existed for the focused group only. Sheen et al. (2009) studied direct written CF but further divided it into focused (article errors only) and unfocused feedback on ESL learners. The key finding was that focused written CF was statistically more effective than the unfocused feedback among these learners.
Research into the effectiveness of WCF on different types of linguistic errors. As Harrasi (2019) pointed out, most published research on WCF examines English articles. Several studies reported favourable effects of WCF on the use of English articles among the learners (Bitchener et al., 2005; Bitchener & Knoch, 2009; Ellis et al., 2008; Sheen, 2007). Conflicting results were found by Shintani and Ellis (2013) and Shintani, Ellis, and Suzuki (2014) who found that written corrective feedback did not help learners improve in accuracy when using the indefinite article. Other error types included the simple past tense which was investigated by Bitchener et al. (2005) and Rummel (2014), both of which found the feedback to be effective. Less responsive to WCF are idiosyncratic structures like prepositions. In Bitchener et al.’s (2005) study, preposition errors were reported to be less treatable under the impact of written feedback. Similar findings were revealed in Guo’s (2015) examination of the effectiveness of WCF on prepositions of space. Generally, the reviewed literature suggests that rule-based structures tended to be more responsive to WCF than less rule-governed ones (Harrasi, 2019).

Overall, the broad spectrum of research foci, treatments, populations, and research designs in this field makes any clear conclusions and generalizations elusive (Hyland & Hyland, 2006). While most empirical research to date tends to suggest that WCF, if appropriately delivered, can facilitate improved writing accuracy (Ferris, 1999, 2003; Kang & Han, 2015), a few researchers have challenged the use of WCF as an instruction approach, debating its value in helping students improve grammatical accuracy. One of the most prominent researchers as such is Truscott (1996, 2004) who consistently voiced his strong disapproval of the benefits of WCF provision. However, the further the field progresses, the more the second language writing community seems to approximate an agreement forwarded by Li, Link, and Hegelheimer (2015) that if non-native writers can improve their linguistic accuracy, this can benefit their expression of ideas. Therefore, as these authors pointed out, the critical question that needs pursuing at this stage is finding the ways to provide students with the most effective CF instead of negating its role in second language acquisition.

Student Response to WCF

Studies into student response to teacher written feedback have investigated student engagement (or processing), uptake, and retention of the feedback. The following sub-parts review research in these three different aspects of student response.

Engagement. Students’ levels of engagement with written corrective feedback relate to the cognitive engagement perspective in Ellis’ (2010) three-dimensional framework. Despite the challenge to access learners’ internal cognitive processes, different research instruments have been
used to elicit information about learners’ engagement, including concurrent think-aloud protocols (e.g. Qi & Lapkin, 2001; Zhang & Hyland, 2018), pair talks (when two students together process and respond to the teacher feedback on their joint written product) (e.g. Storch & Wigglesworth, 2010), as well as the more traditional tools such as questionnaires and interviews. Most studies which examined student engagement with the written feedback provided by the writing instructor focused on students’ depth of feedback processing, as earlier reviewed in sections about engagement and noticing.

**Uptake.** Unlike the qualitative nature of feedback processing, students’ uptake of written corrective feedback is more often measured quantitatively by counting revision points using a framework of categorizing students’ types of revision operations. Min’s (2006) framework for analyzing types, sizes, and functions of revisions remodelled based on that of Sengupta (1998) is presented below as an example of a typology for analysing students’ behavioural engagement with written feedback. According to this framework, there are seven types of text-based changes in response to feedback a student can make, as in Figure 2.2. This includes adding, deleting, substituting, rephrasing, or reordering information in the texts. In addition, the learner can also rewrite the information using larger chunks of language or put together previously separate information, what Min (2006) termed distributions and consolidation, respectively.

<table>
<thead>
<tr>
<th>Surface Changes</th>
<th>Text-Based Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microstructure</td>
<td>Macrostructure</td>
</tr>
<tr>
<td>Changes</td>
<td>Changes</td>
</tr>
</tbody>
</table>

**Additions** *(Reviser adds information)*  
**Deletions** *(Reviser deletes information)*  
**Substitutions** *(Reviser substitutes information)*  
**Permutations** *(Reviser rephrases information)*  
**Distributions** *(Reviser rewrites same information in larger chunks)*  
**Consolidations** *(Reviser puts separate information together)*  
**Reorderings** *(Reviser moves information)*

*Figure 2.2. Min’s 2006 taxonomy of types of revisions with explanations added in parentheses.*
Retention. Despite being a central consideration in most second and foreign language writing classes (Hyland & Hyland, 2006, 2019), earlier studies in the field of teacher feedback were criticized for failing to capture students’ longitudinal changes on new pieces of writing to reflect retention of the feedback. Instead, these studies investigated the efficacy of written corrective feedback using first and revised drafts (i.e. uptake) of the same compositions (Benson & DeKeyser, 2018). Specifically, retention in written corrective feedback can be defined as the long-term impact of the feedback (Storch & Wigglesworth, 2010) or students’ gains in writing accuracy demonstrated in their new compositions. The investigation of retention following the use of written corrective feedback can provide us with a more complete picture of the impact of WCF on students’ writing development.

The measurement of writing development as retention of written corrective feedback can be aligned with efforts to capture second language learners’ general writing development over time. Probably the most popular tools in use in research on writing developments is the triad of complexity, accuracy, and fluency measures (or CAF measures for short). Larsen–Freeman’s (2006) and Storch’s (2009) study are clear examples of how different quantitative measures of CAF are used. Storch (2009), to quantitatively measure accuracy in her study on how studying in an ESL-environment university affects students’ L2 writing development, calculated the number of error-free clauses per clauses (EFC/C), the number of errors per words (E/W), and error-free T-units per T-units (EFT/T). The three measures she used for fluency include the number of words (W), number of T-units (T), and words per T-units (W/T). Complexity was sub-divided into grammatical complexity and lexical complexity, each of which had its own quantitative measures. The former was evaluated by the number of clauses per T-units (C/T) and the number of dependent clauses per clauses (DC/C), while the latter was based on the percentage of academic words used in an essay compared to the Coxhead (2000) Academic Word List and the frequency count of informal expressions in essays.

Similar measures were found in Larsen-Freeman’s (2006) longitudinal investigation of five Chinese EFL students who developed their oral and written production over a ten-month period in a language class in the United States. For accuracy, she used only one quantitative measure that is also used by Storch (2009), the proportion of error-free T-units to T-units (EFT/T). Fluency was manifest through the number of words per T-units (W/T), another measure shared by Storch. Also similar to Storch’s study, complexity was comprised of grammatical (by the average number of clauses per T-units (C/T)) and lexical complexity (by what she called “a sophisticated type-token ratio – word types per square root of two times the words” (p. 596)). Although these two studies did not investigate changed writing development following the use of corrective feedback, their use of
the CAF measures provides a helpful way to quantify observed changes in learners’ linguistic
performance as retention of written corrective feedback.

Over the last few decades, written feedback has always been of great interest to both SLA and
second language writing researchers, although the research focus has gradually shifted. The field
has steadily progressed and one of the recent directions it has taken is the advent of automated
feedback. As an extension to the more established field of teacher feedback, research on automated
feedback has benefited substantially from the available theoretical frameworks as well as research
tools. The following section provides an extensive review of the literature on automated writing
evaluation research, focusing particularly on automated corrective feedback which forms the central
issue in the current study.

**Automated Writing Evaluation**

Starting off with Ellis Batten Page’s *Project Essay Grade* (PEG) proposal in 1965, AWE has
not always experienced constant development efforts and attention. Wresch’s (1993) review of the
earliest studies on the use of computers to evaluate essay responses, 28 years after Page’s PEG
proposal, was discouraging:

> The fact that 25 years have now passed since the original studies were published and no high
> schools or colleges use computer essay grading sums up the reaction to these studies.
> Statistics or no, there is little interest in using computers in this way. (p. 45)

However, recent developments in the field of automated essay evaluation and automated
feedback have proved Wresch’s comment wrong. Beginning with the comeback of research efforts
to upgrade PEG from 1994 to 2000, latecomers in the field soon followed, including ETS *e-rater*,
Pearson *Intelligent Essay Assessor*, and Vantage Learning *IntelliMetric*. More recently, freely
available programs for providing feedback to learner essays such as *Grammarly* and *Pigai* have also
joined the field with increasing numbers of users.

Automated essay scoring systems were originally created to serve high-stakes tests of writing
proficiency to generate summative scores for assessment purposes (Stevenson & Phakiti, 2014). In
fact, as Deane (2013) suggested, the writing construct assessed by current AWE systems
approximates the construct implicit in standardized testing situations. More recently, attention has
been equally focused on automated feedback, including corrective feedback on mechanics,
grammar, usage, and discourse feedback on rhetorical choices, lexical and grammatical complexity
or even content as another critical application of AWE systems. This has made AWE a more robust
tool for the teaching, learning, and assessment of writing not only of non-native writers of English but also of native speakers.

An example AWE system is Criterion, which is an online service developed by Educational Testing Service.Criterion helps students plan, write, and revise essays. It has at its core the automated essay evaluation and scoring machine e-rater which generates holistic scores on essays and the Critique Writing Analysis Tools which provide diagnostic feedback to several types of errors in learners’ writing. The summative scores generated by e-rater are currently officially used along with human ratings in the internet-based Test of English as a Foreign Language (TOEFL iBT), the second score for the Graduate Management Admissions Test (GMAT), and as the check scores for the Graduate Record Examination (GRE) (Shermis, Burstein, Elliot, Miel, & Foltz, 2015). Besides being put to use for these high-stakes tests, e-rater as part of the web-based writing instructional program Criterion, also aims to foster best instructional and formative assessment practices (Shermis et al., 2015).

The processes that e-rater employs to reach a summative score for essays in tests also help generate the formative feedback in Criterion. E-rater’s scoring models are constructed using a set of at least 250 randomly selected essays scored by expert readers. E-rater processes this set of essays to extract the set of feature classes, each of which represents a sub-group of multiple features. The classes of features together make up the construct model used to generate a summative score for essays and include (1) grammatical errors, (2) word usage errors, (3) errors in writing mechanics, (4) the presence of discourse elements in an essay, (5) development of discourse elements at essay level, (6) style errors, (7) content features including word usage in the relevant content domain, (8) and a feature that identifies the correct usage of prepositions and collocations as well as sentence variety (Shermis et al., 2015). Features are conceptually grouped and converted to a list of numerical feature values, which are used for predicting the scores for new essays using a regression modelling approach. Another function in e-rater has advisories which identify off-topic essays and anomalous essays. These essays will not be scored by the system until they have been properly revised.

E-rater has been developed incrementally. Later versions were built up with continuous improvements and modifications to the version e-rater 2 established in 2002. For example, the explanation in Burstein et al. (2015) shows major adaptations from the e-rater v.11.1 version (see Dean, 2013). That is, in the updated version of e-rater, non-native learner essays are taken into account with the addition of a feature that identifies the correct usage of prepositions, collocations, and sentence variety. It seems that there are constant attempts at improving the scoring models built into e-rater, as can be seen in different efforts made by Attali and Burstein (2006), Higgins, Burstein, and Attali (2006), or Ben-Simon and Bennett (2007) to train the system to more reliably
predict the human-assigned scores. Similarly, there have also been attempts among developers such as Burstein, Chodorow, and Leacock (2003), Attali (2004), Han, Chodorow, and Leacock (2006), or Chodorow, Gamon, and Tetreault (2010) at improving the error detection functions of *Criterion*.

**Research into AWE Systems**

Research into AWE systems can be subdivided into two broad areas depending on the focus of the reported results: system-centric and learner-centric. System-centric results refer to findings focusing on the characteristics of the feedback or scoring functions of AWE systems, while learner-centric results report the findings related to learners’ engagement, perception, and response to the feedback they received.

**System-centric results.** Both aspects of AWE systems, including their scoring and feedback generation functions, have been investigated in previous research. The following parts review the literature related to these two aspects, with more focus on the latter as a newer addition to these educational technological tools.

**Research into automated essay scoring.** A large volume of research has been conducted to investigate the performance of different AWE systems and their instructional applications in classroom contexts. Yang, Buckendahl, Juszkiewicz, and Bhola (2002) identified three broad areas of research in the field of automated essay scoring.

First, much of the research on automated scoring systems is devoted to demonstrating the agreement between automated scores and human ratings for the same instrument. For example, Page (1968) and Page and Peterson (1995) reported correlation between human readers and PEG at .71 in 1968 and at .86 in 1995, adding that the system predicted the scores of each pair of human judges better than human judges predicted each other’s scores. Several other studies also indicated very good human and machine scoring correlations, with Pearson $r$ ranging from .66 to .97 (e.g. Attali & Burstein, 2006; Burstein & Chodorow, 1999; Coniam, 2009; Foltz, Laham, & Landauer, 1999; Mikulas & Kern, 2006; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Nichols, 2004; Rudner, Garcia, & Welch, 2006; Weigle, 2010; Williams & Dreher, 2005). It is worth noting that most of the cited studies in this branch of research were conducted by corporate researchers working for companies which develop AWE systems. With the advantage of available access to mass data in the form of several thousand student essays fed into these automated systems, it seems corporate-funded researchers are in a favourable position to conduct studies in this area of work.
The second group of studies investigated the relationship between AWE scores and external indicators. Some of the few studies found in this group include those conducted by Powers, Burstein, Chodorow, Fowles, and Kukich (2002), Weigle (2010), and Koizumi, In’nami, Asano, and Agawa (2016). In their study, Powers et al. (2002) pointed out that both e-rater and human raters modestly correlated with each of the external indicators such as student self-evaluations and their GPA, with e-rater somewhat weaker. Such findings are corroborated by what was found in Weigle’s (2010) study, which shows moderate but consistent correlations between human and e-rater scores with non-test indicators in the form of learner self-assessment, teacher assessment, and scores of student non-test writing samples. Learner self-assessment and its correlation with Criterion holistic scoring was also investigated in the Koizumi et al. (2016) study which shows weak but positive correlation between the two. Other external factors that have been researched are multiple-choice tests and grades in courses dependent on writing (Ben-Simon & Bennett, 2007) or the TOEFL ITP, TOEFL iBT test scores (Koizumi et al., 2016). Compared to the first group of studies, this second line of research is generally not as well investigated, and findings are thus less definitive.

The third approach explores the scoring process and mental models utilized by AWE systems (e.g. Attali & Burstein, 2006; Ben-Simon & Bennette, 2007; Chodorow & Burstein, 2004; Kelly, 2005; Higgins, Burstein, & Attali, 2006; Jones, 2006; McGee, 2006; Mikulas & Kern, 2006; Nichols, 2004). For example, one concern about automated scoring is the weighting of length as a scoring trait in AWE systems. Length was found to be one of the most predictive features in the early versions of PEG (Kukich, 2000; Page, 1968) and e-rater (Chodorow & Burstein, 2004). In Jones’ (2006) study on WritePlacer Plus, ACCUPLACER’s program using IntelliMetric as its automated scoring system, 85% of the variance in essay scores was accounted for by length. See Table 2.1 for a chronology of studies investigating the reliability and validity of AWE scoring of essays.
Table 2.1.

*A Chronology of Empirical Studies on the Scoring of AWE Systems*

<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>AES system</th>
<th>Study sample</th>
<th>Methodology</th>
<th>Main Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968</td>
<td>Page</td>
<td>PEG</td>
<td>276 native speaker essays</td>
<td>PEG scoring process and training model</td>
<td>Five most predictive features: $sd$ of word length, average word lengths letters, number of commas, length of essays in words, and number of prepositions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PEG vs. human rater correlation</td>
<td>PEG vs. human raters correlation: $r = .71$.</td>
</tr>
<tr>
<td>1995</td>
<td>Page and Peterson</td>
<td>PEG</td>
<td>1014 Praxis essays as training input</td>
<td>PEG vs. human rater correlation</td>
<td>PEG vs. single human rater correlation was higher than inter-rater correlations.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>300 new essays for correlation study</td>
<td></td>
<td>PEG predicted the scores of each pair of judges better ($r = .861$) than human judges predicted each other’s scores ($r = .781$).</td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Method</td>
<td>Description</td>
<td>Agreement and Correlation</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>--------</td>
<td>-------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>Burstein and Chodorow</td>
<td>E-rater</td>
<td>510 essays written to two TWE prompts as training input. About 1100 non-native speaker essays for the validation study.</td>
<td>E-rater vs. final human reader mean agreement (exact and adjacent) was 90%; inter-rater agreement was 92%. Correlations between e-rater vs. final human scores and between two human readers were about the same, $r = .75$. E-rater’s features were generalizable from native speaker to non-native speaker essays. The differences between final human reader score and e-rater score across language groups were significant.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Method</th>
<th>Description</th>
<th>Agreement and Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>Foltz, Laham, and Landauer</td>
<td>IEA (LSA)</td>
<td>1393 essays on 13 topics in introductory content classes. 695 GMAT opinion essays. 668 GMAT argument essays.</td>
<td>LSA vs. human raters correlation $r_1 = .80$; inter-rater correlation $r_2 = .83$. LSA vs. human rater 1 correlation $r_1 = .701$; inter-rater correlation $r_2 = .707$. LSA vs. ETS raters correlation $r_1 = .86$; inter-rater correlation $r_2 = .86$. LSA vs. ETS raters correlation $r_1 = .86$; inter-rater correlation $r_2 = .87$.</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>E-rater</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------</td>
<td>---------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Powers, Burstein, Chodorow, Fowles, and Kukich</td>
<td>E-rater</td>
<td>101 to 149 essays for each of 20 GRE argument prompts and 20 issue prompts. E-rater vs. nonetest indicators (self-evaluations, GPA, etc.). E-rater’s scoring models/process. E-rater and human raters modestly correlated with each of the external indicators ($r_1 = .09-.27$ and $r_2 = .07-.38$ respectively), with e-rater somewhat weaker, reflecting similar effects from different aspects of writing proficiency.</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Powers, Burstein, Chodorow, Fowles, and Kukich</td>
<td>E-rater</td>
<td>63 “bad faith” essays written by 27 participants to trick e-rater. E-rater vs. human raters correlation. E-rater’s scoring process. Correlations between human readers were .82 and between e-rater and individual first and second rater were .42 and .37. For exact-plus-adjacent agreement, Cohen’s kappa was .85 between readers, but only .49 and .27 between e-rater and first and second readers.</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Chodorow and Burstein</td>
<td>E-rater</td>
<td>265 TOEFL essays for each of the seven prompts as training input. E-rater vs. human raters correlation. E-rater’s scoring models. Compared to length-based models, e-rater99 accounted for less score variance and had lower exact agreement, while e-rater01 accounted for more variance with higher exact agreement with human raters. E-rater99 relied heavily on counts of argument development, auxiliary verbs, and topical analysis features. E-rater01 relied largely on topical analysis features and lexical complexity measures.</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Tool</td>
<td>Description</td>
<td>Methods</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2004</td>
<td>Nichols</td>
<td>IEA</td>
<td>3244 test takers in a large state assessment program.</td>
<td>IEA vs. human raters correlation</td>
</tr>
<tr>
<td>2005</td>
<td>Kelly</td>
<td>E-rater</td>
<td>1857 GRE essays written to six prompts</td>
<td>$E$-rater vs. human raters correlation</td>
</tr>
<tr>
<td>2005</td>
<td>Williams and Dreher</td>
<td>MarkIT</td>
<td>20 law essays</td>
<td>$MarkIT$ vs. human rater correlation</td>
</tr>
<tr>
<td>2006</td>
<td>Attali and Burstein</td>
<td>E-rater v.2</td>
<td>401 sixth through twelfth grade student essays, GMAT essays, and TOEFL essays to 64 prompts</td>
<td>$E$-rater training and scoring process</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Dataset</td>
<td>Description</td>
<td>Conclusion</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>2006</td>
<td>Higgins, Burstein, and Attali</td>
<td><em>Criterion</em></td>
<td>8000 unexpected topic secondary student essays, 3138 bad-faith GMAT, GRE, and TOEFL essays</td>
<td>Topic-specific scoring models were superior in distinguishing on-topic from off-topic essays. The new generic models did sufficiently well in case topic-specific models were not available.</td>
</tr>
<tr>
<td>2006</td>
<td>Jones</td>
<td><em>WriterPlacer Plus</em> (IntelliMetric)</td>
<td>Several “doctored” student essays, 221 essays</td>
<td><em>IntelliMetric</em> was unable to make judgments about order, word choice, focus, and certain grammar errors. <em>IntelliMetric</em> could not really read essays. Length was overvalued and accounted for 85% of the variance in essay scores.</td>
</tr>
<tr>
<td>2006</td>
<td>McGee</td>
<td><em>IEA</em></td>
<td>Three student essays which were “doctored” to trick IEA</td>
<td>IEA failed to see cohesion and coherence as part of meaning construction in essays. IEA failed to measure “factual content.” IEA failed to measure mechanics properly. IEA cannot grade essays satisfactorily.</td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Tool</td>
<td>Description</td>
<td>Comparison</td>
</tr>
<tr>
<td>------</td>
<td>------------------</td>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2006</td>
<td>Mikulas and Kern</td>
<td><em>IntelliMetric</em></td>
<td>300 middle-school student essays for prompt-specific models</td>
<td><em>IntelliMetric</em> vs. human rater agreement and correlation</td>
</tr>
<tr>
<td>2006</td>
<td>Rudner, Garcia, and Welch</td>
<td><em>IntelliMetric</em></td>
<td>(1) 750 GMAT essays to each of six topics</td>
<td><em>IntelliMetric</em> vs. human rater correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2) 500 GMAT essays to each of 101 topics</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Ziegler</td>
<td><em>E-Write</em> (IntelliMetric)</td>
<td>95 essays written in the Preparation for College Writing Composition I class</td>
<td><em>E-write vs. human raters agreement</em></td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Model</td>
<td>Data Description</td>
<td>Approach/Technique</td>
</tr>
<tr>
<td>------</td>
<td>--------------------</td>
<td>-------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>2007</td>
<td>Ben-Simon and Bennett</td>
<td>E-rater</td>
<td>1255 eighth-grade students</td>
<td>E-rater scoring modeling approaches</td>
</tr>
<tr>
<td>2009</td>
<td>Coniam</td>
<td>BETSY</td>
<td>900 Hong Kong ESL scripts to three prompts</td>
<td>BETSY vs. human raters correlation</td>
</tr>
<tr>
<td>2010</td>
<td>Weigle</td>
<td>E-rater</td>
<td>386 nonnative speakers of English in the U.S.</td>
<td>E-rater vs. external nontest indicators (student self-assessment, instructor assessment, independent ratings of writing samples)</td>
</tr>
</tbody>
</table>
Research into automated feedback. The use of automated scoring systems has long been incorporated into automated writing instructional programs. Accordingly, with e-rater, we have Criterion, with IntelliMetric comes MY Access, and Pearson’s Intelligent Essay Assessor is fed into their instructional tool WriteToLearn. Once becoming part of these programs, automated essay scoring systems extend their functions from pure generation of summative scores to reporting feedback based on their assessments of student essays. Stevenson and Phakiti (2014) observed that the recent years have witnessed a steady increase in the use of AWE to provide formative feedback in writing classes. Different AWE systems generate feedback on different aspects of the writing skill, including global writing skills (e.g. Criterion and MY Access), language use (e.g. Criterion, MY Access, Grammarly, Quick Business Letters (QBL)), content (e.g. WriteToLearn) or rhetorical quality (e.g. Intelligent Academic Discourse Evaluator (IADE)). Compared with automated scoring, automated feedback has recently gained more research attention, a trend that has been boosted by the export of these automated feedback generation tools to Asian contexts characterized by large writing classes. Considering its increasingly popular applications for formative assessment in writing classrooms, automated feedback, especially automated corrective feedback, is examined more extensively in recent studies to supplement our understanding about in-class feedback provision practices alongside teacher written corrective feedback.

Automated corrective feedback. According to Chodorow, Gamon, and Tetreault (2010), the evaluation of a system for detecting and correcting errors “can be performed along at least two dimensions: system-centric evaluation, and user-centric evaluation” (p. 421). System-centric evaluation focuses on issues related to how the system performs such as the number of errors correctly identified by the system while the user-centric examination focuses on how users behave (i.e., their engagement with and response to the feedback, or subsequent textual operations conducted to address the errors) or how the system impacts the users.

There is an increasingly larger number of studies into the automated feedback function of Criterion and similar programs such as MY Access, Writetolearn, or more recently Pigai (a locally developed AWE system in China which generates feedback to student L2 English essays in Chinese) as in Zhang and Hyland’s (2018) comparison of student engagement with teacher versus with automated feedback. Research into AWE automated corrective feedback has diversified to include classroom implementation of such tools. What can be seen from the literature on feedback research so far has pointed to a suggestion that “the impact of automated feedback in AWE on improving student writing is limited, but growing” (p. 52, Heift & Hegelheimner, 2017). The following sections review previous research into automated corrective feedback by focusing on the most researched themes related to the nature of ACF.
Accuracy. Accuracy is the most well researched aspect of the automated corrective feedback generated by different AWE systems. The next sections start with reviewing the way accuracy has been operationalized, followed by relevant results regarding this aspect of automated feedback reported in different research studies.

Defining accuracy. The accuracy of automated feedback has received much attention from both AWE developers and independent researchers and classroom teachers. Yet, Quinlan, Higgins, and Wolff (2009) commented on the scoring machine in Criterion, e-rater, that “we currently have few insights into how accuracy problems may or may not disrupt students in the act of writing. We can assume that the importance of accuracy varies, depending on the specific application of the e-rater scoring engine, whether for providing feedback or predicting human holistic scores” (p. 14).

In an earlier study conducted by Burstein, Chodorow, and Leacock (2003), the accuracy of My Editor, MY Access ’error feedback tool, was evaluated using two statistical indices, precision and recall, and the human annotator served as the gold standard. More specially, precision is the number of cases in which My Editor and the human annotator agreed are errors divided by the total number of cases that the system flags as errors. Burstein, Chodorow, and Leacock (2003) further explain that precision “is equal to the number of the system’s hits divided by the total of its hits and false positives [i.e., the cases the system labels as errors but actually are not in the human judgment]” (p. 6). Recall, the other criterion to measure automated feedback accuracy, is the number of cases agreed on by My Editor and the human annotator divided by the total cases identified by the annotator. In other words, recall measures the rate of errors covered by the system compared to the total number of errors flagged by the human annotator. In their study, Burstein, Chodorow, and Leacock (2003) set a high precision threshold of 90-100% for My Editor feedback to be useful for students’ self-editing activities.

Similarly, Quinlan, Higgins, and Wolff (2009) investigated the automated feedback generated by Criterion’s scoring machine, e-rater. The authors also defined accuracy as having two dimensions, precision and recall. They further commented that as precision takes into consideration the rate of false-positives and that false positives are more obstructive in both essay scoring and Criterion scoring than false negatives (i.e., cases when Criterion missed an error which is detected by the human annotator), e-rater developers aim for high precision over recall, suggesting a precision level of 80% or more. In other words, Criterion developers prioritize precision, believing that it is more acceptable for Criterion to miss certain errors rather than incorrectly tag an error-free structure as erroneous (Burstein, Chodorow, & Leacock, 2004).

Ranalli, Link, and Chukharev-Hudilainen (2017), however, adopted a lower-bound threshold of 70% precision when they evaluated Criterion accuracy, adding that the 10% difference
accounted for additional issues in Criterion ACF not yet considered by its developers when the feedback is put to use and validated as a means of classroom-based formative assessment. By additional aspects of the feedback, Ranalli et al. (2017) referred to Criterion’s textual commentary/metalinguistic explanations and suggested revisions that provide further details about a tagged error to see if these commentaries are generic or specific (i.e., how explicit the feedback is). In a previous study conducted by Lavolette, Polio, and Kahng (2015), the error codes from Criterion were evaluated both by considering whether a real error has been detected (correct error code) and judging if the code assigned to the tagged error is correct. Unlike Ranalli et al. (2017), however, the authors did not seem to consider the metalinguistic explanations that also accompany the error codes but at least they have looked further than whether a real error has been detected to include an extra detail about whether the error code matched the nature of the error being tagged.

Research on the accuracy of ACF. Previous studies have extensively researched the accuracy of the automated feedback generated by different AWE tools. A few studies in this area have investigated both precision and recall of the feedback. For example, Burstein, Chodorow, and Leacock (2003) reported recalls of 40%, 70%, and 71% for subject-verb agreement, possessive marker, and confusable word errors when studying Criterion feedback. ESL Assistant, another AWE tool, was found to have precision and recall of 90% and 40%, respectively, while these statistics for preposition errors were 78% and 18% (Chodorow, Gamon, & Tetreault, 2010).

The majority of studies, especially classroom-based research, focus on the precision of the automated feedback. One common finding is that accuracy varies across error types (e.g. Feng, Saricaoglu, & Chukharev-Hudilainen, 2016; Lavolette et al., 2015; Ranalli, Link, & Hudilainen, 2017: Hoang & Kunnan, 2016). In addition to this, classroom-based research also pointed out differential accuracy levels for the same error types across the AWE tools. For instance, Lavolette, Polio, and Kahng (2015) showed that Criterion was very good at capitalization, missing comma, wrong word, and ill-formed verb errors with 85% plus precision. On the contrary, the system was worst at run-on sentences, wrong article, spelling errors with more than half of the time being miscoded. More recently, Criterion ACF was found to be between 71 to 77% precise when ten most common error types were considered in aggregate, with substantial variation across error types (Ranalli et al., 2017). Ranalli et al. (2017) further commented that Criterion error feedback on several error types was below 70% accurate.

It is worth noting a methodological difference in how accuracy was calculated in Ranalli et al.’s (2017) study. As earlier mentioned above, the inclusion of extra dimensions of the error feedback in this study probably justifies their decision of using a polytomous rather than a dichotomous scale in coding the error tags generated by Criterion. Their scale thus includes three
levels, 1 for not accurate, 2 for partially accurate, and 3 for completely accurate instead of the binary codes of accurate/inaccurate found in earlier studies. Also considering the metalinguistic explanations that accompany each error tag on Criterion, when investigating the accuracy of its form-focused feedback, Lavolette, Polio, and Kahng (2015) subdivided their coded categories into three types of error codes: correct, wrong, and no error (i.e., false positive) instead of reporting a summative precision statistic as done by Ranalli et al. (2017). Their findings include 75% correct, 14% wrong, and 11% no error of all the error tags they coded for their study.

Despite the abundance of research into feedback accuracy of individual AWE systems, not many studies directly compared the accuracy of the ACF generated by two or more AWE systems. Of the small number of such studies is one conducted by Chen, Chiu, and Liao (2009) who compared automated corrective feedback of two AWE programs, Criterion and MY Access. MY Access feedback on spelling, clause errors (fragments), and subject-verb agreement was only 20% precise, and the remaining seven error types had even lower precision with most of the feedback messages being false positives, unclear or not specific enough for learners. Criterion was found to be better with most of the ten studied error types being 70% accurate or more. Similarly, Chodorow, Gamon, and Tetreault (2010) examined the error feedback generated by Criterion and another system called ESL Assistant. Yet this developer-conducted study focused exclusively on the two error types of article and preposition usage for which the authors validated the error detection modules for the two error types above. The reported accuracy was comprised of both precision and recall for two systems. Criterion’s precision and recall for article errors were about 90% and 40% respectively, while the statistics for ESL Assistant’s precision and recall for this error type were 91% and 37%. In addition to this, ESL Assistant also had the error detection modules of preposition errors which performed with 78% precision and 18% recall, respectively.

Notably, most studies reported on overall accuracy of the automated feedback, with precision ranging widely from below 50% up to 90%, depending on the AWE programs. Liu and Kunnan (2016) studied WriteToLearn and found 49% precision and 18.7% recall for its corrective feedback. Similarly, Hoang and Kunnan’s (2016) study of MY Access reported a 75% precision and 39.6% recall for this tool. Among all the investigated AWE tools, Criterion was the most researched and also the one with best reported performances. Its precision is above 50% in all the studies reviewed, with most frequently found accuracy rates bunching around 75%. For example, Chodorow, Gamon, and Tetreault (2010) found Criterion corrective feedback to be 90% precise, and with that, the system missed 60% of all the errors flagged by human annotators. Dikli and Bleyle (2014) reported 72.7% precision for Criterion ACF, which is close to Lavolette et al.’s (2015) 75% precision, while it missed at least 46% of the total errors. More recently, Ranalli et al. (2017) found 71-77%
precision for a range of 10 most frequent error types detected by *Criterion* on a corpus of learner essays.

It is worth noting that not all of the error types have received equally thorough examination in the literature. While classroom based research looked at several error types at a time, with some tending to focus on the errors most frequently made by learners in their studies, earlier research conducted by AWE developers validated feedback on issues considered to be most problematic for second language learners such as articles (e.g. Chodorow, Gamon, & Tetreault, 2010; Han, Chodorow, & Leacock, 2006) and preposition (e.g. Chodorow, Gamon, & Tetreault, 2010). This group of developer-centric studies usually had a much narrower focus, where their research aimed to validate the accuracy of a certain error detection module developed for a specific error type. For example, Burstein, Chodorow, and Leacock's (2003) study reported *Criterion*'s recalls for subject-verb agreement, possessive marker, and confusable word errors being 40%, 70%, and 71% respectively as part of *Criterion* error feedback model development. On the other hand, after developing the error detection module for article usage in *Criterion* and also for preposition usage in *ESL Assistant*, Chodorow, Gamon, and Tetreault (2010) went on to validate its usefulness for native and non-native speakers of English and found specific precision and recall results for these systems as mentioned above.

The range of errors addressed by AWE systems. Although being less extensively researched, the coverage of AWE corrective feedback (i.e., the range of error types being detected by AWE automated corrective feedback) has received some attention in classroom-based research where the use of these systems reveals learners’ actual needs most clearly. A few common findings have been reported about the insufficient linguistic areas covered by AWE corrective feedback in previous studies. Most frequently mentioned is a criticism that AWE tools lack feedback on salient errors made by ESL/EFL populations, one of which being the lack of feedback on verb tenses (Chen, Chiu, & Liao, 2009; Hoang & Kunnan, 2016; Lavolette, Polio, & Kahng, 2015). Lavolette, Polio, and Kahng further observed that *Criterion* seemed to miscode verb tense errors as subject-verb agreement errors. Furthermore, a few studies pointed out the lack of feedback on other error types commonly found among ESL/EFL students, including articles for *WriteToLearn* (Liu & Kunnan, 2016) and for *Pigai* (Zhang & Hyland, 2018), word choice for *WriteToLearn* (Liu & Kunnan, 2016) and for *Criterion* (Chen, Chiu, & Liao, 2009), and preposition for *WriteToLearn* (Liu & Kunnan, 2016) and for *Pigai* (Zhang & Hyland, 2018). Chen, Chiu, and Liao (2009) also indicated that *Criterion* either missed or insufficiently dealt with errors related to word order, modal auxiliaries, collocations, conjunct errors, and pronouns.
In studies which compared teacher and automated corrective feedback, it is reported that the writing instructor provided more feedback on form than AWE systems. For example, Dikli and Bleyle (2014) compared Criterion feedback in mechanics, grammar, usage with teacher form-focused feedback. In this study, Criterion was found to miss several errors which were detected by the instructor. Particularly, some of the unidentified errors were typical of ESL students such as singular/plural and verb tense. Similarly, in examining the feedback generated by Pigai, Zhang and Hyland (2018) reported that the system lacked feedback on big issues such as article and preposition errors and that teacher feedback was able to address more error categories than the automated system did. Overall, what can be seen from the reviewed studies in regard to their coverage of error types is that there is still room for the addition of feedback on certain salient linguistic features that the ESL/EFL student populations are struggling with as well as adjustments to the range of error types covered by these AWE systems to better meet the needs of their targeted users.

Beside the two aspects of accuracy and feedback coverage presented above, the investigation of automated corrective feedback generated by different AWE systems has been extended to include issues related to the metalinguistic explanations that accompany error tags and how explicit these explanations are. Regarding the metalinguistic explanations of automated corrective feedback, previous research has qualitatively examined these error messages and what seems to be consistently reported is that the feedback is long and generic, and that sometimes not usable (Dikli, 2006, 2010). In addition, in comparison with the shorter and more focused teacher feedback, AWE feedback still lacks human interaction and thus, it has less personalized positive reinforcement (Dikli, 2010). A more recent study conducted by Ranalli (2018) investigated the explicitness of the automated corrective feedback generated by Criterion and how it impacts learners’ revision outcomes. By operationalizing explicitness as either “generic” or “specific” to indicate two levels of explicitness found in Criterion metalinguistic explanations, Ranalli (2018) concluded that feedback explicitness is a determining factor in learners’ success at revising their errors.

Although explicitness has long been a central issue in written corrective feedback as demonstrated in a large number of studies on the efficacy of different types of teacher written feedback (i.e., direct and indirect feedback) as discussed previously in this chapter, there is still a scarcity of research into this aspect of automated corrective feedback, which can be partially due to the currently formulaic and limited capacities to generate feedback from AWE systems. In other words, with limited flexibility in the current feedback mechanisms of AWE systems, as reflected through earlier findings that pointed out the genericity in metalinguistic explanations and lack of graduated personalized explanations (Dikli, 2006, 2010, 2014), researchers may have been discouraged from conducting more in-depth research into this aspect of AWE feedback. However,
as the field of automated feedback is making progress, it is worth pursuing this strand of research to deepen our understanding of students’ response to a new mode of feedback on their writing.

Overall, system-centric research into the automated scoring and feedback functions of AWE systems has investigated different aspects of the feedback, in both high-stakes and classroom-based assessment contexts. While being a central issue in earlier research and dominated by corporate-funded researchers, automated scoring has recently received less attention. Instead, research on automated feedback functions of AWE systems has gained in popularity, especially for formative assessment purposes in classroom contexts. The relevance of this strand of research has been established by an increasing number of studies into the characteristics of such feedback. Despite the fact that research tools and methods have become better defined for the investigation of automated scoring and feedback, a more systematic and consistent approach is needed to make reported findings across different studies comparable and thus, better justified.

**User-centric evaluation.** Turning to user-centric research on AWE corrective feedback, previous studies have investigated the learners’ behaviour as well as the impact of such feedback on learners’ learning to write. Stevenson and Phakiti (2014) conducted a comprehensive meta-analysis of studies on the impact of automated feedback on the quality of student writing. Their analysis reinforces Lai’s (2010) three-perspective approach to evaluating automated feedback on students’ learning to write, including: (i) the product perspective (the effects of AWE on written production); (ii) the process perspective (the effects of AWE on revision, teaching and learning processes); and (iii) the perception perspective (learners and teachers’ perception of the use of AWE). Although the number of studies that investigate AWE form-focused feedback has been rather limited up to this point, these studies have employed a variety of research methods and tools to shed light on the impact of AWE on teaching and learning practices. The following sections will critically review the studies in this strand of research based on Lai’s (2010) and Stevenson and Phakiti’s (2014) three-perspective division.

**The product perspective.** Research into the effectiveness of automated feedback has focused mainly on the product perspective to see whether such feedback contributes to writing quality as evidenced in either students’ revised drafts within the same essays or on new compositions. Such research has produced mixed results. The most well researched AWE program is ETS Criterion and its automated corrective feedback. For example, Attali (2004) studied more than 33,000 student essays to investigate the usefulness of Criterion’s error and discourse feedback. The findings from this study showed that the feedback led to a significant error rate decrease for 23 out of 30 error types, and a quarter of the errors were successfully corrected. Criterion also proved to be useful for
its feedback on background, main point, supporting ideas, and concluding elements in student essays.

Other studies supporting Attali’s findings include Chodorow, Gamon, and Tetrault’s (2010) investigation of Criterion article error rate among non-native writers, Choi’s (2010) investigation of both holistic quality and error correction rate among students using Criterion, and Kellogg, Whiteford, and Quinlan’s (2010) comparison of three feedback conditions using Criterion ACF. In Attali et al.’s (2010) study, Criterion’s error detection module could help non-native writers reduce article error rate with the program’s performance levels of 90% precision (i.e., the proportion of cases when the system correctly identifies an ill-formed structure as an error divided by the total cases it flags as errors) and 40% recall (i.e., the percentage of cases when Criterion correctly identifies the errors divided by the total human annotator flagged cases). In the second study, Choi (2010) set up a quasi-experiment of the use of Criterion in writing classes including three conditions. The first group had Criterion fully incorporated in their writing class, the second group could use Criterion but this was considered an optional tool rather than an obligatory part of their writing course, and the third group (comparison group) had no access to Criterion. Choi’s main finding based on the pre, post, and delayed post-test results was that in the writing classes with full incorporation of AWE, students did better with greater gains in both fluency and accuracy measures as compared to the non-AWE and AWE-optional conditions. Similarly, Kellogg, Whiteford, and Quinlan’s (2010) comparison of three feedback conditions (no feedback, intermittent feedback, and continuous feedback from Criterion) among 59 ESL students showed that although holistic scores were not significantly improved, students receiving continuous feedback experienced a decrease in errors of grammar, mechanics, usage, and style.

A more recent study by Li, Link, and Hegelheimer (2015) also reported students’ improvements in linguistic accuracy measured by the comparison of normed error counts (number of errors per 100 words) between the first and final drafts of the same paper using Criterion feedback for revisions. Another study that also examined the impact of Criterion automated corrective feedback on learners’ writing accuracy was conducted by Li, Feng, and Saricaoglu (2017) who found that the use of this feedback resulted in positive short-term effect for eight out of nine examined error types, with short-term effects being measured by the rates of errors from the first to the final drafts of the same papers.

Although most reviewed research suggests a positive impact of Criterion feedback on the learners’ linguistic accuracy and overall writing performance, it is worth noting that with the exception of Choi’s use of a human scorer along with the automated system in measuring the writing quality of student essays in the pre, post, and delayed post-tests, a common weakness of other studies is that the evaluation of the students’ error rates as well as holistic scores was done by
Criterion itself. Therefore, their reported findings are based on the questionable assumption that the automated system’s error feedback and scoring are valid measures of learners’ performance.

Other programs being studied for their contribution to the quality of students’ written products include MY Access (e.g. Lai, 2010), ESL Assistant (e.g. Chodorow, Gamon, & Tetrault, 2010), Intelligent Essay Assessor (e.g. Foltz, Laham, & Landauer, 1999), QBL (e.g. Warden, 2000), and Writing Roadmap (e.g. Wang & Wang, 2012). All of these studies report that the researched AWE programs performed well when helping students improve their writing quality. One exception is Lai’s (2010) finding which shows that students adopted peer feedback more often than AWE feedback, leading to greater improvement in their essays. However, the design of these studies can pose a threat to the validity of their findings. For example, Lai’s study did not use objective measures of the writing quality but the findings were merely based on students’ self-evaluation of their use of the feedback from their peers and MY Access through responding to a questionnaire. Similarly, Wang and Wang’s (2012) study, which compared Writing Roadmap feedback to teacher feedback, only had a whole-class feedback session giving general comments on student writing. Students received a few form-focused feedback points on their essays but feedback on content and other discourse level issues were neglected, not surprising considering the number of large writing classes the teacher had to teach each semester. Due to this lack of control of the amount and quality of feedback students received either from the writing instructor or the AWE system, any interpretations or generalizations of the results should be done with some caution.

Retention as changed accuracy. Some studies have been conducted to investigate how learners’ accuracy changes following the incorporation of Criterion automated corrective feedback. The studies varied in how they operationalized accuracy change. For example, in a recent study by Li, Feng, and Saricaoglu (2017), change was calculated by comparing error reduction rates from first drafts of the first paper to first drafts of subsequent papers. Their study showed that students improved in terms of one error type, run-on sentences, but not in the remaining eight error types where the participants experienced improved accuracy from first to revised drafts of the same essay. Using Criterion error reports to make conclusions about students’ accuracy changes was also what Attali (2004) and Kellog, Whiteford, and Quinland (2010) did in their studies. Attali (2004) found that the rates of errors in learners’ grammar, usage, mechanics, and style decreased by .15 to .27 of the respective standard deviations. Reduction in error rates in these aspects of writing in a delayed test was also reported by Kellog, Whiteford, and Quinlan (2010). However, what differentiates this study from Attali (2004) is that it compared three groups in three different feedback conditions: no feedback, intermittent feedback (i.e., students only received automated feedback in week four of the semester), and continuous feedback (i.e., students received automated feedback on all of the three
practice essays in week 3, 4, and 5). Findings from this study pointed out that continuous feedback condition was better with reliable reduction in errors of grammar, mechanics, usage, and style on the test essay. Also comparing different treatment groups, yet using a pre-test post-test design, Choi’s (2010) study compared three treatment groups (no-Criterion, optional Criterion access, and integrated-Criterion) spanning a period of eight weeks. Results from this study indicated that treatment conditions significantly affected writing quality among students, with the AWE-integrated group having better fluency and accuracy measures.

Overall, these studies suggest improvement at differential levels in learners’ accuracy over time thanks to the use of automated corrective feedback from Criterion. Results from these studies, however, should be interpreted with some caution because of the way accuracy change was calculated. That is, these studies disregarded the question about how accurate the automated corrective feedback was when they chose to use Criterion error reports and its automated scores as standards to judge learners’ accuracy changes over time. This methodological choice poses a potential risk to their validity as the automated feedback itself is still open to flaws and needs further validation. By using it as the gold standard for measuring the changes observed in learners, the researchers based their calculations on a questionable assumption that Criterion corrective feedback provided accurate measures of L2 learners’ levels. In addition, for feedback to have impact on the students’ accuracy development, longitudinal studies are desirable when students have more frequent access to the automated feedback over an extended period of time. Except for Choi’s (2010) research which spanned eight weeks, the reviewed studies tended to make non-systematic observations of learners’ accuracy changes using first and revised drafts of the practice essays to investigate the participants’ L2 developmental trajectories.

**The process perspective.** Students’ use of automated feedback and its contribution to developing student independent editing skills is another focus of research into AWE. In this part of the literature review, there are two subsections that focus on two different aspects of students’ interaction with the feedback they received: (i) their cognitive engagement with the feedback and (ii) behavioural engagement as evident in their revision processes using AWE feedback.

Regarding cognitive engagement, a comprehensive search of the literature highlighted the paucity of research into this dimension of student engagement with the feedback from AWE systems. The search found only two such studies. Firstly, a study recently conducted by Zhang and Hyland (2018) adopted a case-study approach to investigating the influence of individual learners’ factors on their engagement and use of Pigai automated feedback and how that compared to teacher feedback. The second study conducted by Zhang (2017) examines an individual learner’s
engagement with the feedback generated by an AWE system developed by a Chinese technology company, although the names of the company and the AWE system were not mentioned. Both studies approached the construct of engagement using the triad of cognitive, behavioural, and affective (or emotional as referred to in the latter) dimensions. While Zhang and Hyland (2018) provided an in-depth analysis of students’ engagement with Pigai feedback in relation to a range of individual and contextual factors, the roles of learner proficiency and repertoire of learning strategies were highlighted as determining elements. Zhang’s (2017) study was less convincing when providing idiosyncratic results about a single case using a less systematic approach to data analyses. However, its contribution was undeniable in the sense that this study was the first to draw the link between the established construct of engagement with the use of AWE feedback in an EFL classroom context.

It comes as no surprise that the nature of student cognitive engagement with AWE feedback has been under-researched, as even with the more established field of teacher feedback, there has been much less research into learners’ processing and engagement with the feedback compared to their revision operations and practices (or behavioural engagement as in Ellis’ (2010) framework). According to Storch and Wigglesworth (2010), this is probably due to the difficulty of accessing learners’ internal cognitive processes.

On the contrary, there has been more ample research into uptake behavior (i.e., editing and revising) using the automated feedback. Some previous research showed middling uptake rate of AWE feedback. For example, in Dikli’s (2006) study of MY Access, despite the fact that MY Editor provided twice as many usable feedback points compared to the teacher, the twelve participating students used only half of them for their revisions. Similarly, Chapelle, Cotos, and Lee’s (2015) examination of Criterion ACF shows that nearly half of all the feedback did not lead to any revised forms whether the feedback was direct or indirect. On the other hand, in Choi’s (2010) investigation, with Criterion full integration in the writing courses in both ESL and EFL contexts, students revised their drafts more often than other classes which had either no or optional AWE integration. In a similar vein, Ebyary and Windeatt (2010) found that Criterion feedback encouraged the EFL learners to write and revise more, although the planning function on Criterion (with its pre-writing tools) was hardly used at all. Looking closely into revision practices using MY Access feedback, Fang (2010) found that the system’s feedback on form was more useful than its feedback on content for essay revisions.

Besides looking solely at learners’ revisions using the automated feedback, a few studies have combined user-centric and system-centric evaluations by examining the impact of feedback quality and error type on student revisions. For example, Chodorow, Gamon, and Tetrault’s (2010) study on ESL Assistant showed that the users were able to distinguish correct from incorrect suggestions
from this program, leading to well informed choices for their revisions. In addition, Chapelle et al. (2015) suggested that there is a relationship between student revision success rate and the accuracy of the feedback. Error type is another factor that was examined in Lavolette, Polio, and Lee’s (2015) study. Students’ revisions were reported to vary depending on specific error types. Specifically, low response rates (50%) were recorded for missing comma, preposition, and spelling, while much higher rates (about 85%) were recorded for the grammatical error types of ill-formed verbs, proofread, and subject-verb agreement.

Another aspect that emerges from the review of the literature is that the use of AWE feedback positively helps students develop certain self-regulatory revision strategies. For example, with the implementation of IADE in a graduate academic writing course in the US, IADE feedback was found to make the 105 international students in this study notice and focus more on the discourse form (i.e., the rhetorical moves in the introduction to research articles). Wang and Wang (2012) also reported that the students receiving Writing Roadmap in their study could exercise self-correction and revision effectively by reading Writing Roadmap formative, diagnostic and summative feedback. Li et al. (2015), on the other hand, investigated the amount of students’ writing practice (i.e. how often they resubmitted their drafts to receive feedback) when they had access to Criterion. Although the findings based on students’ number of submissions for each paper suggest that Criterion feedback encouraged students’ engagement in writing practices, interview results with teachers and students revealed that other variables such as error types or instructional methods may have played a role in students’ enhanced engagement in writing practice.

There have also been negative results about the impact of AWE feedback on students’ learning to write. For instance, negative evidence was found in Warden’s (2000) study concerning students’ revision activity when they were provided with computer-generated feedback from QBL. The study compared three different treatment groups: Group A received computer-generated feedback but no redraft was required, the treatment for Group B was similar to A but included submission of a redraft, while Group C received teacher-based non-sentence level feedback with redrafting. Based on keystroke analyses of students’ performance during their first draft and redrafting on new compositions, students in treatment group B (who were required to redraft after receiving specific automated error feedback from QBL before having their grade assigned) had lower levels of editing time. That is, the chance to redraft seemed to negatively impact how these students approached their first drafts as they demonstrated a lack of direction and waited instead to receive the feedback to take actions. Their redrafting (i.e., revision) time was taken up by removing the errors pointed out to them. All of the three measures of editing for the first draft (editing time, times printed, and QBL start) revealed that Group B performed little revision. Such a finding seems
to support Truscott’s (1996) argument against the provision of corrective feedback which he suggested as both time-consuming and detrimental to students’ interlanguage development.

Studies into the impact of AWE on students’ engagement and revision processes using AWE feedback have utilized various methods. Data collection tools range from students’ self-reports in questionnaires (e.g. Lai, 2009) to text analysis of writing specimen from first to second drafts, and retention rate at a delayed test (e.g. Choi, 2010; Dikli, 2006; Wang & Wang, 2012). Other methods used include Camtasia screen recordings (e.g. Cotos, 2011), keystroke analysis (e.g. Warden, 2000) and think-aloud protocols (e.g. Cotos, 2011). While heterogeneous approaches to unveiling students’ engagement and processing of the feedback can inform future research of relevant methods and tools, they show a lack of consistency in approaching issues in this field. This also makes interpretation of research results about learners’ cognitive and behavioural engagement with automated corrective feedback problematic and inconclusive.

The perception perspective. A number of studies have investigated students’ perception of the use of AWE systems in writing classes. Most of these studies did not single out the automated corrective feedback for examination. Instead, they predominantly use students’ self-reports in questionnaires and interviews to investigate users’ general perception of the AWE systems and have produced mixed results. For example, in Chen and Cheng’s (2006) study, a moderate percentage of 68 third-year EFL English majors in Taiwan (55%) found the use of MY Access either slightly or moderately helpful. These students, however, were less happy with the writing and revision functions of the program. In another study in Taiwan, the majority of students were positive about MY Access as a writing tool and expressed satisfaction with MY Access use for revising forms, but not for revising content. Ebyary and Windeatt (2010), however, found that the Egyptian trainee EFL teachers in their study expressed generally improved perceptions about feedback thanks to the incorporation of Criterion automated feedback in the writing course. Such results were echoed in Li et al. (2015) as well as Li, Feng, and Saricaoglu’s (2017) studies of Criterion where intermediate to advanced ESL learners in the US found AWE corrective feedback to be a helpful resource for their language related issues. However, students in the former study did not rate Criterion AWE feedback on content, organization, and rhetorical issues highly, while those in the latter had a few issues understanding and following the feedback provided to revise their essays. In a similar vein, Cotos (2011) found that 105 master and PhD international students enrolled in a graduate academic writing course in the US were motivated to revise their drafts using IADE. Additionally, Wang (2015) reported average to high satisfaction levels with Criterion’s instant scoring function, error analysis of usage, feedback for organization and development among 53 English majors at a Taiwanese university. However, students in this study were much less happy with Criterion’s
scoring rubric, its style error analysis and the planning tool on the system. A recent study conducted by Huang and Renandya (2020) to investigate the automated feedback generated by Pigai also produced very positive response from low-proficiency learners. One of the key findings is that none of the surveyed participants expressed doubts about the accuracy of Pigai automated feedback.

An exception is Li, Link, and Hegelheimer (2015) who focused exclusively on the automated corrective feedback generated by Criterion and found some perceived benefits of using this feedback in learning to write. Among these are increased grammatical consciousness among learners, the learning of metalinguistic terms and clearer awareness about the distinction between content versus form feedback. In addition, students in this study commented that the automated corrective feedback encouraged them to use external tools for error correction besides using Criterion ACF itself as a resource for fixing the errors in their essays. On the other hand, students expressed negative comments about the inaccurate feedback for some errors and the fact that Criterion failed to detect all the errors made. Also, they questioned the value of Criterion ACF for longer term error reduction as they were not sure if the feedback would help them reduce errors over time.

In studies where AWE feedback was used along with other forms of feedback, AWE has been rated of lower value than peer feedback (e.g. Lai, 2009) and instructor’s feedback (e.g. Choi, 2010; Dikli, 2010; Dikli & Bleyle, 2014). For example, in her study of Criterion, Dikli (2010) noted that the automated feedback generated by MY Access and teacher feedback were different in nature. While the former was long and generic, the latter was more concise and cumulative. The author especially noted the lack of human interaction and thus the system’s failure to personalize the positive reinforcement in its feedback. Similarly, both the ESL and EFL students in Choi’s (2010) study rated the feedback provided by the teacher clearer as well as more encouraging than Criterion automated feedback. The students having access to Criterion in his study also held negative opinions about the amount of feedback they received from the system. Four years later, Dikli and Bleyle (2014) investigated the feedback from Criterion and despite the fact that the ESL students in this study held positive perceptions of the usefulness of the automated feedback, they still thought more highly of the teacher feedback. Generally, learners’ less favourable perception of AWE feedback compared to teacher feedback tends to come from two sources: (i) AWE systems’ failure to accurately detect certain error categories to a satisfactory level; and (ii) their provision of generic, long and redundant feedback as opposed to the more “humanistic” teacher feedback which is personalized, specific, consistent, and thus more useful.

**Learner factors.** Despite being an extensively researched area and the most thoroughly studied psychological aspect in second language acquisition (Dornyei, 2005), individual learner
differences have received little attention in research into automated corrective feedback. The review of the literature showed scant evidence about how individual differences impact learners’ engagement and use of automated feedback. Apart from a small number of recent studies (e.g. Zhang, 2017; Ranalli, Link, & Chukharev-Hudinailen, 2017; Zhang & Hyland, 2018) which touched on this issue, the discussion about learner factors in automated feedback still remain hypothetical surrounding the role of students’ proficiency level and its influence on how much they could make sense of the feedback and the long-term impact on linguistic acquisition. Zhang and Hyland (2018) adopted a multiple case studies approach to compare two Chinese EFL learners’ engagement with teacher and automated corrective feedback. Besides suggesting the central role of the learner’s proficiency, the authors discussed different learner goals and learning strategies as crucial factors as explaining each individual’s engagement with the two types of feedback.

Adding to the evidence about the role of learner proficiency, Ranalli, Link, and Chukharev-Hudinailen (2017) studied two different level classes comprising of a total of 82 ESL learners. Ranalli et al. (2017) found that students in the lower level course were able to use Criterion automated feedback in correcting errors better than their higher course counterparts. Unlike Ranalli et al. (2017), Huang and Renandya (2020) studied the impact of Pigai automated feedback on 67 lower-proficiency students whose CET-4 scores were below 425 (approximately 50 on the TOEFL iBT). These students were from two writing classes, one serving as the experimental group who received Pigai automated feedback while the other being the control group. Huang and Renandya (2020) did not find any significant difference in the quality of revisions between the two groups. Although Ranalli et al. (2017) accounted for the two classes’ differential capacities to make use of the AWE feedback for revisions by referring to the different pedagogical foci of the two courses, while Huang and Renandya (2020) attributed the lack of the groups’ differences to the short duration of study time, proficiency levels deserve deeper investigation at individual level rather than group level to show how it determines not only revisions but also long-term benefits each learner can have out of using the feedback.

Overall, the review of the literature accentuates the need to examine automated feedback from multiple perspectives to provide deeper and more systematic insights into how different factors interact to impact students’ learning to write using a new writing platform and a new channel of feedback provision. Through the several studies that have been mentioned, a few factors have emerged as worthy of more examination. From a system-centric perspective, this includes error types and feedback types while from a user-centric perspective, some exemplary factors are individual learner factors of proficiency, learning goals and strategies. Putting a number of different factors into a coherent research study promises to provide us with better understanding about this
automated tool of feedback provision as a starting point in our pursuit of maximizing learning outcomes for learners.

**Limitations of Previous Research**

The review of the literature has showed certain limitations of previous research on automated feedback and automated corrective feedback in particular. Firstly, most studies are either one-shot examination of the impact of ACF on students’ revisions or short-term implementation of these tools. As Koizumi et al. (2016) commented, previous studies adopted only two time points to collect information related to students’ accuracy change over time, which failed to sufficiently capture long-term changes among learners. Secondly, with the exception of Choi’s (2010) research, all reviewed studies based their judgement of changes in writing quality on holistic scores and of accuracy changes on the number of error tags generated by AWE systems themselves. This poses a threat to the validity of the studies due to the fact that the feedback generation modules and scoring engines in AWE systems are not free from errors, and therefore, the golden standards to determine changes in writing quality or accuracy gains should still be well-trained human annotators and raters (Lavolette et al., 2015).

In addition, although accuracy changes and measurements of overall writing quality judged by holistic scores assigned to student essays have formed the bulk of research into automated feedback, there is a paucity of insights into student engagement with the feedback and how that impacts decisions on revisions as well as long-term writing developments. In relation to engagement as a research issue, certain potentially explanatory factors have not been sufficiently investigated. Although there have been efforts at studying student engagement with, response to, and use of the automated feedback whilst taking into consideration the influence of learner proficiency levels or feedback explicitness (e.g., Ranalli et al., 2017; Ranalli, 2018), the feedback in these studies were provided in a simulated condition where the errors and the accompanying error tags had been separated and presented in isolation, one by one. This isolated response format, although more convenient for research purposes, may fail to capture students’ actual response to automated feedback as they would have done if they had worked on their drafts and revised essays in their entirety on AWE systems. Therefore, more naturalistic research design would be a better choice to capture the realities regarding students’ use and response to the feedback.

Overall, Warschauer and Ware’s (2006) comment seems to still hold true, that the existing AWE research seems to be too outcome-oriented while not enough attention is paid to the learning and teaching processes involved in the implementation of AWE systems and feedback in the classroom (as cited in Stevenson, 2016).
Where Should We Go from Here?

In seeking explanations for some L2 learners’ failures to benefit from WCF, Bitchener (2017) noted that studies on corrective feedback mainly focused on examining improved written accuracy in revised texts and new pieces of writing, while the actual cognitive processing that is required to produce modified output received little attention. Therefore, we are still left in the dark regarding the factors that may explain why learners have not been successful in producing modified output. In Bitchener’s (2017) words, “little attempt has been made to explain why some learners fail to either engage with the feedback and/or successfully process it across the various cognitive stages that have been hypothesized as essential for text modification and the ongoing production of accuracy in new pieces of writing” (p. 129).

More directly addressing issues in the field of automated feedback, Zhang and Hyland (2018) commented that most current AWE feedback research has focused on students’ written products while it neglected students’ writing and revision processes. Despite the importance of this new feedback channel in the second language writing classroom, Zhang (2017) expressed his concern that AWE feedback has not been thoroughly studied in previous research. Storch (2018) added to this point by outlining future directions in this field: “Clearly what we need is well-designed studies […] which examine the impact of these tools on processes: on the nature of the feedback provided, on how learners engage with that feedback, and ultimately on L2 learning outcomes” (p. 270). Storch’s (2018) comment echoed the concern earlier mentioned by Qi and Lapkin (2001) and Storch and Wigglesworth (2010) that there is an urgent need for studies on how individual learners are engaged with the automated feedback during their revision processes and for more classroom based research that reflects teaching/learning realities instead of experimental settings where AWE feedback is implemented.

Gaps to Be Filled…

The review of previous research informs and shapes part of the aims and design of the current study. A common recommendation made in previous studies is that there should be greater emphasis on the process perspective in research on automated feedback (Kim et al., 2019; Stevenson, 2016; Zhang & Hyland, 2018). More specifically, Kim et al. (2019) suggest further examination of students’ cognitive processing of the AWE feedback which could provide insight into the factors that make feedback effective in both the short and long terms. In addition, Stevenson (2016) called for research that goes beyond pedagogical implications of process-oriented research. Instead, he was concerned about a theoretical framework that considers both the cognitive and social aspects of the revision processes and the impact of AWE systems on such processes. The implementation of such a theoretical framework helps to situate AWE feedback in the larger picture
of written feedback and writing processes in general. Zhang and Hyland (2018) also stressed the central role of learners’ engagement through the examination of their cognitive processing of the feedback in second language writing research. Specifically, they recommended the investigation of both the feedback itself and individual learner factors and how both of these shape the revision processes.

Bringing Stevenson’s (2016) suggestion to reality and also addressing the major concern raised by Zhang and Hyland (2018) regarding the under-researched aspect of student engagement with automated feedback, Ranalli et al. (2017) used an argument-based approach to validate the applications of Criterion for formative assessments. In order to do this, the authors used the evaluation and utilization inferences in Chapelle, Cotos, and Lee’s (2015) argument-based framework. To enhance understanding about student engagement processes with the feedback, Ranalli (2018) zoomed in on the explicitness, course level, and the accuracy of the corrective feedback provided by Criterion as potential factors that mediate students’ revisions (behavioral engagement) and perceptions (affective engagement). Beside contributing some original perspectives on the automated corrective feedback, these studies have also highlighted the need for research on underexplored aspects including the reasons why learners disregard the feedback.

All of the above recommendations are well justified, as without in-depth research on different aspects of student engagement with the feedback, it is almost impossible to either reveal the underlying cognitive processes that determine the revision and learning outcomes addressed in the product perspective or to account for student perceptions in self-report tools such as questionnaires and interviews. Guided by previous research, the current study aims to answer the five questions below using a main study and an embedded multiple-case study to address the issues identified in the review of the literature.

Research Questions
1. What is the nature of Criterion automated corrective feedback?
2. How accurate is Criterion’s automated corrective feedback?
3. How do EFL students engage with and respond to Criterion ACF?
4. How does EFL students’ accuracy in L2 writing change after a semester’s access to and use of Criterion ACF?
5. What are students’ perceptions of Criterion ACF?

In the next chapter, the methodology for the study is presented in detail to seek answers to these five research questions.
Chapter 3
METHODOLOGY

Brief Overview of the Study

The current study uses a mixed-methods approach to investigate the impact of Criterion’s automated corrective feedback on EFL students’ writing accuracy over a semester as well as their engagement with this input as they applied it to essay revisions. Criterion was implemented in two intact writing classes comprising 72 tertiary EFL learners in central Vietnam. Quantitatively, accuracy of student essays in the pre-test, post-test, and delayed post-test was calculated to examine their accuracy gains (or the lack thereof) as evidence of uptake and retention of the corrective feedback. Qualitatively, the think-aloud protocols of 14 students were recorded during the time they revised their essays on Criterion. Students’ think-aloud protocols were triangulated with the analysis of their revisions by looking at the first and revised drafts submitted to Criterion during practice sessions on the software. These were followed by four stimulated recall interviews among four TAP students, and four focus group interviews among other students towards the end of the semester. Three participants were also selected for inclusion in the embedded multiple-case study for a more in-depth picture regarding students’ perception, engagement, and processing of automated corrective feedback. These learners’ individual proficiency levels and learning goals were factored in as a way to see how individual factors played a role in the types of revisions made and the learning outcomes produced. An overview of the structure of this study is found in Figure 3.1.
Engagement with automated feedback

Revision operations

Correct revision

Effective response to feedback? (Uptake)

Incorrect revision

Pre-test

Gained accuracy over time (Retention)

Post/Delayed post tests

Figure 3.1. Validation of the efficacy of automated corrective feedback for EFL learners
Context of the Study

The study took place in a university in the city of Hue, central Vietnam. Students who participated in this study were English majors in the English Department which is the largest department in the university. Students learned the four English skills of listening, reading, speaking, and writing along with other subjects related to their major such as theories and practice in translation, pedagogical skills, or semantics and pragmatics.

Technological applications in teaching and learning in the English Department were quite popular but limited to simple uses. The university did not subscribe to any online learning management systems such as Blackboard or Moodle and most online communications between teachers, students, and other staff took place through emails and some social networking sites like Facebook or teacher-created websites. Furthermore, simple applications of technology took place in classes of reading, listening, speaking, but were rarely seen in writing courses.

For each subject, students met once a week for a 100-minute session. Writing lessons were mainly delivered in the traditional mode with the writing instructors presenting knowledge and skills. In Writing 4, the course taken by participants for the current study, students learned English academic writing skills with a focus on three types of essays: problem-solving, opinion, advantage/disadvantage essays. After knowledge about the specific genre of writing and the relevant writing skills were presented by the instructor, students were assigned practice activities when they hand-wrote essays on paper for submission to their instructors for feedback. This practice limited writing instruction to the 100 minutes’ meeting in class each week, while teacher feedback was usually delayed a week or more when students had probably forgotten what they had written on certain topics and what linguistic difficulties (‘holes’) they encountered during the composing process. Student learning was assessed on a regular basis but was still heavily dependent on the final assessment score in the end-of-term examination (60 percent of the total score). The other 40 percent went to attendance and participation in class activities, presentations, and one or two mid-term tests.

Against this backdrop, the current study introduced ETS automated writing evaluation program, Criterion, in a quasi-experiment to investigate how its automated corrective feedback could potentially benefit EFL learners in their learning to write in English.

Criterion Feedback

Although Criterion generates both surface-level and content-level feedback, the only focus of the current study is on automated surface-level feedback. This research choice was based on the findings of previous research commenting on the lack of usefulness of content, organization, and essay development feedback offered by the system (see the review in the previous chapter). Quinlan,
Higgins, and Wolff (2009) commented that these areas of feedback from Criterion need further development. I also agree with Chen, Chiu and Liao’s (2009) assumption that “most of the current AWE systems can only provide feedback at the sentence-structure level” (p. 6). This is not to overlook more recent attempts at feeding new functions into the systems like Criterion or Writetolearn to improve their performance on discourse and content features of essays. While program developers have made claims on the systems’ capabilities to deal with higher-order writing skills such as content and rhetorical feedback, there is still not enough convincing evidence regarding its usefulness for learners’ writing skills.

Adapting another way of looking at teacher written feedback by Peterson, Childs, and Kennedy (2004) to automated feedback, feedback can be classified as editive or revision. Criterion feedback on grammar, mechanics, and usage fall into the first level as it encourages students to make local changes to their text including mechanical, grammatical, simple and straightforward lexical and syntactic changes. These changes, as is observed from the use of Criterion, are always at the sentential and phrasal levels, while cross-sentential and discoursal revisions are not generated by the system. Other than the feedback on grammar, usage, and mechanics, Criterion feedback on higher-order traits of writing skills of organization, development, and style is generic and repetitive from one essay to another. For example, for organization and development feedback, the first sentences are always highlighted with the following feedback: “Is this part of the essay your thesis? The purpose of the thesis is to present the main idea of your essay clearly and concisely. Check the Writer's Handbook for ways to improve your thesis”, while the first sentence in each paragraph from the second to the second-to-last paragraphs always get the following feedback message:

You have used three or more main ideas to support your thesis statement. Does each of your main ideas help the reader better understand the overall point of your essay? Each main idea should be introduced by a topic sentence and should have its own paragraph. Use examples, stories, and details to support and explain your main ideas. Check the Writer's Handbook for ways to develop main ideas.

Similarly, the final paragraph of all essays generates the same message asking students if they have rephrased their thesis in a different way or brought their ideas together.

For organization structure feedback, transitional words and phrases used in all the essays are flagged to signify to students what transitional devices they have used, and to ask them to self-check by referring to the Writer’s Handbook uploaded on Criterion.

Style feedback messages include those giving advice on the choice of passive voice over active voice structures in some cases, and especially the flagging of words or phrases which are repeated in
the essays. The issue is that those flagged repetitive words are not limited to content words in the categories of verbs, nouns, adjectives, and adverbs. Instead, it includes also high-frequency non-content words like articles or prepositions, which makes feedback on style much less informative and useful than expected.

With the considerations of how relevant and potentially useful Criterion feedback is for student writing skills at this stage in the development of Criterion and similar programs, the current study only focused on the form-focused feedback. This is further broken down to feedback on (1) grammatical errors, (2) word usage errors, (3) errors in writing mechanics. Figure 3.2 (taken from Dean, 2014) illustrates error categories generated by Criterion.
Figure 3.2. Criterion’s error types based on its scoring engine e-rater v.11.1

For each error identified by Criterion, the specific word or phrase will be highlighted. Once students drag the pointer to this highlighted word/phrase, there will be a pop-up screen giving explanations of the errors as in Figure 3.3 below.
Based on the feedback, students can choose to revise their essay by clicking on “Revise” which initiates a split screen with the right half being an interactive section for keying in corrections, as in Figure 3.4.

Figure 3.3. Example screenshot of Criterion feedback in pop-up screen

Figure 3.4. Criterion’s split screen for revision activities
Structure of the Study

By adopting an argument-based approach to investigating students’ engagement with ACF, the current study aims to provide systematic insights into different aspects using varied evidence that supports or rejects a case for the value of Criterion ACF for EFL learners. Adapting Chapelle et al.’s (2015) validity arguments for automated writing evaluation into the use of Criterion automated corrective feedback in EFL writing classrooms, the current study seeks evidence for four most relevant inferences from Chapelle et al.’s (2015) framework: Evaluation, explanation, utilization, and ramification. The warrants for these inferences become the focus of validation of Criterion ACF in the current study:

- **Evaluation**: Criterion feedback provides students with accurate information to target relevant areas for revision/improvement/learning.
- **Explanation**: Diagnostic results given by Criterion reflect the aspects of a construct of academic writing ability targeted by the writing course.
- **Utilization**: Diagnostic results on the grammaticality and accuracy in student essays obtained from Criterion are useful for students to make decisions on revisions.
- **Ramification**: Use of Criterion ACF is beneficial for students’ L2 learning and acquisition.

The assumptions underlying these warrants were subsequently operationalized in the following research issues:

- The **appropriateness** of Criterion ACF for EFL learners
- The **accuracy** of Criterion ACF
- The **coverage** of Criterion ACF relevant to the EFL learners’ linguistic issues
- The **usefulness** of Criterion ACF for students’ decision-making about revisions

The research issues were more specifically addressed in the five research questions below:

1. What is the nature of Criterion automated corrective feedback?
2. How accurate is Criterion’s automated corrective feedback?
3. How do EFL students engage with and respond to Criterion ACF?
4. How does EFL students’ accuracy in L2 writing change after a semester’s access to and use of Criterion ACF?
5. What is students’ perception of Criterion ACF?
Instruments

To obtain the data that provide answers to the research questions presented earlier, four instruments were used. First, being quasi-experimental, this study uses pre, post, and delayed tests to investigate any observed accuracy change in students’ writing after a semester. The qualitative tools to elicit students’ engagement with ACF include think-aloud protocols, then stimulated recall interviews and focus group interview protocols.

Pre-post tests. The prompts used for the pre, post, and delayed post tests were pre-determined and agreed on by the writing instructors and the researcher. These timed writing essays were forty-five minute in-class tests. Three writing prompts were counter-balanced across the pre, post, and delayed post conditions for the three groups in the study to control for prompt effects. All the writing prompts were taken from Criterion preloaded topics for the TOEFL test and purposefully selected in alignment with the syllabus of participating classes. During the tests, students did not have access to reference materials. The specific prompts used for the tests can be found in Appendix A. An example of the test paper delivered to students can also be found in Appendix A. In the pre-test paper, a subsection at the beginning was used as an elicitation tool to collect students’ bio-data and some relevant information about them.

Writing tasks for practice sessions on Criterion. Pre-loaded essay prompts on Criterion were selected for the purpose of the study and to match the course syllabus of the participating writing classes. After aligning the Criterion pre-loaded prompts to the class sessions, three writing prompts were finalized for the three practice sessions when students composed on Criterion to receive the system’s scoring and corrective feedback. Table 3.1 presents the selected writing prompts.

Table 3.1
Writing Prompts for Practice Sessions on Criterion

<table>
<thead>
<tr>
<th>Practice session</th>
<th>Genre</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Problem solving essay</td>
<td>Reducing Pollution: There are many kinds of pollution. What can you do to help reduce one kind of pollution in your community? Use examples and specific details to explain your answer.</td>
</tr>
<tr>
<td>2</td>
<td>Opinion essay</td>
<td>Money and Success: Do you agree or disagree with the following statement?</td>
</tr>
</tbody>
</table>

56
Think-aloud protocol. The think-aloud protocols served as the primary data elicitation tool, which was supplemented by the stimulated recall interviews to seek clarifications and extensions where applicable. Cohen (2000) pointed out that as self-reports (e.g. interviews or questionnaires) are removed from the mental event being described, the validity of these research tools can be questionable. In this study, the think-aloud protocols were recorded to provide data about learners’ engagement with automated corrective feedback, especially as demonstrated in the participants’ noticing of the feedback received. As a form of self-reports, the think-aloud protocol (TAP) as used in the current study may also be vulnerable to a criticism pointed out by Schmidt (2001) that studies using TAP to shed light on ‘noticing’ as a cognitive process falsely assume that the protocols recorded everything that was noticed.

The above challenges were partially addressed using clear protocols for the administration and collection of TAP recordings, with triangulation from stimulated recall interviews conducted with the participants of the think-aloud procedures. The combination of these two introspective research tools was especially meaningful for the multiple-case study which aims at analyzing explanatory factors for L2 development patterns recorded for the learners. From an SLA perspective, such introspective methodologies “seem the best available options for accessing metalinguistic operations” (Norris & Ortega, 2003, p. 731) and how it relates to long-term L2 learning. Although the think-aloud protocol is not without its own problems, consideration of several possible research tools boiled down to TAP as most potentially productive and plausible for the current research. It is expected to do a better job of revealing what learners actually do, not what they believe they do. In a similar vein, Ericsson and Simon (2003) recommended the collection of what Cohen termed self-revelational data (i.e., from online think-aloud protocols) which directly investigate learners’ cognitive processes over self-observational data from introspective or retrospective verbal reports. The use of think-aloud protocols is still open to debate regarding their validity as a data collection tool. One opposing view challenged the value of verbal reports by

| 3 | Advantage/Disadvantage essay | Change Job or Not: Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Discuss the advantages of each choice. Which do you prefer? Use reasons and examples to explain your choice. |
arguing that most language learning takes place unconsciously, and thus learners are unable to report it (Seliger, 1984).

**Target level of verbalization.** Following Ericsson and Simon’s (1984) classification of three levels of verbalization in doing think-aloud protocols, the current study aims to maintain the subjects’ thinking aloud at the second level, which is defined as involving “description, or rather explication of the thought content” (p. 79). Ericsson and Simon assign to this level “verbalizations that do not bring new information into the focus of the subjects’ attention, but only explicate or label information that is held in a compressed internal format or in an encoding that is not isomorphic with language (e.g., information about odors)” (p. 79). They went on to review several studies on the impact of think aloud protocols on subjects’ processing of different tasks. The conclusion is that thinking aloud does not significantly affect the structure of the cognitive processes if the experimental conditions comply with Level 1 or 2 verbalization criteria. However, they added that “since explication or recoding requires processing time for the subjects but does not replace other processing involved in the task performance, a subject who is verbalizing at this second level can be expected to take more time for the task than one who is not verbalizing” (p. 70). Such a position is questionable from Smagorinsky’s view (1998) who argues that speech activities provide the ways to organize cognition. The second level of verbalizing is thus distinguished from the third level by the fact that the latter requires the subject to explain his/her thought processes or thoughts, which, according to Ericsson and Simon, can alter the participant’s internal structure of thoughts as it directs his/her attention to the procedures involved.

However, it is worth noting that Ericsson and Simon’s discussion was mainly based on studies on problem solving skills to see the number of moves and time duration taken by the participants in reaching a solution to a problem. This presents a fairly different picture from how verbalization can impact the cognitive processes involved in processing linguistic input in the form of corrective feedback in a second language. If being considered a task to be solved, revising the first draft using the different feedback points generated by Criterion is not a continuous process where the previous move leads onto the next. In other words, there is no such closely interrelated link between the processing of a feedback point and another. Instead, they are better viewed as different pieces of information that are independently processed and the final product is not a summative solution to a whole problem as is discussed in most studies in Ericsson and Simon’s (2003) review.

As the current study used TAP as a research tool to elicit data about students’ engagement and processing of ACF rather than problem solving skills, Ericsson and Simon’s interpretation of second level of verbalizing is adapted to ensure relevance to the discussion about processing form-
focused feedback among English language learners. More specifically, in this study, the second level of verbalization can involve the participant’s self-thoughts regarding how or why he/she chooses to respond to or ignore a feedback point, or his/her justification for arriving at a correction in response to the error being attended to. What counts as threatening to change the internal cognitive process being involved is any effort to explain to the researcher who is present during the time a participant is doing the think aloud protocol. To avoid this, the instructions for TAP conduct delivered to the participants include the direction for them to avoid trying to communicate any thoughts or explanations to another person.

The following instruction was thus used for the conduct of students’ think aloud protocols in the current study.

Main part of instruction:

*Try to think aloud. Imagine you are alone and read the feedback to revise your essay. Just let your thoughts speak when you are working on your revisions.*

Complementary instructions: (This is to make sure students are not trying to explain things to another person, considering the fact that the researcher is present during the time TA protocols are recorded to offer prompts where necessary.)

*Do not try to explain anything to me or anyone else. Pretend there is no one here but yourself. You can talk in either English or Vietnamese, whatever you are comfortable with.*

**Stimulated recall interviews.** The stimulated recall interviews, conducted with four of the TAP students, were semi-structured interviews. The interviews included both pre-formulated and spontaneous questions. Guiding questions for the interviews were constructed based on things verbalized in the TAPs, along with the expected issues relevant to the research questions. In other words, the TAPs provided a backbone for main questions to be raised during the stimulated recall interviews. However, for the interaction between the researcher and students to be more informative and relaxing, the researcher also used spontaneous prompts and probes to encourage students to expand their answers. Semi-structured research interviews of the type that allow room for certain sharing and negotiating world views, according to Grindsted (2004), can minimize the threat to validity in research.

Although questions for each stimulated recall with individual students varied, focus areas for these interviews were usually as below:
1. Error types that students notice and types of revision they make in response to automated corrective feedback
2. The reasons for certain corrections they make or for feedback points they ignore
3. Whether the feedback is understandable and easy to follow
4. Difficulties in implementing revisions using the feedback
5. Whether students think Criterion feedback is helpful for improving their accuracy in writing essays in English
6. The expected level of direct/indirect error coding or the level of explicitness of the feedback to errors (e.g., does the student need a suggested correction?)
7. Error types students expect to have most feedback on, and those they do not think feedback is necessary

**The language of the interviews.** The interviews were conducted in either English or Vietnamese, as long as the students were comfortable when they expressed thoughts, concerns, and feelings. This was done by having students sign up for either English or Vietnamese interviews once they had agreed to participate in the case study. Also, taking their varied competence levels in communicating in English into consideration, the interviews were expected to be more informative and data richer if students were given the choice if they wanted to discuss in their mother tongue.

**Focus Group Interviews.** Focus group interviews were structured to help reveal students’ perceptions of Criterion as well as what Storch (2018) listed, their beliefs, goals, and preferences that contribute to students’ behavior as they work on their writing using automated corrective feedback. As a general guide, there were two broad themes for discussion during these interviews, although further topics and reflections were welcome. The two broad themes were actually derived from the pilot study for this research when the focus group students were given a chance to discuss any issue they wanted to mention during the interview after they had completed two sessions on Criterion. Their discussion was much less structured as participants were given the chance to freely comment on different aspects of Criterion. Upon completion of the interview, the researcher examined the issues pointed out and looked for the broad themes participants tended to focus on, which guided the focus group interviews in the main study. Table 3.2 provides the two broad themes followed specific questions used in the focus group interviews.
Table 3.2

*Themes and Guiding Questions in Focus Group Interviews*

<table>
<thead>
<tr>
<th><strong>Theme 1: Students’ use of Criterion feedback</strong></th>
</tr>
</thead>
</table>
| 1 | How do you approach the feedback? Why?  
*Prompt:* Error type by error type from left to right? By the order of importance you place on types of error? Or by the number of errors in different types? |
| 2 | Do you correct all the errors pointed out by Criterion? |
| 3 | Do you read the pop-up screen for each identified error?  
*Probe:* Is there any difference between the first time using Criterion feedback and the second time when you do the TAP? |
| 4 | What strategies do you employ to revise an error pointed out by Criterion?  
*Prompt:* Do you refer to online resources? A dictionary? Check with someone who may know the correct forms? |

<table>
<thead>
<tr>
<th><strong>Theme 2: Students’ perception of Criterion feedback</strong></th>
</tr>
</thead>
</table>
| 1 | What do you think of the way Criterion gives feedback on the grammar errors in your essay?  
*Probe:* Do you expect Criterion to provide corrections for your errors or prefer it just to point out the error, then you can correct it by yourself? |
| 2 | Is Criterion language of feedback easy to understand and follow? |
| 3 | Is Criterion ways of giving feedback easy to understand and follow? (The way it highlights the errors, then give further explanations in pop-up screens, and split screens during revision time) |
| 4 | What kinds of error do you think Criterion gives most effective feedback on? Which kinds of error do not need feedback on as you can correct them yourselves? |
| 5 | Are there any things that can be improved on Criterion grammatical feedback? |
| 6 | Generally, do you think Criterion feedback is accurate? Helpful? |
| 7 | Any further comments and reflections on using Criterion feedback to revise your essays? |

**Participants**

Data for the study came from a total of 104 students, with participants in different parts overlapping. Table 3.3 presents the exact number of participants for different parts of the study. All participants came from three classes of writing which were selected for the study based on the researcher’s previous connections with the writing instructors in the English Department. Of this convenience sample, student numbers in each class ranged from 32 to 48. They were second-year English majors at a university in central Vietnam where they were doing a four-year course in
English Teacher Education Program. At the time data were collected for this study, they were in the 4th semester out of a total of eight semesters towards their B.A. degree.

Table 3.3

Participants for Different Parts of the Study

<table>
<thead>
<tr>
<th>Part of the study</th>
<th>Comparison group</th>
<th>1st + 2nd Experimental groups</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre to post-test accuracy change analysis</td>
<td>37</td>
<td>38</td>
<td>75</td>
</tr>
<tr>
<td>Revision analysis</td>
<td>0</td>
<td>39 (from F’s class) and 28 (from M’s class)</td>
<td>67</td>
</tr>
<tr>
<td>Think-aloud protocol analysis</td>
<td>0</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Stimulated recall analysis</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Focus group analysis</td>
<td>0</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>

The classes were taught by two experienced EFL writing instructors, one male and one female. To allow for maximum variation, the teachers were targeted based on their gender, experience in teaching the writing skills to English majors, and experience as well as their comfort with using technology in the classroom. The male instructor was in charge of one of the two experimental groups in the study. On the other hand, the female teacher instructed two writing classes for this study, one comparison and one experimental group. The male teacher had 16 years of teaching experience in the tertiary EFL context, and he earned a PhD in Applied Linguistics from an Australian university about half a year before he took part in this study. He was quite familiar with the use of technology in the classroom and expressed enthusiasm about the implementation of an automated program like Criterion to support his teaching of the course. He often communicated his course contents with students through an online learning platform. His course materials as well as other references related to the writing course in this study were all uploaded on the website, along with the use of Criterion as a supplement to teacher instruction for the writing course.

Turning to the female instructor, with an M.A. in TESOL, she had more than 20 years’ experience in teaching writing in the tertiary EFL context. Yet unlike the male teacher, she was more into traditional approaches in writing instruction, where the use of automated essay evaluation was a totally new concept to her. She admitted being intimidated by such recent applications in writing instruction and assessment and was also generally not confident about using technology in her teaching. The variable qualities of these two writing instructors in terms of gender, experience, and exposure to technology in the writing classroom made them the optimal choices for the study.
The writing class taught by the male teacher (hereafter referred to as M) had 32 students, with only three males and 29 females. Attendance was not always good for this class, with the number of students present being around 26 to 28 students per session (about 84%). The experimental group instructed by the female teacher (referred to as F hereafter) included 40 students, only one of whom is male. The comparison group, also taught by F, had 48 students (45 females and 3 males). Students in both groups taught by F attended most sessions, with attendance rate always exceeding 95%.

**Participant Inclusion Criteria**

The final number of participants was different for different parts of the study, and the selection criteria for each part is explained in the sections below.

**Participants for pre to post-test accuracy change analyses.** Students for this part of the study came from the three participating classes (subdivided into the experimental and comparison groups) whose number totaled 120. However, the number of students remained at 75 for inclusion in the accuracy analyses from pre to post and delayed post-tests. The data of students who missed one or more of the tests (pre, post, or delayed post-tests) were omitted from this analysis. For those who completed all of the three tests, pre-test essay length statistics revealed some outliers where certain students in the two groups produced very short essays (less than two standard deviations from the group mean). A further look at the essays showed students’ incomplete answers, short response to the prompt, indicating lack of a serious effort in writing the essay. This resulted in the decision to omit a few cases in the comparison and experimental groups. Initial comparison of two groups’ pre-test overall accuracy scores also indicated significant difference. Therefore, the next step was to further clean up the data by looking at individual students’ pre, post, and delayed post-tests’ accuracy scores. In an effort to make two groups more comparable and minimize skewed data during the actual accuracy analyses, four cases in the experimental group who performed highly erratically from pre to post and delayed post tests were omitted from the data. These students had significantly different pre, post, and delayed post test accuracy scores. The total number of students remained at 75 for accuracy analyses in this study.

The information about these 75 participants can be found in Table 3.4. The information in this table includes both the participants’ bio-data elicited from the short questionnaire completed by students in a section on the pre-test paper they received, and the word length counted in total words of their pre-test essays.
Table 3.4
Participants’ Demographics for Pre-post Test Change Analyses

<table>
<thead>
<tr>
<th></th>
<th>Comparison group</th>
<th>Experimental group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total participants (81)</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Gender distribution</td>
<td>3 males</td>
<td>2 males</td>
</tr>
<tr>
<td></td>
<td>34 females</td>
<td>36 females</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>SD</strong></td>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>Age</td>
<td>19.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Years of learning English</td>
<td>10.3</td>
<td>2.0</td>
</tr>
<tr>
<td>IT skill (self-rating 1-9)</td>
<td>4.6</td>
<td>1.4</td>
</tr>
<tr>
<td>Word length in pre-test</td>
<td>166</td>
<td>48</td>
</tr>
</tbody>
</table>

The vast majority of participants were female. The fact that female students far outnumber their male counterparts in this study reflects a typical gender distribution in many English departments across different universities in Vietnam. The two groups, comparison and experimental, are comparable in terms of other criteria including age, years of learning English, and their self-ratings of IT skills. However, pre-test essay length for the comparison group averaged at 166, while that for the experimental group is 218. Based on this criterion alone, different average essay length can be indicative of some differential proficiency levels, taking into consideration the fact that the pre-test was administered during the second week of the semester for both groups under similar conditions such as time constraints and writing modality.

**Participants for revision analyses.** This part of the study included all the students who submitted at least one revised draft on *Criterion* during one of the three in-class practice sessions. They attended at least one practice session, and as long as there was a first draft followed by a revised draft, they were included in the total data for revision analyses. The rationale behind this was to collect as much information as possible regarding how learners generally used and responded to *Criterion* ACF, irrespective of their performance or level of engagement with the system. In other words, the ultimate aim was to obtain a picture which is as rich, diverse, and accurate as possible to reveal how the feedback was perceived, used, and processed by learners. Using this criterion, the total number of students included in this part of the study was 67.

**Participants for think-aloud protocols.** The students selected for think-aloud protocol sessions were primarily based on voluntary participation as well as teachers’ recommendations.
After 12 students volunteered to be involved in the think-aloud protocol procedures, the female instructor recommended three more participants who she thought would be interesting cases to examine further based on her previous experience working with these learners. A total of 15 students were invited to participate in this procedure, and 14 agreed to be involved. This group included one male and 13 females. All the students attending the TAPs were remunerated for their participation.

**Participants for stimulated recall interviews.** Based on the TAP recordings, certain participants were subsequently contacted to attend a stimulated recall interview one week after their think-aloud protocol session. Criteria for inclusion in this interview were the richness of the data in students’ TAPs, as well as the necessity to clarify certain issues that emerged in the recordings that might potentially contribute meaningful information to the results. Of the total 14 students who completed the TAP procedures, four students had a stimulated recall interview.

**Participants for focus group interviews.** At the beginning of the semester, after being introduced to the research, a subsample of 20 students from the three participating classes were randomly selected and invited to attend the focus group interviews which would take place towards the end of the course. A total of 19 students (one male and 18 females) agreed to attend the interviews and they were divided into 4 focus groups. One group had four students while the remaining three had five students each. All the students attending the focus group interviews were remunerated for their participation.

**Ethics**

Before the start of the project, ethics approval was sought from the Human Research Ethics Committee of The University of Melbourne. In November 2016, approval for conducting the research was granted and Approval ID is 1648139 (see Appendix G).

**Data Collection Procedures**

Overall procedures for data collection among the experimental group throughout the second semester of the academic year 2016-2017 can be found in the following timeline (Table 3.5). The whole semester ran for 15 weeks and the classes met once a week, with each session lasting 100 minutes.
Table 3.5
Data Collection Timeline for the Experimental Group

<table>
<thead>
<tr>
<th>Week</th>
<th>Class activity</th>
<th>Research plan</th>
<th>Data collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td><em>Researcher collected preliminary data about students in two writing classes through two pre-course meetings with the writing instructors, regarding the total numbers of students in each class and their general overall proficiency levels. During these meetings, the researcher also discussed the alignment of data collection timeline with class schedules with the two teachers in charge.</em></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Introduction to the course</td>
<td>Getting to know participants and introduction to the research project.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Diagnostic test</td>
<td>Pre-test (with detailed bio data of participants as part of the test paper where students input their information on the test paper)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture 1</td>
<td><em>Criterion</em> training and account setup</td>
<td></td>
</tr>
<tr>
<td>2’</td>
<td>Practice session 0 (Homework): Students logged on <em>Criterion</em> using their account to do a self-practice session on a pre-loaded prompt at home, taking time with their composing and revising on the system</td>
<td>Researcher logged on <em>Criterion</em> to see the number of student submissions for this trial session.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Lecture 2</td>
<td>Researcher followed up with further instructions on how to use <em>Criterion</em> in future in-class practice sessions</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Lecture 3</td>
<td>Training of TAP procedures among a potential group of students</td>
<td>Trial think-aloud protocols (TAPs) among 2 students</td>
</tr>
<tr>
<td>5</td>
<td>Intervention 1: In-class writing exercise 1 (Practice session 1)</td>
<td>Students do their composing and revision with <em>Criterion</em> for the first time as an in-class writing exercise. The researcher answers questions on any problems students have as they interact with the system. The researcher collects students’ first and revised drafts.</td>
<td>Students’ first and revised drafts for Essay 1. Conducting TAP procedures among 4 students (two from each experimental group).</td>
</tr>
<tr>
<td>6</td>
<td>Lecture 4</td>
<td>Conducting stimulated recall interviews with selected TAP students</td>
<td>One stimulated recall interview.</td>
</tr>
<tr>
<td>7</td>
<td>Lecture 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Intervention 2: In-class writing exercise 2 (Practice session 2)</td>
<td>Collecting students’ first and revised drafts. Conducting think-aloud protocols for 5 students (from both experimental groups).</td>
<td>Students’ first and revised drafts for Essay 2. 5 TAPs.</td>
</tr>
<tr>
<td>9</td>
<td>Lecture 6</td>
<td>Conducting stimulated recall interviews with selected TAP students</td>
<td>Two stimulated recall interviews.</td>
</tr>
<tr>
<td>10</td>
<td>Lecture 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Intervention 3: In-class writing exercise 3 (Practice session 3)</td>
<td>Collecting students’ first and revised drafts. Conducting think-aloud protocols for 5 students (from both experimental groups).</td>
<td>Students’ first and revised drafts for Essay 2. 5 TAPs.</td>
</tr>
<tr>
<td>12</td>
<td>Lecture 8</td>
<td>Conducting stimulated recall interviews with selected TAP students</td>
<td>One stimulated recall interview.</td>
</tr>
<tr>
<td></td>
<td>Midterm Revision and Course Evaluation</td>
<td>Posttest</td>
<td>Posttest essays</td>
</tr>
<tr>
<td>---</td>
<td>--------------------------------------</td>
<td>----------</td>
<td>----------------</td>
</tr>
<tr>
<td>13</td>
<td>Debriefing interviews with four groups of students other than the TAP students (4 or 5 students in each group)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>No class for four weeks when students were away on their defense training program (part of the study curriculum when students are required to learn basic skills and knowledge related to military and defense for youth.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Lecture 9 Revision test</td>
<td>Delayed post test</td>
<td>Delayed posttest essays</td>
</tr>
<tr>
<td></td>
<td>Course wrap-up and evaluation</td>
<td>Focus group interviews</td>
<td>Four focus group interviews</td>
</tr>
</tbody>
</table>

*Note.* Highlighted rows show the sessions when students used *Criterion* for the lesson and teaching/learning took place in the computer lab.

Data collection in the comparison group ran in parallel with that for the experimental group, yet without the three practice sessions taking place in the computer lab when students worked on the computer to compose and revise essays using *Criterion* feedback. Instead, comparison-group students did their practice essays on the same topics on paper in their normal classroom. They submitted their handwritten essays to the teachers and no revised drafts were required. No think-aloud protocol procedure nor stimulated recall/focus group interview was conducted for this group either.

**Pre-post test collection.** As can be seen from the timeline, the pre-test was conducted in the second week of the semester, the post-test took place in the 13th week, and the delayed post-test a month later. There was no class during the four weeks between the post and delayed post-tests as students had to attend the compulsory national defense training trip as part of the learning curriculum for all university students in Vietnam. The delayed post-test was taken as both the in-class midterm test for assessment purposes of the course and a revision exercise before students sat their end-of-semester examination.

The pre-test, post-test and delayed post-test were carried out in paper-based format. All the tests were timed essays and students were given a total of 45 minutes to compose an essay of
between 250 and 300 words in response to a prompt which was relevant to the genres of persuasive and expository essays they had learned in the semester. Students were informed that their essays for these tests would account for 30% of their final grade for the course, 10% each. By aligning student essays for the study with their in-class ongoing assessment tasks (counted as the mid-terms and revision tests), the researcher attempted to shorten the gap between experimentally elicited data needed for the study and what Ellis and Barkhuizen (2005) termed naturally occurring samples which are supposed to be more valid evidence of learner interlanguage. Ellis and Barkhuizen argued that timed essays written in examination provide naturally occurring samples of learner data as an examination “constitutes a ‘natural’ context for learners to use the L2 and data so obtained have not been designed for purposes of research” (p. 50). The tests were done in paper-based format as the essays from these tests are used as data to measure students’ linguistic accuracy gains (or the lack thereof) over a five-month period after they had used Criterion corrective feedback in their writing course. Considering the teaching and learning context where data for this study was collected, most students were used to composing their essays in hand-written format (for both regular practice activities and tests). The choice of this traditional writing modality, therefore, aims to rule out the interference of typing skills as students write, focusing more on their actual productive skill of writing. Thus, students’ different computer literacy levels were controlled for.

**Criterion training.** Time constraints and the need to stick to the 15-week teaching schedule somewhat limited extensive in-class training of the use of Criterion. Nevertheless, the researcher attempted to familiarize both the instructors and students in the experimental group with Criterion. As seen in Table 3.5, for the writing teachers, the researcher spent half of the first meeting with them to introduce them to Criterion and get their accounts set up after they discussed the research project as well as agreed on the data collection timeline. For learners, during the second session, the researcher was given about 20 minutes to introduce Criterion. Also during these 20 minutes, students’ emails and relevant information were collected so that after the session, the researcher could create individual students’ accounts and send them the link for account activation. This was followed by a homework task which required students to log into their accounts for a self-practice session (treated as practice session 0) on Criterion. This was a change from the original plan as time did not allow for an extra in-class practice session on the program. To make up for this, the researcher followed up by checking student submissions of the homework task. Information about students who submitted essays for practice session 0 gave the researcher an idea of learners’ actual work on Criterion so that further help in the following sessions could be offered. The researcher came in each class during the third meeting to answer questions about the problems students encountered as they started using Criterion. Other more practical questions about using the system
were answered during the first in-class practice session when students were required to compose essays under time constraints and to revise their first drafts incorporating Criterion feedback. The researcher was also present during the remaining practice sessions to offer prompt technical help with Criterion.

During the three official practice sessions, students had few issues with composing and revising on Criterion. There were no major problems that affected their performance with the assigned writing tasks, and issues were mainly about occasional slow responses from the system due to internet connection speed. Otherwise, all of the three practice sessions went smoothly.

**Procedures for in-class practice sessions.** For the four in-class writing tasks (the first one being a trial for home practice), students interacted with Criterion, the online writing instructional program, to compose essays by typing their drafts and feeding them into Criterion. Students were given 45 minutes to write their first drafts. After Criterion provided instant corrective feedback on their first drafts, students revised their essays (after a 5-minute break) and resubmitted the second drafts to the system. Revision time was at maximum another 30 minutes. Students were informed that these second drafts would be assessed by the teacher as part of the evidence for their writing portfolios which accounted for 10% of their final grade. The researcher tried to align students’ work for data collection with the assessment of this subject as it helps to make students take Criterion feedback seriously while improving their drafts/writing.

**TAP training procedures.** The students doing TAPs for this study were trained prior to actual TAP recording. Each selected student did their TAP once but there was one trial session at the beginning to help them feel comfortable with the process, thus preventing any issues when the real sessions took place.

During the trial session, the researcher explained TAP procedures to students and showed them a Youtube video clip in which the think-aloud protocol was used as a data collection tool for demonstration. After a few questions raised by the participants about TAP, the researcher gave instructions on doing TAPs, followed by her demonstration on conducting the think-aloud protocol before students did it themselves. Then, she asked two of the participants to think aloud as they went through their drafts with Criterion feedback available. When students trialed their think-aloud procedures, the researcher gave reminders if they lapsed into silence for 20 seconds or more. Reminders were as simple as “Keep talking” or “Go on”. Other more intrusive prompts (e.g., “What do you think about this point?” or “What are you thinking about?”) were not used to avoid diverting the subject’s thoughts to a different direction. Through this trial session, the researcher aimed to help the participants get used to the act of verbalizing their thoughts as they went along revising
essays. Therefore, participants were allowed to pause to ask questions regarding this activity whenever they had difficulty. This was followed by the whole group questions and answers time for clarifications of the procedures.

**TAP collection.** The students selected to do TAPs were recorded in a room nearby on the same day of the in-class practice sessions. Their think-aloud protocols were conducted in either English or Vietnamese depending on their personal choice. They were given a maximum of 35 minutes to revise their essays. In each practice session, the five TAP students did the composing of their first drafts with their peers in the writing lab under the supervision of their writing instructors. After completing the first drafts, while other students continued with revisions in the computer lab, these five students would take turns to do their revisions in a separate room using a wifi-connected laptop. Each of these students revised their essays and verbalized their thoughts in the presence of the researcher. The TAP students were given five extra minutes compared to others not doing TAPs as talking aloud while processing the feedback may add more time to the completion of tasks, though, as pointed out in quite a few studies, it does not necessarily change the internal processes involved (Ericsson & Simon, 1993; Leow & Morgan-Short, 2004; Sachs & Polio, 2007).

Before actual revising and thinking aloud, the students were given a few minutes as warm-up time to get used to the experimental situation (check the microphones, tape recorders, etc.). Afterwards, the researcher just sat next to the student to offer minimal reminders where necessary. Other than that, any interruptions or interventions were avoided.

The think-aloud protocols were recorded using the Open Broadcasting System (OBS at https://obsproject.com/) software to record both students’ verbalizations as well as their activities on the computer screen as they worked on the essays. Despite training, however, there were three participants who had difficulty with their first think-aloud session and requested having the chance to re-do it. They were allowed to take part in the next TAP procedures and their second TAP recordings were taken as data for the study. One student, who did the TAP twice, had both of her recordings selected as she provided an excellent example of conducting the think-aloud procedures and the researcher recommended that she did a second recording as a sample session in the presence of those who had difficulty verbalizing their TAPs and thus needed more guidance doing it. This sample recording was also included in the data.

**Stimulated recall interview collection.** After listening to students’ recorded TAPs, the researcher proceeded with selecting one or two students from the batch of five TAP students of the week for the stimulated recall interview. The criteria for this selection were the richness of data in TAPs and interesting issues that had emerged from students’ verbalized thoughts. These issues
became the backbone against which the researcher constructed guiding questions for the interview to take place a week later. A week’s gap between the TAP and interview could have threatened the value of the stimulated recall when further clarifications on certain points in the TAPs were very much dependent on their memories of their processing of Criterion feedback and the freshness of the composing experience. However, this lapse of time offers a reasonable time gap for constructing the questions and negotiating a meeting time with the selected students.

Moreover, the effects of memory constraints and reconstructive processes (Leow & Morgan-Short, 2004) which make students’ reflections during stimulated recalls less reliable for research purposes can be made up for by the fact that during the interviews, students were able to watch the video recording of their work on Criterion the week before. This video showed their on-screen interaction with the program. When the researcher wanted to focus on a certain portion of verbalized thoughts, she would show the relevant part of the video when the student made revisions to their text. This helped with refreshing students’ memory as they went along with the interviews. All the stimulated recall interviews were audio-recorded.

**Focus group interview collection.** The focus group interviews at the end of the semester were conducted among four groups of students other than those doing the TAPs. These students gathered in a classroom after the final session of their writing course (Week 15) and were asked open questions regarding their comments and reflections after using Criterion feedback for a whole semester. During these focus group interviews, students were given time to discuss with their peers relevant issues around the two broad themes set out by the researcher in the focus group interview protocol (see page 61 above for details). The interview was informal and the researcher’s main role was delivering guided questions for the group discussion as well as facilitating the smooth flow of their conversation through giving reminders and prompts for discussion.

**Data Analysis Procedures**

**Data cleaning.** After a semester-long implementation of the treatments using either the teacher-automated feedback combined conditions or the teacher-only feedback condition, the researcher sorted through the data to look for cases of students who completed all the base tasks related to the treatment conditions. This means that for accuracy change analyses, students needed to complete all the pre, post, and delayed post tests. For other analyses, each student needed to attend at least one of the three writing practice sessions and had composed essays in response to the writing prompts. For students in the experimental group, they needed to submit at least one revised draft of their essay using Criterion feedback for each practice session. From these criteria, nine students in M’s group and 35 of F’s experimental group were selected, making a total of 44 students
for the combined experimental group, while 37 students in the comparison group met the criteria. Further omission of students from the accuracy change analyses took place following the screening of outliers in the experimental group. This resulted in a sample of 75 students. Individuals who took part in the think-aloud protocols and focus group interviews were accordingly drawn from this sample, with a few exceptions when students who volunteered to participate in the focus group interviews missed one task in the three practice sessions or missed one of the three tests. One TAP student missed the pretest, and another attended all the practice sessions but failed to submit a revised draft for one session. These two cases were still included in the sub-sample due to their contribution to the whole data as being quite significant, sitting two TAPs each.

**Hand-written essays.**

**Retyping.** Altogether, there were a total of more than 200 handwritten essays collected for the pre, post, and delayed post-tests. These essays were typed up by a research assistant and checked for accuracy. After all the retyping was finished, the typed-up essays were checked the second time by the researcher to make sure there were no mistakes. There were some issues related to unclear handwriting and this required some guessing work from the typist and the researcher. Considering the fact that retyping took place a while after the data collection period was over, contacting individual students for clarifying their writing was not feasible. Therefore, in case some ambiguity was experienced, the typist discussed any issues with the researcher to agree on a resolution.

**Coding clauses in essays.** Coding the typed-up essays in the pre, post, and delayed post-tests went through two stages. In the first stage, all the essays were coded for clauses using a coding scheme (See Appendix B for the *Guidelines for Coding Clauses*). After preliminary coding of a sample of student essays using mainly Evans, Hartshorn, Cox, and de Jel's (2014) guidelines for coding clauses, initial rules were modified to reflect the issues which arose as the researcher went through the first set of essays. Guidelines for coding clauses were later further adapted using certain guidelines for clauses from Polio (1997) and Cumming, Kantor, Baba, Eouanzoui, Erdosy, and James (2006). Examples provided in Appendix B are thus drawn from the data besides those provided by Polio (1997) and Cumming et al. (2006). A sample of 20 essays were double coded, making roughly 10% of the total essays in this part of the analysis. Both intra-rater reliability and inter-rater reliability were calculated to make sure satisfactory agreement was reached before coding of all the essays took place.

In the second stage of coding, errors in clauses in student essays were coded. The coding guidelines for errors was constructed by incorporating rules from a range of sources, some of which being Polio’s (1997) guidelines for coding errors and Storch and Tapper’s (2009) study, in
combination with issues that emerged from the data of the current study (see Appendices C and D for Guidelines for Error Coding). This part of coding is explained in detail in the next section.

**Coding learner errors.** Linguistic accuracy in student writing over the five-month period was calculated by different measures using two major approaches in dealing with learner errors. For a holistic approach to error analysis, the most frequently used tools for measuring linguistic accuracy include holistic scales, error-free T-units (EFTs), and error count. Polio (1997) suggested that “in cases of homogeneous populations, a more fine-grained measure of accuracy such as an error-count may be a better option” (p. 117). She argued that an EFTs approach disregards the number of errors in each unit as well as the severity of each error. In the meanwhile, a holistic scale assesses linguistic or grammatical accuracy along with other components making up the scoring rubric of writing skills. Such an approach is hardly informative in terms of the actual errors made, not to mention the number of errors in each grammatical category. However, as second language learners’ errors are notoriously diverse and complex, error coding using Polio’s error-count approach proved to be a formidable task for the current study. In an empirical paper she later co-authored, Polio admitted coding individual errors could be a great challenge “in that depending on how a sentence is corrected, the number and types of errors can vary” (Lavolette, Polio, & Kahng, 2015, p. 58). Foster and Wigglesworth (2016), on the other hand, suggested,

> For most research purposes, global measures based on a syntactic unit can provide a better evaluation of accuracy in L2 performance than can local measures, and the best tool in the measurement kit is currently the error-free clause because it combines a reliably defined and valid unit with a finer-grained analysis than offered by a whole T-unit or AS-unit analysis. (p. 104).

Therefore, with the dual aims of yielding a more nuanced measure of overall accuracy and yet addressing the practical challenge of being consistent in coding L2 learner errors, this study adopted Foster and Wigglesworth’s (2016) weighted clause ratio (WCR) approach. WCR takes the use of error-free clauses (EFC) as an index of learners’ overall accuracy demonstrated in their written texts. Following this approach, clauses were coded as error-free (entirely correct) or incorrect. Incorrect clauses were further coded as being at level 1, 2, or 3 depending on the gravity of the errors they had. In other words, based on the errors’ impact on the comprehensibility of the clauses, they could be coded at “1” for having only minor errors that did not compromise meaning; “2” for containing serious errors, but the meaning is recoverable, though not always obvious; and “3” if the clause has very serious errors that made the intended meaning far from obvious and only partly
recoverable. Further details and examples for coding errors using WCR can be found in Appendix C.

**Initial checking of coder reliability.** For each stage of coding the test essays (i.e., coding clauses and coding errors), coding began with the researcher doing intra-rater reliability check by conducting the coding twice with a gap of ten days between two coding times to trial the guidelines. After intra-rater reliability reached a satisfactory level, coding proceeded to testing inter-rater reliability which involved a second coder on a sample of about 20 percent of the total data. This was done to check for differences in interpretation of the guidelines and to resolve any ambiguity or misinterpretation before the whole set of essays were coded.

For clause coding, intra-rater reliability was calculated using percent agreement which was at 95% and inter-rater reliability at 91% agreement. Error coding, as expected, was a much more challenging task and a lot of issues emerged, among which were the initially overwhelming number of error types targeted or the two annotators’ different interpretations of certain codes based on their different reformulations of the incorrect structures. After a reduced list of targeted error types was finalized, intra-rater reliability was 80%, just good enough for coding of both annotators to take place. Yet after three rounds of double coding, with a lot of discussion and adaptations to the original guidelines, inter-rater reliability statistic was still low at 46%. This resulted in a change in error coding approach from the initial error count to using an alternative approach discussed below.

**Checking coder reliability.** Inter-coder reliability for coding the errors related to the targeted morphemes in the Obligatory Occasion Analysis is at 90.5%, and that for the general accuracy using the weighted clause ratio is 81%. As shown in previous research (e.g., Lavolette, Polio, & Kahng, 2015; Polio, 1997) as well as hands-on experience in the initial error coding stage of the current study, coding errors in learner essays is seen as a great challenge. It is, therefore, reasonable to adopt Shrout’s (1998) suggested standards for interpreting reliability statistics where .81 to 1.00 can be considered substantial reliability. The inter-rater reliability estimates of 90.5% and 81% for double-coding of the errors in this research tend to satisfy the generally accepted baseline of 80% inter-coder agreement.
Target forms in obligatory occasion analyses. Beside the holistic measure, for a more in-depth investigation of whether Criterion corrective feedback facilitates improved accuracy in specific grammatical morphemes most relevant to the students in the current study, obligatory occasion analysis was used on a smaller number of error types. These errors were shortlisted based on Criterion’s counts of errors in all of the experimental students’ drafts from the three practice sessions. As seen from Table 3.6, these five most frequent errors include spelling, missing or extra article, subject-verb agreement, missing comma, and determiner-noun agreement. This analysis provided more information regarding students’ interlanguage development as they incorporated automated feedback into their learning process. According to this approach, spelling, article, subject-verb agreement, comma use, and determiner-noun agreement scores were calculated as the percentage of target-like use in an obligatory context.

Table 3.6
Five Most Frequent Error Types in Students’ Essays Tagged by Criterion

<table>
<thead>
<tr>
<th>Category</th>
<th>Error Type</th>
<th>Total Criterion error tags</th>
<th>Operationalised as…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanics</td>
<td>Spelling</td>
<td>1656</td>
<td>Spelling</td>
</tr>
<tr>
<td>Usage</td>
<td>Missing or Extra Article</td>
<td>1539</td>
<td>Use of indefinite articles <em>a</em>, <em>an</em> and definite article <em>the</em></td>
</tr>
<tr>
<td>Mechanics</td>
<td>Missing or Extra Comma</td>
<td>587</td>
<td>Missing comma, extra comma</td>
</tr>
<tr>
<td>Grammar</td>
<td>Subject-Verb Agreement</td>
<td>442</td>
<td>Third person singular, singular and plural copulas</td>
</tr>
<tr>
<td>Usage</td>
<td>Determiner Noun Agreement</td>
<td>322</td>
<td>Plural s/es with pre-modifying quantifiers</td>
</tr>
</tbody>
</table>

Statistical analysis. To determine what statistical analysis was most suitable for the data, pre-test scores on different accuracy measures of the two groups were compared. Specifically, independent-samples T-tests were conducted to see if there was any statistical difference between the two groups’ pre-test accuracy scores in terms of the targeted structures. Table 3.7 presents the relevant results.
Table 3.7

Testing Statistical Difference in Two Groups’ Pre-test Scores for Different Accuracy Measures

<table>
<thead>
<tr>
<th>Accuracy measure</th>
<th>Independent samples t-test result</th>
<th>Statistical difference?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall accuracy</strong></td>
<td>$t(73) = 4.56, p &lt; .001$</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Article</strong></td>
<td>$t(73) = .896, p = .373$</td>
<td>No</td>
</tr>
<tr>
<td><strong>Third person singular</strong></td>
<td>$t(48.11^*) = 3.640, p = .001$</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Copula use</strong></td>
<td>$t(70) = .466, p = .643$</td>
<td>No</td>
</tr>
<tr>
<td><strong>Plural s/es</strong></td>
<td>$t(72) = .813, p = .419$</td>
<td>No</td>
</tr>
<tr>
<td><strong>Comma use</strong></td>
<td>$t(73) = -1.458, p = .149$</td>
<td>No</td>
</tr>
</tbody>
</table>

* Equal variances not assumed

The statistical procedure for analyzing results related to student accuracy changes for all of the six measures was Repeated Measures Analysis of Variance (RM ANOVA, Fisher, 1932, 1935, cited in Rutherford, 2000). As seen in Table 3.7, there was no significant difference in the pre-test scores for article, copula use, plural s/es, and comma usage. For these four measures, therefore, RM ANOVAs were conducted using accuracy scores obtained for the pre, post, and delayed post-tests. The between subject effect was group, and the within subject effect was time. Mean difference between the post-test and pretest is interpreted as intervention effect on the experimental group (or acquisition), and that between the delayed post-test with the post-test as retention.

Regarding the two remaining measures of general accuracy and third person singular use, a significant difference in the two groups’ pre-test scores was detected. As it is important that all analyses adjust for baseline levels, for these two measures, gain scores were used for inferential analyses using RM ANOVA. Such a statistical choice aims to eliminate confounds (Field, 2009), in this case being student differential baseline levels before the intervention was implemented. Two gain scores were calculated. Gain score 1 was obtained by subtracting pretests from post-tests to show the intervention effect, while the second gain score was post-tests subtracted from delayed post-tests to evaluate retention. Therefore, intervention effect and retention are used interchangeably with Gain Score 1 and Gain Score 2 hereafter.

Meeting statistical assumptions. Assumptions required of repeated measures ANOVA are presented below.
**Normality, linearity, range.** In general, accuracy scores for general accuracy, use of article, third person singular, singular copula, plural copula, plural s/es, and comma usage met the necessary assumptions regarding normal distribution, linearity, and range. Scores were slightly negatively skewed for most sets of accuracy scores, showing that the frequent scores were more clustered towards the higher end of the scale, but no issues with kurtosis were detected (Field, 2009; Larsen-Hall, 2010).

**Sphericity.** Another assumption for repeated measures ANOVA analyses to be conducted, sphericity, was also met. Mauchly’s tests on all the sets of pre-test, post-test, delayed post-test scores for each accuracy measure, as well as gain scores for the two measures of overall accuracy and third person singular use indicated that the assumption of sphericity has been met for the main effects of the variables under study (Field, 2009; Larsen-Hall, 2010).

Based on these exploratory analyses, it was decided that the statistical assumptions for inferential statistics were met by the data.

**Student first and revised drafts on Criterion.** First and revised drafts written by students during practice sessions on Criterion were used for three different yet related analyses: the verification of Criterion error tags for feedback precision, students’ revision operations, and their revision success rates. How each of these analyses was conducted is presented in detail below.

**Verification of Criterion feedback.** According to Ranalli, Link, and Chukharev-Hudilainen (2017), the measurement of the quality of automated corrective feedback generated by systems such as Criterion is usually operationalized in the two statistics of precision and recall in many studies that aim to evaluate the quality of automated corrective feedback. In the current study, to measure how accurate Criterion form-focused feedback is, only the statistic precision was calculated. The exclusion of recall was based on one of the foci of the current study, which is students’ engagement with the feedback. The verification of the actual feedback points students received helps to contribute to the explanation of their revisions and perception of Criterion ACF as a whole. On the other hand, recall, which indicates how many errors Criterion failed to detect, could be an interesting extra piece of evidence, yet it had little to do with what the study set out to investigate.

In this study, precision was adopted based on the Verification Approach used by Gamon, Gao, Brockett, Klementiev, Dolan, Belenko, and Vanderwende (2008, as cited in Leacock, Chodorow, Gamon, & Tetrault, 2014) to investigate the feedback points generated by Criterion in student first drafts from three practice sessions. This approach is “a method of simply checking the
acceptability of a system’s error flags or corrections compared to what the learner wrote”, and thus, “it does not permit estimation of recall” (Leacock et al., 2014, p. 90). In the current study, two annotators, with the researcher being one of them, annotated the corpus of all the first drafts for 
Criterion feedback points as:

- **Correct error code (CC):** to indicate that an error was correctly identified by Criterion.
- **Incorrect error code (IC):** to indicate cases when Criterion appropriately coded a structure as incorrect but gave it a wrong code (adapted from Lavolette et al., 2015), or when Criterion offered a confusing error message that failed to pinpoint the nature of the problem.
- **False positive (FP):** to indicate that the system provided a false alarm by flagging a correct structure in essays as an error.

Sometimes the coding can be challenging as in the following case:

**Student text:** Nowadays, some people take the option to change jobs during their careers while *another* want to stay in the same job.

**Error message:** *These two words belong together. Write them as one compound word.*

The system’s suggestion to write “an other” as one word, “another”, makes sense. The error code can be considered correct but judged by a usefulness criterion, a useful feedback point should direct learners to using “others” instead of “another” to make the whole sentence grammatically correct. In this case, as well as other similar cases, judgement in coding should prioritise consistency between the error code and its explanatory message. If Criterion can indicate an error at the local level, correctly identify its type of error, and then suggest the appropriate change to fix that issue, then the error code is correct. This does not take into account the impact of the error code on the whole clause or sentence, and thus draws our attention to a fact that ACF can be correct but may not be useful all the time.

**Calculation of precision.** Precision equals the total CCs divided by the sum of CCs, ICs and FPs. Where there was a disagreement between the two annotators, they would discuss the difference and its reasons to reach an agreement; if no agreement could be reached, further consultation would be sought from native speakers of English. Double annotation of a subpart of all the error tags retrieved from Criterion (about 10 percent) produced an inter-coder reliability of 89 percent agreement. After all the differences between the two annotators for this 10 percent subpart were resolved, coding of all the error tags proceeded. This annotation approach does not take into consideration cases when Criterion missed an error in students’ writing (i.e., the recall value). Such
a methodological choice was founded on a fact that Criterion ACF should prioritize precision over recall as the aim for improvement to the system because the developers espoused the policy that it is better for Criterion to miss an error than flag a correct structure as erroneous (Burstein, Chodorow, Leacock, 2004). This piece of information regarding Criterion’s level of precision in generating feedback was triangulated with the data about students’ processing, uptake, and revision as well as their perception of the feedback which might emerge in the stimulated recall interviews.

**Coding students’ revisions from first to second/final drafts.** This part of coding deals with students’ revisions (or lack thereof) made to their first drafts in response to error tags and feedback from Criterion. The coding was conducted on all the students’ practice essays. The following example illustrates how an error is tagged by Criterion:

Student text:  
*As a result, this method of making living can not bring us success and happiness*¹ but receiving our society’s disrespect.

Error message:  
¹You may need to use a comma after this word.

In response to this Criterion error feedback, the student chose to insert a comma after happiness, which is coded as an incorrect revision as the revised form did not resolve the fragment following the coordinating conjunction but. See Appendix E, *Guidelines for Coding Revision Practices*, for details about coding students’ revisions from first to second/final drafts.

**Coding revision practices (i.e., text operations).** Student revision practices in this study were investigated using data from all the essay drafts collected from Criterion practice sessions of the whole experimental group. Coding student revision practices involves analyses of the first and revised drafts of essays written in the practice sessions to see what changes were made to the text and was both theoretically and data driven. Categories from Han and Hyland’s (2015) list of revision practices were adapted for initial trial-coding, which was followed by adjustments to produce a workable list applicable to the actual data. In the revised coding scheme, five categories from Lee and Hegelheimer (2012, as cited in Chapelle et al., 2015, p. 389) on students’ text revision operations in response to ACF were adopted.

In the finalized coding scheme, student revisions were coded on two levels: the kinds of operations done and whether such operations led to correct or incorrect revised text. All the revisions were coded by error types. In the first level of coding, the list of categories for text operations includes:
• **No change:** Students did not make any change to their text. In other words, the original text remained unchanged as the student avoided addressing the identified error or decided to keep the original correct form after consulting the automated feedback which is a false positive.

• **Removal:** Students deleted the part of the essay (phrase, clause, or whole sentence that included the highlighted error) in the revision process.

• **Addition:** Students revised the essay by adding pre/suffixes to word(s) or phrase(s) in response to the error message.

• **Deletion:** Students revised the essay by deleting pre/suffixes from word(s) or phrase(s) in response to the error message.

• **Reordering:** Students revised the essay by reordering certain parts of a clause or swapping the positions of phrases in a sentence.

• **Substitution:** Students revised the essay by changing forms of word(s), phrase(s), or sentence(s) with alternatives.

In the second level, the changes were further coded based on the outcome using the binary coding of (1) correct vs. (2) incorrect revision. The third category, **avoidance**, is comprised of No change or Removal cases above to indicate cases when students refrained from addressing the tagged error and thus, not responding to ACF. Cases where students chose not to make changes to their texts after receiving false positives from **Criterion** were coded under **retention of the correct forms**. As a result, this second tier of coding had four codes:

• **Correct revision (CR):** Instances when students were successful at addressing a tagged error by producing a correct revised form.

• **Incorrect revision (IR):** Instances when students failed to address the tagged error, resulting in an incorrect revised form.

• **Avoidance (AVOID):** Instances when students avoided addressing the tagged error by either making no change to an incorrect structure or removing the part of text that included the tagged error.

• **Retention of correct form (RE):** Instances when students decided not to make any change to a tagged error as it was a false positive generated by **Criterion**. In other words, the originally correct form is retained despite the error code.

Being consistent with the coding approach for verifying error codes, revision outcomes were judged by the result from addressing the error message generated, not by the total result from that revision on the whole phrase or clause, as in:
First draft:  For example, recently I have thought a gift by hand made, it has used to give my boyfriend and I have successed on my complete by hard work but not by money.

Error message:  These two words belong together. Write them as one compound word.

Revised text:  For example, recently I have thought a gift by handmade, it has used to give my boyfriend and I have successed on my complete by hardly work but not by money. [Correct revision]

The revised sentence shows a lot of grammatical problems. Yet, as we are looking at only students’ ability to deal with his/her error on compound words using the correct error code from Criterion, (s)he could fix this single issue and thus satisfy the condition of making a correct revision in response to the automated feedback.

**Coder reliability analysis.** After familiarizing the second coder with the list of codes for (a) revision practices and (b) revision outcomes, the researcher and the second coder both coded a sample of about 15% of the total revision points (i.e., each point is comprised of the first and revised texts in response to a Criterion error tag) taken from practice essays written by the experimental group. Inter-coder reliability statistics were 90% and 93% agreement for (a) and (b), respectively. These figures satisfy the condition for coding of all the revision points to begin.

**Calculation of revision success rate.** The calculation of students’ success rate was based on the number of correct forms elicited by Criterion’s feedback out of the total revisions (correct and incorrect revisions combined) to provide the first picture regarding revision efficacy. The second measure with regards to revision is the rate of successful revisions out of total error feedback points generated by Criterion. In other words, this more conservative measure takes into account non-text operations of no change and removal in its calculation, lowering the success rate compared to that when only actual text operations are counted.

In Chapter 7, *The Multiple-case Study*, with more detailed analyses regarding a focused group of participants, the calculation of students’ revision success rate was based on the number of correct forms elicited by Criterion’s usable feedback messages. In other words, revision success rate was the percentage of successful revisions out of Criterion’s total usable feedback points. Success rate based on the total Criterion’s error feedback messages, including those which misled rather than helped participants arrive at a correct form for their revisions (i.e., ICs and FPs), was also
calculated. These two statistics were compared to see how the nature and quality of ACF influenced students’ revision efficacy.

This was then triangulated with not only the kinds of revisions made but also the verbalized thoughts as the feedback points were being processed and revisions being done.

**Think-aloud protocols.** All the think-aloud protocols were recorded using the free software OBS which helps capture both the students’ verbalizations of thoughts and their on-screen operations as they went through the revisions.

**Transcription.** All of the 15 TAPs collected were transcribed by an assistant using Inqscribe, a transcription software that makes it easier for manual transcription. After the files on Inqscribe were exported to Word format, the researcher double checked the transcripts for accuracy before further analyses took place.

**Coding.** Following Cohen’s (2000) procedures for conducting think aloud protocols and adapting Sachs and Polo’s (2007) two-tiered coding system, the current study attempted to map what happened in students’ think-aloud recordings to revision episodes by quantifying the two processes. Each feedback point on an error in a student’s first draft was coded as noticed (+N) or not noticed (-N) according to what was verbalized in his/her TAP. This was cross-referenced to the student’s revised draft to see if the error was revised (+R) or not revised (-R). The revised error continued to be identified as a successful or unsuccessful revision.

Simplifying Sachs and Polo’s (2007) second tier of coding when the quality of noticing is assessed, this study investigates the level of students’ processing of the noticed feedback points by coding them as substantively or perfunctorily noticed. According to Qi and Lapkin (2001), a feedback point receives substantive noticing if it is noticed (mentioned in the TAP) followed by a reason, while one has only perfunctory noticing if it is noticed but no further reason is provided. This study extends on Qi and Lapkin’s definition of “substantive” and “perfunctory noticing” and incorporates certain descriptions of “extensive vs. limited/no engagement” with advice from Storch and Wigglesworth’s (2010) study to more comprehensively encompass students’ responses to feedback points under these two broad levels of noticing. Briefly, a point is substantively noticed if its mention is accompanied by some cognitive search for metalinguistic knowledge to reflect on the error before correcting it, or when noticing is followed by the mention of some consequence that may occur to the meaning of the text if a change is made; or the mention triggers actions like looking up reference materials for correct forms. On the other hand, a feedback point is
perfunctorily noticed if it is simply mentioned and followed by the student taking up the suggested correction (if available in Criterion feedback) or a very quick change of the form to another form.

These broad concepts in defining students’ processing of feedback points from a cognitive perspective have guided the construction of a Coding Scheme for the TAPs (see Appendix F for Guidelines for Coding Think-aloud Protocols). The TAP transcripts were fed into NVivo 12 for coding and further analyses.

**Coder reliability analysis.** To examine the reliability of the list of codes for coding TAPs, the process began with testing intra-coder reliability which resulted in an agreement rate of 94.9%. With this statistic, inter-coder reliability testing could begin.

After familiarizing the second coder with the codes, the researcher gave the second coder a sample coded text including two transcripts which were the two think-aloud protocols used for intra-rater reliability check. This helped the second coder get a better sense of the codes being used. After the training session, a transcript was prepared for piloting inter-rater reliability. After that, coding went more smoothly and inter-coder reliability was at 90% agreement, satisfying the condition for coding of all TAPs to begin.

**Interviews.** A total of eight interviews were audio recorded for this research, with four stimulated recall and four focus group interviews.

**Transcription.** The recorded stimulated recall interviews and focus group interviews were transcribed by an assistant using Inqscribe. After the files on Inqscribe were exported to Word format, the researcher double checked the transcripts for accuracy before further analyses took place.

**Coding.** The current study adopts a thematic analysis approach to deal with the stimulated recalls as well as focus group interviews. The researcher worked on all the interviews to look for patterns that emerged from these four students’ reflections on their own revisions.

While the contents of the stimulated recall interviews were varied across the four participating students depending on issues that emerged in each student’s verbalized protocols, the four focus group interviews tended to be more structured and consistent across the four groups with two broad themes pre-determined before the interviews took place. Despite this difference, analyses of the interviews were based on an emergent approach (Polio & Friedman, 2017) when the researcher read through the transcripts several times in search of some patterns. After general patterns were noticed, themes were identified. These themes structured the categories to be used for coding. The list of
categories was then generated and trialed on two transcripts in each category of interviews to see if it covered the content of the interviews. After some minor revisions, the finalized list of categories includes:

- **Advantage of ACF**: Students’ mention of a benefit brought about by Criterion ACF
- **Disadvantage of ACF**: Students’ mention of a detrimental effect or a drawback about Criterion ACF
- **Trust in ACF**: Students’ mention of adoption of a usable suggested correction or mention of their belief that Criterion ACF is useful for their learning, or high trust level by quoting a percentage
- **Distrust in ACF**: Students’ mention of their lack of trust in Criterion ACF, refusal to adopt its suggested correction, or low trust level in ACF by quoting a percentage
- **Expectation/Suggestion**: Students’ expression of some expectations from Criterion ACF
- **Comparison with teacher CF**: Instances when learners compared Criterion ACF to teacher corrective feedback, either in favor or against it
- **Other relevant comments**: Any comments not covered in the above categories

After the list of categories was confirmed, all the interviews were fed into NVivo 12 for data coding and subsequent analyses.

In addition to these categories, the stimulated recalls also had extra categories to cover its contents regarding learners’ reflective thinking on their revisions depending on each individual’s verbalizations.

**Coder reliability analysis.** The researcher worked in corporation with a second coder to trial the coding scheme on a focus group interview. After resolution to differences on the coding had been reached, coding on one focus group and one stimulated recall began for checking inter-coder reliability. Agreement was recorded at 89% for double coding of the interviews, which satisfied the intercoder reliability requirement.

**Putting Things Together**

To summarize the ways answers were found for research issues set out in the study, Figure 3.5 presents the inter-relationship between the data sources and each research question.
<table>
<thead>
<tr>
<th>Data sources</th>
<th>Coded categories</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion error messages</strong></td>
<td>Correct error code</td>
<td>ACF nature</td>
</tr>
<tr>
<td></td>
<td>Incorrect error code</td>
<td></td>
</tr>
<tr>
<td></td>
<td>False positive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seeking peer/teacher feedback</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drawing on stored metalinguistic knowledge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Looking up online resources</td>
<td><strong>Substantive noticing</strong></td>
</tr>
<tr>
<td></td>
<td>Translating into L1</td>
<td></td>
</tr>
<tr>
<td><strong>TAP recordings</strong></td>
<td>Getting cues from ACF message to revise</td>
<td><strong>Engagement</strong></td>
</tr>
<tr>
<td></td>
<td>Guessing the correct form</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quickly revising the form using ACF as a red flag</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adopting ACF</td>
<td><strong>Perfunctory noticing</strong></td>
</tr>
<tr>
<td></td>
<td>Noticing without attending to the error</td>
<td></td>
</tr>
<tr>
<td><strong>Comparing first and revised drafts</strong></td>
<td>Addition</td>
<td><strong>Revision</strong></td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reordering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Removal</td>
<td></td>
</tr>
</tbody>
</table>

**Revision operations**

RQ1

RQ2

RQ3
<table>
<thead>
<tr>
<th>Revision outcomes</th>
<th>Revision success rate</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct revision</td>
<td></td>
<td>RQ3</td>
</tr>
<tr>
<td>Incorrect revision</td>
<td></td>
<td>RQ4</td>
</tr>
<tr>
<td>Avoidance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention of correct form</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No revision</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interviews</th>
<th>Perception of ACF</th>
<th>Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantage of ACF</td>
<td></td>
<td>RQ5</td>
</tr>
<tr>
<td>Disadvantage of ACF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trust in ACF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distrust in ACF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectation/Suggestion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison with teacher CF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other comments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3.5. Connecting multiple data sources with the research questions of the study*
Chapter 4

Criterion Automated Corrective Feedback: Its Nature and Accuracy

Introduction

This chapter presents the findings that answer the first two research questions about the nature and accuracy of Criterion automated corrective feedback. Multiple data sources provide answers for each of these research questions. Regarding the nature of Criterion ACF, Criterion-generated error tags and their accompanying metalinguistic explanations are presented first. After this general overview of the feedback, all of the error tags in response to all the practice essays written by the entire group of participants in this study were tabulated by error category. For the second research question, Criterion ACF accuracy is measured using the statistic precision through the verification annotation approach (Leacock et al., 2014) on specific error types as well as the total errors. The chapter ends with a section that focuses on Criterion’s unhelpful feedback which is comprised of its false positives and incorrect error codes. With the subsections above, Chapter 4 serves the dual purposes of familiarizing readers with the kind of feedback Criterion generates and prepares them for Chapter 5 which zooms in on learners’ interaction with such feedback.

Research Question 1. What is the nature of Criterion automated corrective feedback?

Findings for this research question will be presented in two subsections. Firstly, Criterion error messages (or metalinguistic explanations) for different error types in the three broad categories of grammar, usage, and mechanics are provided in Tables 4.1, 4.2, and 4.3. Included in each table are the names of error types and the metalinguistic explanations that follow the error tags. Based on the judgement of the explicitness of metalinguistic explanation(s), each error type is categorized into either generic or specific feedback. Data-driven examples are also provided to illustrate each error type as well as the metalinguistic explanation.

The second subsection presents the numbers of Criterion error tags generated in response to all the learners’ practice essays in the current study. The error tags are reported in bar charts for each category of grammar, usage, and mechanics, which are then further broken down into specific error types.

Error messages generated in response to student texts. Alongside the error tags on errors in student essays, Criterion generated formulaic metalinguistic statements to guide student corrections. For each error type, one single statement is produced. For instance, with all the errors receiving a Proofread this! tag, one formulaic message is generated:
This part of the sentence contains an error or misspelled word that makes it hard to understand what you mean.

Twenty-five of the 29 error types with Criterion error flags in the current study received one formulaic metalinguistic explanation. Each of the remaining four error types, Missing or Extra Articles, Possessive Errors, Preposition Errors, and Confused Words, had more than one metalinguistic explanation. More specifically, errors flagged with the Missing or Extra Articles tag received one of the four messages below:

a. You may need to use an article before this word.
b. You may not need to use an article here.
c. You may need to use an article before this word. Consider using the article a.
d. You may need to use an article here. Consider using the article the.

Similarly, depending on the specific errors, one of two metalinguistic explanations will be generated by Criterion for errors receiving the Possessive Errors, Preposition Errors, and Confused Words tags. These metalinguistic messages can be found in Tables 4.1, 4.2, and 4.3 which present the error tags in each category of Grammar, Usage, and Mechanics in greater details. Included in these tables are error types, metalinguistic explanations that students receive in addition to the error tag, as well as specific examples drawn from the data for illustration. An additional piece of information in the tables relates to the level of explicitness of the error explanations to be found below each error type name. Using Ranalli, Link, and Chukharev-Hudilainen’s (2017) categorization of error messages, the current study makes the distinction between generic and specific metalinguistic explanations which gives us a broad overview of how explicit or implicit the feedback generated by Criterion is. In this distinction, generic feedback indicates messages that are identical in response to an error type without reference to any specific word or phrase in student writing. Such feedback tells students what and where the error is. On the other hand, specific feedback consists of some component of students’ text in its wording, either in the form of:

- a recommendation to use a word instead of another word used by the student.
- a specific suggested text operation with reference to a highlighted portion in students’ texts.

With such information, specific feedback answers both the what and where questions about the error, then adds the information regarding how to revise it.
**Grammar errors.** Grammatical feedback on *Criterion* was mainly generic with seven out of nine grammatical error types receiving formulaic generic feedback. The system generally stated that something was wrong about the verb or the sentence structure through indicating the location of the errors. This was done by highlighting the relevant words, phrases or whole clauses in sentence structure errors. Two exceptions, *possessive errors* and *wrong/missing word* error tags, are followed by specific metalinguistic explanations which gave students a suggestion to add or remove the apostrophe (the former) and to replace the word with another word (the latter), making the revision task seemingly easier to implement for students.
Table 4.1
Criterion ACF Feedback on Grammar Errors with Cited Examples from This Study

<table>
<thead>
<tr>
<th>Error type</th>
<th>Explicitness</th>
<th>Metalinguistic explanation(s)</th>
<th>Example(s)</th>
</tr>
</thead>
</table>
| Proofread this!    | Generic      | This part of the sentence contains an error or misspelled word that makes it hard to understand what you mean. | - As a result, it creates for you 'ability' to deal with difficulties in your work.  
- She 'study' college of tourism.                                                                                                           |
| Subject-Verb       | Generic      | The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb. | - Nowadays the economy is growing so it 'give' us a lot of choices of jobs.  
- There 'are' some advantage of change job.                                                                                                   |
| Agreement          |              |                                                                                               |                                                                                                                                                                                                           |
| Ill-formed verbs   | Generic      | You may have used the wrong form of this verb. Match the subject to the verb to decide whether you have used the verb correctly. | - Finally, find a new job brings people a novel routine, they have dissimilar tasks and events that they 'may' never 'met' in advance, it creates for them a meaning of life that they desire and they have 'try' their best to achieve. |
| Possessive errors  | Specific     | (1) You may need to use an apostrophe to show that someone owns something.                     | - Many 'people' choice some jobs for them, then they just do a jobs that they feel comfortable.                                                                                                           |
|                    |              | (2) You may need to take out the apostrophe to make this word a plural noun.                   | - Besides, you also have to find a person you love to get happy from 8p.m to 6a.m next day'.                                                                                                          |
| Run-on sentences | Generic | This sentence may be a run-on sentence. You may need to add a linking word, add punctuation, or make two separate sentences. | - It means that, the more you change your job, the more experience you get from your previous jobs.¹  
- The last solution, individual must to have knowledge about the pollution water in this local and each person should be prevent from another things which is danger to the river to sum up, self awareness of each people is best solution to reduce this pollution because it decrease many activity polluted the river.¹  
- After graduating with a high mark.¹ She decided to apply for a foreign company that had headquarters in Viet Nam…  
- Because it also have benefits same to the changing job.¹ |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragments</td>
<td>Generic</td>
<td>This sentence may be a fragment. Proofread the sentence to be sure that it has at least one independent clause with a complete subject and predicate.</td>
<td>- Personally, I would like change my work at least one times eventhough must to have another difficults.¹</td>
</tr>
<tr>
<td>Garbled sentences</td>
<td>Generic</td>
<td>This sentence is confusing because it contains several grammatical or spelling errors.</td>
<td>- They kill the¹ endanger animals to take the parts of their body.</td>
</tr>
<tr>
<td>Wrong or missing word</td>
<td>Specific</td>
<td>Did you type the instead of they, or have you left out a word? You have a spelling or typing mistake that makes this sentence seem to have a wrong or missing word.</td>
<td>- They kill the¹ endanger animals to take the parts of their body.</td>
</tr>
<tr>
<td>Pronoun errors</td>
<td>Generic</td>
<td>You may have used the wrong pronoun.</td>
<td>If a worker switched jobs every couple of years, it may be difficult for they¹ to form strong relationship that endure after they stop working.</td>
</tr>
</tbody>
</table>
**Mechanics errors.** Unlike its feedback on grammatical errors, *Criterion*’s feedback to mechanics errors in students writing was specific with the only exception of spelling. All other mechanics error types receive suggested text operations for corrections such as deletion or addition of punctuation marks (e.g. missing final punctuation, missing comma, extra comma) or rewriting of certain words pointed out (e.g. capitalize proper nouns, hyphen error, and compound words). Table 4.2 provides more detailed information for each error type in this category.
<table>
<thead>
<tr>
<th>Error type</th>
<th>Explicitness</th>
<th>Metalinguistic explanation(s)</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing final punctuation</td>
<td>Specific</td>
<td>You are missing a punctuation mark at the end of this sentence.</td>
<td>- In conclusion, people is successful or being not successful, i think it depends on the way of their life.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Spelling</strong></td>
<td>Generic</td>
<td>This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.</td>
<td>- Success does not just depend on the wealth but happiness and contentment.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Missing comma</strong></td>
<td>Specific</td>
<td>You may need to use a comma after this word.</td>
<td>- Nowadays the economy is growing so it give us a lot of choices of jobs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Capitalize Proper Nouns</strong></td>
<td>Specific</td>
<td>If this word is a proper noun, you need to use a capital letter.</td>
<td>- I prefer to stay in the same job because need have stability in my life, don't like to change in all things of life.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hyphen error</strong></td>
<td>Specific</td>
<td>You may need to add a hyphen between these two words.</td>
<td>- If the employee wants to get close to the hometown or to a metropolitan or for any reason, he may seek lacation change with all the other above mentioned benefits.</td>
</tr>
</tbody>
</table>
| **Missing initial capital letter in a sentence** | Specific | Remember to capitalize the first word of each sentence. | - Nowadays, many people in a lot of orbits such as technology, medical, science… try to bring convenient and peaceful for people.¹  
- for example, my sister is a doctor with a pretty high salary.¹ |
| **Extra comma** | Specific | You may need to remove this comma. | A worker who constantly switches jobs may have difficulty getting jobs in the future¹ because they will have to train for a long time before working officially. |
| **Fused words** | Specific | This word is not a compound word. Write it as two words. | In brief, noone¹ can help you to choose which career is suitable for you. |
| **Duplicates** | Specific | (1) You have typed two **identical words** in a row. You may need to take out one of them.  
(2) You have typed two **articles** in a row. You may need to take out one of them.  
(3) You have typed two **pronouns** in a row. You may need to take out one of them. | As the world is changing and¹ and² developing day by day, expectations for jobs of people are also different.  
Besides that, changing jobs or move an¹ another² city to work that will help me know more about that place and search more culture here.  
Turn off your faucets when you don’t using water, such as when you brushing my¹ your² teeth. |
<table>
<thead>
<tr>
<th>Missing Question Mark</th>
<th>Specific</th>
<th>Remember to use a question mark at the end of a sentence that is a question.</th>
<th><strong>What's more amazing.</strong> 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compound Words</strong></td>
<td>Specific</td>
<td>These two words belong together. Write them as one compound word.</td>
<td>To sum up, I believe we need some 1 times 1 change our jobs because of our soul.</td>
</tr>
</tbody>
</table>

1. The numbers in the text indicate the position of the text in the document.
**Usage errors.** Usage errors received mostly specific feedback from *Criterion*, with a mixture of precise indication of the error locations (e.g. wrong form of word, confused words, prepositions) and suggested text operations to deal with the errors found (e.g. article errors, faulty comparisons). The two cases of generic feedback were determiner-noun agreement and negation errors. In response to determiner-noun agreement errors, *Criterion* generally asked students to double check the noun and the determiner that modified it without giving further suggestions on how to correct the error or what exact noun or determiner to examine. Similarly, for negation errors, *Criterion’s* metalinguistic explanation simply indicated that the sentence contained an incorrect use of a double negative and left it to the students to figure out the way to revise. Table 4.3 provides more details regarding each error type with exemplary flagged errors from the student essays.
<table>
<thead>
<tr>
<th>Error type</th>
<th>Explicitness</th>
<th>Metalinguistic explanation(s)</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing or extra article</td>
<td>Specific</td>
<td>(1) You may need to use an article before this word.</td>
<td>- The last one is building reputation(^1) of dependability.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) You may not need to use an article here.</td>
<td>- In my opinion, both changing or staying the(^1) jobs have some advantages.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) You may need to use an article before this word. Consider using the article a.</td>
<td>- Third, changing careers helps some people gain expertise in a new area and can make they suitable for wide(^1) range of job opportunities in the future.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) You may need to use an article here. Consider using the article the.</td>
<td>- Same(^1) work throughout the career is tedious to many people.</td>
</tr>
<tr>
<td>Determiner Noun Agreement</td>
<td>Generic</td>
<td>You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.</td>
<td>- However, there are several(^1) advantage of stay. Firstly, one(^1) of the biggest advantage of stay, you have ability to promote location from small location to large location. - Another(^1) benefits of staying at the same job for a long time is that you will have strong work relationship.</td>
</tr>
<tr>
<td>Preposition error</td>
<td>Generic</td>
<td>(1) You may be using the wrong preposition. Generic</td>
<td>- They have passion on(^1) the certain job so they can do it well.</td>
</tr>
<tr>
<td><strong>Confused words</strong></td>
<td>Specific</td>
<td>(2) You may need to remove this preposition. Specific</td>
<td>It is not all of people who are a talent in the first time they have gone to work.</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>----------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
|                     | Specific | (1) You have used … in this sentence. You may need to use … instead. **Note:** The error message is adapted to the specific error being tagged by referring to words from learner texts. See the two examples in the next column for details. | - In addition, you remarkably increase your earning power in another company which appreciates your ability and strengths. **Error message:** *You have used in this sentence [sic]. You may need to use you’re instead.*  
- Both of these choices have its own advantages  
**Error message:** *You have used it’s in this sentence. You may need to use its instead.* |
|                     | Specific | (2) You have misspelled a word near the word that makes look [sic] like a confused-word error. | - It suggests you an opportunity to get in a completely new environment in which you can get rid of difficulties and start everything again.  
- The new social relationship which might help to escape some negative feelings as spending too much time at the boring workplace that would be counted. |
<p>| <strong>Wrong form of word</strong> | Specific | You may have used the wrong form of this word. | Firstly, I think we should walk, ride our bike whenever possible or use public transportations. |</p>
<table>
<thead>
<tr>
<th>Issue</th>
<th>Specificity</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong part of speech</td>
<td>Specific</td>
<td>You may need to use a verb instead of a noun. <em>This same message is generated irrespective of the actual part of speech of the word being flagged.</em></td>
<td>- The <strong>finally</strong> advantages from changing the jobs is reducing boredom.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Beside, the government should have <strong>to</strong> emphasis on cleaning energy resources.</td>
</tr>
<tr>
<td>Wrong article</td>
<td>Specific</td>
<td>(1) You have used a. You may need to use the instead.</td>
<td>- To be concluded, staying in a <strong>same</strong> job or changing different jobs has it own advantages.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) You have used the. You may need to use a instead.</td>
<td>- They have passion on the <strong>certain</strong> job so they can do it well.</td>
</tr>
<tr>
<td>Faulty comparisons</td>
<td>Specific</td>
<td>You cannot use these two words together to make a comparison. You need to take out &quot;more&quot; or “most.”</td>
<td>- Our lifes has become <strong>more ‘worse’</strong> by worse due to many kind of pollution.</td>
</tr>
<tr>
<td>Negation error</td>
<td>Generic</td>
<td>Your sentence contains an incorrect use of a double negative.</td>
<td>It can’t not be denied that there are advantages of changing job or not so that people keep doing what they prefer.</td>
</tr>
</tbody>
</table>

**Correction:**

- The **finally** advantages from changing the jobs is reducing boredom.
- Beside, the government should have **to** emphasis on cleaning energy resources.
- To be concluded, staying in a **same** job or changing different jobs has it own advantages.
- They have passion on the **certain** job so they can do it well.
- Our lifes has become **more ‘worse’** by worse due to many kind of pollution.
**Error tags on different categories and error types.** The number of error tags received by all of the participating students will be presented in this section. Error flags for each category of grammar, usage, and mechanics will be shown in three separate Figures 4.1, 4.2, and 4.3, respectively. Despite the fact that the version of *Criterion* used for the current study could generate feedback on a total of 31 error types, over the whole duration of the research, students received feedback on 29 out of these 31, including nine error types for Grammar, nine for Usage, and eleven for Mechanics. The two error types of *Nonstandard Word Form* (Usage) and *Missing Apostrophe* (Mechanics) did not record any errors committed by the participating students, probably due to the very low chance of use related to such aspects in student writing.

Figure 4.1 summarizes the numbers of *Criterion*-generated error tags in response to students’ first and revised drafts in three practice sessions. Subject-verb agreement was the most frequent error type among all the grammatical errors detected by *Criterion*, followed by ill-formed verbs. The least frequent error types were wrong/missing word and pronoun errors which only had 7 and 11 errors, respectively.

![Figure 4.1. Numbers of *Criterion*-generated error tags by individual error type in grammar.](image)

Figure 4.2 below provides details about the error tags for the second category, usage. In this category, missing or extra article was the most common issue for students in this study with more than 1500 errors detected. With 322, 213, and 200 counts respectively, determiner-noun agreement, preposition, and confused words were the second, third, and fourth most commonly tagged usage errors. All other types of usage errors had much lower numbers of error tags from *Criterion*. 
Figure 4.2. Numbers of Criterion-generated error tags by individual error type in usage.

Regarding the third category, Figure 4.3 summarizes the numbers of error tags provided by Criterion in response to students’ errors in mechanics. Spelling was by far the most common error with 1656 counts, followed by missing comma as the second most frequently found for participant students according to Criterion’s feedback. Missing initial capital letter in a sentence, probably due to the new composing mode of typing, was also very common at 272 counts.

Figure 4.3. Numbers of Criterion-generated error tags by individual error type in mechanics.
Summary of findings for RQ1. To answer the first research question, *Criterion* error tags on each error type of the three broad categories of grammar, usage, and mechanics were tabulated. The metalinguistic explanations for 29 error types that students in the current study received *Criterion*’s ACF on were also presented and analyzed for their level of explicitness.

Figure 4.4 provides a visual illustration of error distribution among these three broad categories of grammar, usage, and mechanics. The largest proportion of errors detected by *Criterion* belongs to mechanics (48%), followed by usage errors (42%), while the rate of grammar errors which were flagged was much lower, at 10% of the total.

![Figure 4.4. Error distribution among grammar, usage, and mechanics.](image)

In addition, error tags on the ten most frequently flagged error types across the three broad categories are summarized in Table 4.4 below. The results showed that *spelling* (for mechanics) and *missing or extra articles* (for usage) were by far the most frequently recurring error types among the learners.

Turning to the explicitness of the metalinguistic explanations for the errors, seven of the nine grammatical error types were provided with generic feedback, while the reverse was true for error types in the mechanics category with its ten out of eleven types receiving specific feedback. Similar to grammar errors, usage error types were mostly accompanied by generic feedback, and only two of its total nine received specific feedback from *Criterion*. 
Table 4.4

**Error Types with Most Frequent Occurrences in Participating Students’ Essays**

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Error Type</th>
<th>Total Criterion Error Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mechanics</td>
<td>Spelling</td>
<td>1656</td>
</tr>
<tr>
<td>2</td>
<td>Usage</td>
<td>Missing or Extra Article</td>
<td>1539</td>
</tr>
<tr>
<td>3</td>
<td>Grammar</td>
<td>Subject-Verb Agreement</td>
<td>442</td>
</tr>
<tr>
<td>4</td>
<td>Mechanics</td>
<td>Missing Comma</td>
<td>421</td>
</tr>
<tr>
<td>5</td>
<td>Usage</td>
<td>Determiner Noun Agreement</td>
<td>322</td>
</tr>
<tr>
<td>6</td>
<td>Mechanics</td>
<td>Missing Initial Capital Letter in a Sentence</td>
<td>272</td>
</tr>
<tr>
<td>7</td>
<td>Grammar</td>
<td>Ill-formed Verbs</td>
<td>216</td>
</tr>
<tr>
<td>8</td>
<td>Usage</td>
<td>Preposition Error</td>
<td>213</td>
</tr>
<tr>
<td>9</td>
<td>Usage</td>
<td>Confused Words</td>
<td>200</td>
</tr>
<tr>
<td>10</td>
<td>Mechanics</td>
<td>Extra Comma</td>
<td>166</td>
</tr>
</tbody>
</table>
Research Question 2. How accurate is Criterion’s automated corrective feedback?

To answer the second research question about the accuracy of Criterion ACF, all Criterion error tags generated for the first drafts of participating students’ practice essays were verified following the Verification Approach used by Gamon et al. (2008, as cited in Leacock et al., 2014). The choice of the verification approach was suitable for the purpose of the current study. It only allows estimation of precision while ignoring the recall statistic. This means that the verification approach only determines whether the error tags generated by the system are correct or incorrect while not including the estimation of the number of errors which have not been detected by the system. With the focus of the study on learners’ engagement with and subsequent revisions in response to the feedback they received from Criterion, the question of whether Criterion has missed certain errors became irrelevant.

The accuracy of Criterion ACF in this study was examined using three categories of correct error code, incorrect error code, and false positive. In addition, a more detailed examination of the incorrect error codes and false positives from Criterion (to be grouped under the broad category of unhelpful feedback later on) provides further information for this question. Therefore, the following presentation of findings on Criterion ACF’s accuracy will be divided into two parts: (1) Findings about the accuracy of Criterion ACF based on the verification of total error codes and verification of error codes in each error type, and (2) detailed analysis of the unhelpful feedback from Criterion.

**Criterion ACF accuracy.** A total of 2787 Criterion error tags were extracted from all of the practice essays submitted to Criterion by the 43 students in the experimental group. The verification of these error tags using the three coding categories above revealed that Criterion was satisfactorily precise overall in detecting errors for learners to attend to issues in their essays. Table 4.5 presents the numbers of error tags being coded as correct, incorrect, or false positive, with the corresponding percentages below each number.

<table>
<thead>
<tr>
<th>Correct Error Code</th>
<th>Incorrect Error Code</th>
<th>False positive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2291 (82.2%)</td>
<td>171 (6.1%)</td>
<td>325 (11.7%)</td>
<td>2787 (100%)</td>
</tr>
</tbody>
</table>
As seen from Table 4.5, about 82 percent of all error messages were correct tags in response to errors in students’ essays, while the remaining 18 percent were either false positives (11.7%) or incorrect error codes (6.1%) that did not help students correctly identify the errors they made, thus not enabling effective revisions.

For more details, Table 4.6 presents the precision of Criterion ACF for specific error types. Some error types were merged due to their common nature and the small number of errors found in each type. For example, Missing initial capital letter in a sentence and Capitalize proper nouns were collapsed under Capitalization. In this table, the error types are ordered from those with the highest to the lowest precision rates.

Table 4.6
Verification of Criterion ACF for Specific Error Types (N = 2787)

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Number of error codes</th>
<th>Correct error codes (%)</th>
<th>Incorrect error codes (%)</th>
<th>False positives (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicates</td>
<td>4</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faulty comparisons</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garbled sentences</td>
<td>20</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negation errors</td>
<td>3</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong article</td>
<td>22</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proofread this</td>
<td>59</td>
<td>98.3</td>
<td></td>
<td>1.7</td>
</tr>
<tr>
<td>Spelling</td>
<td>600</td>
<td>97.4</td>
<td></td>
<td>2.56</td>
</tr>
<tr>
<td>Determiner noun agreement</td>
<td>167</td>
<td>97</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Subject-verb agreement</td>
<td>264</td>
<td>91.7</td>
<td>5.3</td>
<td>3.0</td>
</tr>
<tr>
<td>Compound words</td>
<td>10</td>
<td>90</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Capitalization</td>
<td>176</td>
<td>88.6</td>
<td>10.8</td>
<td>0.6</td>
</tr>
<tr>
<td>Missing final punctuation</td>
<td>23</td>
<td>87</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Pronoun errors</td>
<td>7</td>
<td>85.7</td>
<td>14.3</td>
<td></td>
</tr>
<tr>
<td>Ill-formed verbs</td>
<td>120</td>
<td>83.3</td>
<td>15</td>
<td>1.7</td>
</tr>
<tr>
<td>Run-on sentences</td>
<td>6</td>
<td>83.3</td>
<td></td>
<td>16.7</td>
</tr>
</tbody>
</table>
Table 4.6 shows that Criterion was most accurate in detecting duplicates, faulty comparisons, garbled sentences, negation errors, and wrong articles where 100% of all flagged errors were correct identifications by the system. One thing worth noting is that for all these error types, there were quite small numbers of error messages generated, with wrong articles having 22 tags as the highest. Spelling, Subject-verb agreement, and Determiner-noun agreement, with large numbers of errors detected, were among the group of error types that Criterion was very good at detecting (at 90% or more precision). Proofread this!, a more general way of dealing with clauses/sentences with more than one error that makes it hard for the system to give a specific tag, was understandably among this group with almost 98% precision.

Taking the widely accepted criterion of 80% precision for being a useful system (Quinlan, Higgins, & Wolff, 2009; Ramineni & Deane, 2017), Criterion was below expectations in terms of fragments, compound words, preposition errors, extra comma, hyphen errors, and particularly well below 50% precise with possessive errors, missing comma, and missing question mark errors.

Unhelpful feedback. The unhelpful error messages provided by Criterion come in the forms of false positives and incorrect error codes presented in Tables 4.5 and 4.6, as well as some systematic faults in some of its error explanations. Specifically, both the error codes and the metalinguistic explanations that accompany them were examined to initially discover why they may be of little value or even detrimental to students’ revisions.
**False positives.** As earlier defined in Chapter 3, an error tag is considered a false positive when *Criterion* incorrectly flags a correct structure (word, phrase, or clause) and punctuation in students’ essays as erroneous, as in:

Student text:  
*In addition, you remarkably increase your earning power in another company which appreciates your ability and strength.*

Error message: *You have used your in this sentence. You may need to use you’re instead.*

In the current study, 11.7% of the total error tags generated by *Criterion* were false positives. Regarding specific error types, five out of seven error tags on missing question marks were false positives. A frequent source of false positives in this error type comes from *Criterion*’s failure to recognize the adverbial phrase *what’s more* which it mistook for a question starting with a Wh-word. For example:

Student text: *What’s more, you can have the opportunity to pursue bigger challenges.*

Error message: *Remember to use a question mark at the end of a sentence that is a question.*

Spelling is the error type with the largest number of false positives, with more than 150 cases. Most of the words falsely tagged as spelling errors were proper nouns which *Criterion* did not recognize, as in:

Student text: *After a long time changing jobs, she has identified her own passion which is trading goods made from “Moringa oleifera”.*

Error message: *This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.*

Next on the list is missing comma with the second highest number of false positives, at 124 cases. Commonly found were *Criterion*’s wrong suggestion of inserting a comma before a restrictive relative clause, such as the following example:

Student text: *They have to do works they don’t have passion for.*
Error message: You may need to use a comma after this word.

Similarly, extra comma errors had 16 of the total 70 being false positives. The example below illustrates a Criterion false positive when it suggested deleting the comma before money or your health.

Student text: Think twice then decide, which is much more significant money or your health?
Error message: You may need to remove this comma.

Other error types which had about 15 to 20 percent of their total counts being false positives include preposition, possessive, compound words, run-on sentences, fragments, and pronoun errors. Feedback on subject-verb agreement, ill-formed verbs, proofread this, and capitalization had minimal rates of false positives, all with below 3 percent of the total error tags in each category.

Also falling under the category of false positives are redundant error messages. An error tag is considered redundant when an error code is correct but not necessary because the student’s text is grammatically/mechanically acceptable without the need to revise at all, as in the example below:

Student text: If so, their aims in life must be the aims of getting houses, cars and every facilities for their own or for their families.
Error message: You may need to use a comma after this word.

Judged by strict academic writing rules, the insertion of a comma before “and” in a list is needed. However, such a strict rule does not hold in the current fluid state of writing conventions (e.g., see The Blue Book of Grammar and Punctuation by Strause, Kaufman, & Stern, 2014, 11ed.). In addition, a large number of redundant tags in the missing comma error type were due to the system’s prescriptive treatment of comma usage that asks for a comma between clauses connected by coordinating conjunctions like but or and, as in the next sentence:

Student text: I consider that changing job consists of potential risks but it has some advantages.
Error message: You may need to use a comma after this word.
**Incorrect feedback.** As mentioned in Chapter 3, a *Criterion* feedback point is incorrect when the error message either mislabels the error type or misdiagnoses the nature of the error. In the vast majority of coded incorrect error tags, despite correct identification of an erroneous structure in the student’s essay, *Criterion* both mislabeled and offered a misleading message about the error being highlighted. The following is an example when *Criterion* appropriately coded a structure as incorrect but gave it a wrong code:

Student text:  
*In term of keeping the current job, it might have two major strengths to support this trend.*

Error message:  
*You may need to use an article before this word. Consider using the article the.*

In the above example, *Criterion*’s highlighting of the word *term* correctly draws the student’s attention to an issue in her writing. However, the omission of ‘s’ in the word ‘term’ can be considered a spelling error rather than what is explained in *Criterion*’s error code, an article error. The suggested correction offered by *Criterion* was even worse as its adoption would result in a more severely incorrect structure. Instead of helping students correct the problem by a simple insertion of ‘s’ after ‘term’, things were made confusing with the automated feedback point.

In the second example below, *Criterion* was precise in categorizing an error, but its detection failed to get to the bottom of the problem, as removal of the comma after *choose* was just the first step towards correcting this run-on sentence.

Student text:  
*People often believe that a career is a life choose some people have choose a profession they will keep doing the same job all throughout their life.*

Error message:  
*You may need to remove this comma.*

Of the total error tags generated by *Criterion* to participating students, 6.1% belonged to this category of incorrect error codes. Error types with the largest numbers of incorrect error codes include missing or extra article (56), capitalization (19), ill-formed verbs (18), extra-comma (17), subject-agreement (14), missing comma (10), confused words (8), and preposition errors (8).
**Systematic faults in Criterion error tagging.** Beside the feedback points specifically generated by Criterion to individual errors found in learners’ essays, there were inherent issues in two metalinguistic explanations from the system that can compromise the usefulness of Criterion ACF despite its correct identification of errors of two types below.

**Confused words error tag.** In response to confused words errors, one of the two Criterion’s automated metalinguistic explanations is that “You have misspelled a word near the word that makes look [sic] like a confused-word error”. The explanation is problematic in two ways. Firstly, it is grammatically faulty as there is no object for the verb “makes”. Secondly, the expression in the message telling the writer that she has “misspelled a word near the word” was ambiguous as to which word is the real error. The example student sentence below received this error tag and explanation.

Student text:  *The new social relationship which might help to escape some negative feelings as spending to much time at the boring workplace that would be counted.*

The word “to” was highlighted by Criterion. If following the error tag type of “confused words”, we tend to think that “to” has been confused with “too” and the latter seems the plausible alternative. Yet, if reading Criterion’s metalinguistic explanation, it seemed that the system wanted to draw the learner’s attention to the word adjacent to “to”, which is either “spending” (preceding word) or “much” (succeeding word). However, such interpretation is not helpful for a correct revised form. Generally, the grammatical error together with an ambiguous expression in this particular metalinguistic explanation makes the feedback less useful than expected.

**Wrong part of speech.** The second systematic issue with Criterion error tagging relates to its metalinguistic explanation for Wrong Part of Speech errors, which is “You may need to use a verb instead of a noun”. Irrespective of the actual part of speech of the word being flagged across different errors in this error type, this exact same message is generated, which potentially misleads students’ revisions. For example,

Student text:  *The finally advantages from changing the jobs is reducing boredom.*
In the above sentence, a helpful metalinguistic explanation would be “You may need to use an adjective instead of an adverb” instead of the one-size-fit-all message “You may need to use a verb instead of a noun”. However, Criterion was not able to customize its error explanation to specific errors in this particular type, making it potentially confusing for learners during their revisions.

Summary of findings for RQ2. The findings show that Criterion performed differently across different error types. Overall, the automated corrective feedback was satisfactorily accurate with 21 error types exceeding the 80% precision threshold. Below this cut score was feedback on the eight error types of fragments, compound words, preposition errors, extra comma, hyphen errors, possessive errors, missing comma, and missing question marks. In addition to the proportion of correct error codes, this section also provides the rates of incorrect error codes and false positives for a more complete picture of Criterion ACF. The largest numbers of false positives were found in mechanics error types of spelling and comma usage, while the grammatical errors of subject-verb agreement, ill-formed verbs, and ‘proofread this’ had few false positives. Incorrect error codes were also found common for mechanics errors including capitalization, extra and missing comma, as well as the two grammatical error types of ill-formed verbs and subject-verb agreement.
Chapter 5
Learner-centric Perspectives on Criterion ACF: Learners’ Engagement, Revisions, and Perceptions

Introduction

Departing from the exclusive focus on the characteristics of Criterion automated corrective feedback, this chapter presents findings related to learner-centric issues including students’ engagement, perception, and revision practices using the feedback, as well as observed accuracy changes over the studied period, which are the focus of the next three research questions of the study. Research Question 3 focuses on students’ engagement (RQ3.1) and subsequent revisions (RQ3.2) in response to Criterion ACF. Research Question 4 is about the acquisition of specific linguistic morphemes as well as overall accuracy change patterns among the participants, while Research Question 5 presents the learners’ perceptions of Criterion automated corrective feedback.

Multiple data sources were analysed to provide answers for each of these research questions. Insights into students’ engagement with automated corrective feedback came primarily from the think-aloud protocols recorded for 14 learners when they were revising essays using Criterion ACF and their follow-up stimulated recall interviews. Data to answer RQ3.2 about student revisions were the comparison of first and revised drafts written by all of the learners, backed up by relevant comments and reflections during the interviews. Findings for RQ4 about accuracy changes came from the overall accuracy scores recorded for students’ pre, post, and delayed post-test essays as well as accuracy scores for the five specific morphemes of article, third person singular, copula, plural, and comma usage. Finally, students’ perceptions of Criterion ACF were based on thematic analyses of stimulated recalls and focus group interviews.

Research Question 3.1. How do students engage with Criterion ACF?

Findings for this research question are presented in the following order. First, students’ cognitive engagement with Criterion ACF is provided through the analyses of their online verbal TAPs. This is triangulated with the TAP students’ comments on their engagement with the feedback during their stimulated recall interviews, as well as reported cognitive strategies used to process the feedback identified by students in the focus group interviews. Finally, more in-depth analyses of the learners’ engagement with unhelpful feedback are provided.
**Students’ cognitive engagement with ACF.** Analyses based on 14 students’ think-aloud protocols show that all of the 270 error tags were noticed by students. Their “noticing” of the feedback, however, suggests more limited than extensive engagement episodes with *Criterion ACF*. Of the total 270 noticing episodes coded, 183 were perfunctory (68%) while the remaining 87 (32%) were substantive. In other words, more than two thirds of students’ engagement episodes with the feedback generated by *Criterion ACF* were at surface levels demonstrated through perfunctory strategies in dealing with the errors made. Table 5.1 presents further details regarding engagement activities conducted by students as they were processing *Criterion ACF*.

Table 5.1
**Student Engagement Activities through TAP Analyses**

<table>
<thead>
<tr>
<th>Engagement level</th>
<th>Engagement activity</th>
<th>Total coded episodes</th>
<th>Coded episodes in percentages</th>
<th>No. of students involved*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perfunctory noticing</strong></td>
<td>Quickly revising the form using ACF as a red flag</td>
<td>98</td>
<td>36%</td>
<td>14</td>
</tr>
<tr>
<td>183</td>
<td>Notice without attending to the error</td>
<td>35</td>
<td>13%</td>
<td>11</td>
</tr>
<tr>
<td>(68%)</td>
<td>Guessing the correct form</td>
<td>27</td>
<td>10%</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Quickly adopting ACF</td>
<td>23</td>
<td>8%</td>
<td>9</td>
</tr>
<tr>
<td><strong>Substantive noticing</strong></td>
<td>Looking up online resources</td>
<td>37</td>
<td>14%</td>
<td>9</td>
</tr>
<tr>
<td>87</td>
<td>Drawing on stored metalinguistic knowledge</td>
<td>28</td>
<td>10%</td>
<td>10</td>
</tr>
<tr>
<td>(32%)</td>
<td>Getting cues from <em>Criterion</em> error message to revise</td>
<td>10</td>
<td>4%</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Translating into L1</td>
<td>10</td>
<td>4%</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Seeking peer/teacher feedback</td>
<td>2</td>
<td>1%</td>
<td>2</td>
</tr>
</tbody>
</table>

*The total TAP students were 14, and each of them had different engagement activities when processing ACF*

As can be seen from Table 5.1, more than one third of revision episodes were quickly executed using *Criterion* highlighted error flags as a signal. This means that students did not
need to go further than looking at the highlighted words to realize they had made a mistake, implicating the surface-level nature of the errors pointed out. In engagement episodes of *quickly revising the form using ACF as a red flag*, students did not read or refer to *Criterion* metalinguistic explanations either. Other forms of limited engagement with ACF, *noticing without attending to the error, guessing the correct form* and *quick adoption of ACF suggested corrections provided in Criterion metalinguistic explanations*, were also commonly found instances in the data.

On the other hand, the most frequently used activity that represents extensive engagement with ACF was *looking up online resources* (37). This is an expected result due to the fact that for students working on *Criterion*, a web-based learning program itself, other online resources were easily accessible by simply pulling up another webpage on the screen. The second most common engagement strategy with 28 coded episodes was *drawing on stored metalinguistic knowledge*. During these episodes, engagement with *Criterion* error feedback positively triggered learners’ stored knowledge as they searched for a solution to flagged errors, as in the following example from Quan, one of the TAP students in this research:

| Student text: | But the problem is, how each person define the word “success” for their own. |
| Criterion message: | The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb. |
| TAP excerpt: | Subject-verb agreement. What’s the problem? Oh. *Is said that only people who earn a lot of money are successful. But the problem is, how each person define the word “success” for their own. I meant “những vấn đề ở đây là làm sao mỗi người có thể định nghĩa được từ thành công cho chính họ” [Translation of his writing into Vietnamese]. For each person. How each person. Each person. Define. Why? Ahh, how each person, each person is a third person singular. Define should have an “s”. OK. I will add “s”. Each person is a third person singular pronoun, so there is an “s” after the verb that follows. |

Due to *Criterion*’s error tag and its accompanying metalinguistic message, the student’s attention was drawn to a subject-verb agreement error in his text. He initially had difficulty...
seeing the problem that needed fixing. However, after more careful consideration of the error code and reference to his previously acquired knowledge about subject-verb agreement for third person singular pronouns, he finally realized the error he made and corrected it successfully.

Turning to the number of students involved in the range of engagement activities, all the 14 TAP students were perfunctorily engaged with Criterion ACF with varying numbers of limited engagement episodes, ranging from five up to 32 episodes per student. Limited engagement was demonstrated through activities like using error flags as a trigger for self-editing, guessing the correct form, quickly adopting the suggested corrections, or noticing without attending to the error. Similarly, all of the students were substantively engaged with the feedback through various strategies listed in Table 5.1. Among those strategies, most frequently used engagement strategies among the learners were drawing on stored metalinguistic knowledge (10 students) and looking up online resources (9). Other engagement strategies were used by far fewer learners, including getting cues from Criterion error messages (4), translating into LI (4), and consulting a peer or the teacher (2).

The above results from TAP analyses find support from relevant excerpts from the stimulated recall interviews and focus group interviews. These will be discussed in more detail in the next section.

**Reported cognitive strategies to correct errors.** During the focus group interviews, students mentioned different strategies they used to process the errors pointed out by Criterion. All of the reported strategies find their equivalent forms among the engagement activities presented above. For example, students said that if they felt unsure about any errors, they would look up the dictionaries or google search/translate the answers (6 students), seek peer or teachers’ help (6), and read Criterion’s error explanations carefully to find clues (4). Less common was the approach taken by Thuy, a student participating in Focus Group Interview 3. Thuy made the following comment,

> *Normally, I examined the error to see what error type it was. If it was about a sentence structure I would look carefully at the sentence again…. I would look at the textual clues around the sentence too, then relate them to my learned linguistic knowledge in English [emphasis added]. I would determine if it was a grammatical or spelling error, then find a solution either from what I had previously learned, a dictionary, or an online resource.*
Or, being faced with the inability to correct an error, Hoa from Focus Group Interview 4 commented on her approach, which was,

*For me, if I couldn’t correct an error, I would replace the whole sentence with a new one. A new sentence which was sure to be a familiar sentence to me.*

Despite a few exceptions as above, the large majority of reported strategies were basically the same as the engagement activities found through the analyses of students’ think-aloud protocols conducted on a sub-sample of 14 students.

**Engagement with unhelpful feedback.** As Bitchener (2017) pointed out, what is often neglected in the field of written corrective feedback is the examination of factors that could explain learners’ failure to benefit from such feedback and accordingly their unsuccessful revised output. Inspired by this comment, this section focuses on how students were cognitively and behaviorally engaged with unhelpful feedback, one of the hypothesized factors behind learners’ failure to produce improved output. Unhelpful feedback includes both false positives and incorrect error feedback points categorized in the verification of *Criterion* ACF, accounting for 15% of the total error codes generated by *Criterion* in this data set.

Students’ TAP recordings reveal an interesting pattern where incorrect error codes were more likely to trigger substantive than perfunctory noticing among students. For example, An, during her two think-aloud protocols recorded for the study, was substantively engaged with all of the five false positives she received in her essay. The excerpt below provides one of her substantive noticing episodes in response to *Criterion* error message:

<table>
<thead>
<tr>
<th>Student text:</th>
<th><em>On the other hand, several learners try to pull down a large income because of fame or social position, which makes them to live meaninglessly.</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Error message:</td>
<td><em>This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.</em></td>
</tr>
<tr>
<td>TAP excerpt:</td>
<td><em>The next is meaninglessly. Meaninglessly is wrong? Meaninglessly? Meaningless? Meaningless is the synonym of meaning, and it is an adverb. So it has a LY in the end. Meaning ...less ...ly. MEANINGLESSLY. Lewis? Why it’s wrong? I don't know. I think this is right. Meaninglessly.</em></td>
</tr>
</tbody>
</table>
[After making no change, the student got the same error code on this word after the second submission, so she continued to process it] And meaninglessly. Meaninglessly? M E A N I N G. Umm. I don’t know why the meaninglessly are wrong. Meaninglessly? After live is a verb, so I use an adverb. So meaninglessly is right. Why it’s wrong? I need a second to look for dictionary. M E A N I N G L E S S L Y. Meaninglessly. I don’t think it’s wrong.

In response to false positives (as in the example above) or incorrect error codes, students most frequently resorted to the two follow-up strategies to confirm their doubts, which are reflecting on stored metalinguistic knowledge as above and looking up online resources as in the next example cited from Phu’s data set:

Student text: In aditions, It would be received the honor, cherishment through recognising the loyalty and rewarding their service.

Error message: This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.

TAP excerpt: It would be received the honor, cherishment. I think I’m right, right? [check on google translate] Cherish... ment. It right? Cher... C H E R I S H M N... M E N T.

I will check. I will check the Vietnamese. Be admired? Be loved. But I love the word cherish. I wanna use this word. The noun from cherish is. Cherish is the verb. Cherish, cherishing... no. Cherishing, just only. I don’t want to change to. No, I don’t want to change.

While the most common strategy was making no change, especially in response to false positive messages, if students did try to revise their text in response to the unhelpful feedback points, none of their revised text was a correct revision. One such example comes from An’s data and it was in response to a Criterion incorrect error code.

First draft: As a result, this method of making living can not bring us success and happiness but receiving our society’s disresect.

Error message: You may need to use a comma after this word.

Revised text: As a result, this method of making a living can not bring us success and happiness, but receiving our society’s disrespect.

Students’ reported strategies to deal with error codes they believed to be incorrect. During the focus group interviews, there were nine students who voiced their doubts about the accuracy of certain Criterion error tags they received. Three of these students explicitly expressed their refusal to revise their text if they considered the feedback to be false positives or incorrect. The following citation came from a student who reported that she ignored the error messages and moved on,

Over time, I realized that a few correct structures still got flagged by Criterion. Therefore, I just ignored the feedback and made no change. If I think there was no error, I just left the sentence unchanged. (Nhi, Focus Group Interview 1)

The other strategy, much less commonly found, was trying to adopt the suggested change provided by Criterion, as in the comment below:

Sometimes I was not sure what the error was, like an occasion when Criterion suggested that I should not have a comma in a sentence. Therefore, I deleted a comma, but felt that something was wrong there. I was not sure if I was right in the beginning or Criterion gave me a wrong error code. (Thanh, Focus Group Interview 4)

Even after adopting the suggested correction, Thanh still felt unsure about the revised sentence. Generally, the interviewed students tended to exercise some caution in adopting Criterion feedback. The question remains whether the processing and engagement with incorrect and false positive feedback codes from Criterion does good or harm to learners’ second language writing development. This question could find some further explanations and findings
in qualitative analyses using think-aloud protocols as well as stimulated recalls, especially among the multiple case-study students reported in Chapter 6.

**Research Question 3.2. How do students respond to Criterion ACF?**

“Response” to *Criterion* ACF refers to the EFL learners’ behavioural engagement with the feedback. More specifically, this research question focuses on two aspects related to learners’ response to *Criterion* ACF: (a) textual operations conducted by the learners to correct the errors and (b) their revision outcomes. Data for (a) come from the comparison of all the participating students’ first and revised drafts in the three practice sessions. The analysis considered textual operations such as deletion, addition, or reordering of texts. The revisions were further coded as correct, incorrect, removal, or retention of correct forms in the case of false positives, which make up the findings for part (b) in this section.

**Textual operations in response to Criterion ACF.** This part reports on the actions taken by learners, either in the forms of corrections that corresponded to *Criterion* error codes or operations other than addressing the error such as removal of the section including the error flag or ignoring the feedback by making no change. A summary of all the participating students’ revision operations is presented in Figure 5.1 and Table 5.2 below.

![Figure 5.1. Students’ text operations in response to Criterion ACF.](image-url)
As seen from Figure 5.1, when receiving Criterion ACF, students conducted five different textual operations, including (i) substitution when they changed words, phrases, or sentences with alternatives; (ii) addition when they added words, phrases, or sentences; (iii) deletion when they deleted words, phrases, or sentences from their texts; (iv) removal when they deleted parts of the essay that included the highlighted error; and (v) reordering when they reordered certain parts of a clause or swapped the positions of phrases in their essay. Another response was “no change”, meaning that no alteration was recorded in students’ essays irrespective of Criterion error tags. Among these six kinds of responses, removal and no change can be grouped as avoidance strategies because students did not seem to respond to the suggestions offered by Criterion. Table 5.2 provides further details of learners’ textual operations by error type.

Table 5.2

Students’ Textual Operations in Response to Criterion ACF by Error Type

<table>
<thead>
<tr>
<th>Error type</th>
<th>No change</th>
<th>Removal</th>
<th>Addition</th>
<th>Deletion</th>
<th>Reordering</th>
<th>Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capitalization</td>
<td>71</td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
<td>87</td>
</tr>
<tr>
<td>Duplicates</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faulty comparisons</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Garbled sentences</td>
<td></td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Negation errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Wrong article</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Proofread this</td>
<td>17</td>
<td></td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>Spelling</td>
<td>223</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>664</td>
</tr>
<tr>
<td>Determiner-noun agreement</td>
<td>57</td>
<td>14</td>
<td>32</td>
<td>32</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Subject-verb agreement</td>
<td>54</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>193</td>
</tr>
<tr>
<td>Missing final punctuation</td>
<td></td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronoun errors</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Ill-formed verbs</td>
<td>33</td>
<td>2</td>
<td>18</td>
<td>19</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>Run on sentences</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Fragments</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compound words</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preposition errors</td>
<td>48</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra comma</td>
<td>2</td>
<td>5</td>
<td>30</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyphen errors</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possessive errors</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing comma</td>
<td>104</td>
<td>3</td>
<td>75</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing question mark</td>
<td>6</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confused words</td>
<td>46</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing or extra articles</td>
<td>297</td>
<td>11</td>
<td>253</td>
<td>94</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>977</td>
<td>80</td>
<td>443</td>
<td>212</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>PERCENTAGE</td>
<td>32.3</td>
<td>2.6</td>
<td>14.6</td>
<td>7</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

As seen from Table 5.2 in combination with the visual illustration of Figure 5.1, the largest number of text operations was in the form of substitution of a word/phrase with another word/phrase. A closer look at all the substitution cases shows very few replacements of phrases. Instead, replacements took place at individual word levels of lexical items. For example,

First draft:   *Sometime, they are not keen on* there job anymore.
Revised draft: *Sometime, they are not keen on* their job anymore.

‘No change’ is the second most common category found in response to *Criterion* feedback. Besides the expected “no change” response to false positive error tags from *Criterion* which accounted for 11.7% of the total feedback points (equivalent to 325 false positive tags in raw counts), this figure of 32.3% “no-change” responses (977 of total responses) is high. This shows that in response to at least 600 *Criterion* error flags (correct and incorrect combined), students chose to keep their texts unchanged irrespective of their errors. The reasons why students decided to make no changes to their text after investigating the automated CF offered by *Criterion* based on different error codes will be described further in the qualitative analyses in response to Research Question 5 as well as case-study insights in Chapter 6.

Although not as commonly adopted, addition and deletion are the next two types of revision operations among these learners, at 14.6 and 7 percent respectively. Like substitution,
although addition or deletion took place mainly at the word level, there were also phrasal and even clausal additions or deletions, depending on the error types. The examples below illustrate how learners revised their texts using addition (a) and deletion (b):

(a) Addition
First draft:  
*If we have knowledge, we can business and could be earn a lot of money more.*
Error message:  
*You may have used the wrong form of this verb. Match the subject to the verb to decide whether you have used the verb correctly.*
Revised draft:  
*If we have knowledge, we can do business and could be earn a lot of money more.*

First draft:  
*One day, she met my aunt about her loneliness.*
Error message:  
*You may be using the wrong preposition.*
Revised draft:  
*One day, she met my aunt to talk about her loneliness.*

(b) Deletion
First draft:  
*As a consequence, they will be always selected for the rewards and recognitions so their chances of promotions are more greater.*
Error message:  
*You cannot use these two words together to make a comparison.*
Revised draft:  
*As a consequence, they will be always selected for the rewards and recognitions so their chances of promotions are greater.*

First draft:  
*Example, Vietnam use vietnam dong, the US use dollar, if we want to go to the US, we must change vietnam dong to dollar.*
Error message:  
*This sentence is confusing because it contains several grammatical or spelling errors.*
Revised draft:  
*Example, if we want to go to the US, we must change vietnam dong to dollar*
Reordering, a textual operation that takes more than simple rearrangement of the words in a clause or sentence, was least used by the participating learners, accounting for a mere 0.5% of the total revisions. The following is one of the few cases in this category.

First draft:  

I think successful person is person who have a good characteristics and be looked up to by everyone.

Error message:  

This sentence is confusing because it contains several grammatical or spelling errors. [Garbled sentences]

Revised draft:  

I think those who has good characteristics and be looked up to by everyone is a successful person.

Overall, there seems to be a relationship between the nature of Criterion ACF and students’ revision operations. As presented earlier in Chapter 4, Criterion tended to address superficial level errors rather than inter-clausal or inter-sentential problems. Therefore, textual operations in response to Criterion ACF conducted by students in the current study mainly related to local issues that addressed lexical changes (e.g. substitution, addition, or deletion of words). Much fewer instances of substantive changes such as reordering or substitution/addition of phrases/clauses were recorded. In addition to these, students sometimes chose avoidance strategies in the forms of removal and ‘no-change’ to deal with the automated corrective feedback they received from Criterion.

**Revision outcomes.** Revision outcomes refer to the efficacy of revisions made by students in the current study. With all the text operations presented in the previous section, this part considers how effective these textual operations are by coding them as correct or incorrect revisions. This section is subdivided into a presentation of overall revision outcomes by counting
the total number of correct/incorrect revisions across all categories, followed by revision success rates for each error type.

**Overall revision outcomes.** Based on students’ text operations, overall revision outcomes were further coded. Figure 5.2 provides a visual breakdown of students’ revision outcomes in response to Criterion ACF.

![Pie chart showing revision outcomes]

**Figure 5.2.** Overall revision outcomes in response to Criterion error tags.

As seen from Figure 5.2, there were four coded categories for students’ revision outcomes, including (i) **correct revision** indicating instances when students were able to produce a correct revised form; (ii) **incorrect revision** when students failed to provide a correct revision; (iii) **avoidance** when students avoided addressing the highlighted errors by either making no change to an incorrect structure or removing the part of their essay that contained the tagged error; and (iv) **retention of correct form** to indicate instances when students decided to keep their texts unchanged because the error tag they received was a false positive from Criterion. Table 5.3 provides further details about revision outcomes for specific error types. The table is ordered by the total number of total revisions in each error type.
### Table 5.3

**Summary of Revision Outcomes**

<table>
<thead>
<tr>
<th>Error type</th>
<th>Correct Revision</th>
<th>Incorrect Revision</th>
<th>Avoidance*</th>
<th>Retention of correct forms**</th>
<th>TOTAL revisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error type</td>
<td>Correct Revision (%)</td>
<td>Incorrect Revision (%)</td>
<td>Avoidance (%)</td>
<td>Retention of correct forms (%)</td>
<td>TOTAL number of errors</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>---------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Spelling</td>
<td>68.6</td>
<td>5.6</td>
<td>23.3</td>
<td>2.6</td>
<td>900</td>
</tr>
<tr>
<td>Subject verb agreement</td>
<td>66.4</td>
<td>12.6</td>
<td>20.2</td>
<td>0.8</td>
<td>262</td>
</tr>
<tr>
<td>Garbled sentences</td>
<td>60</td>
<td>55</td>
<td>25</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Wrong article</td>
<td>59.1</td>
<td>18.2</td>
<td>22.7</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Capitalization</td>
<td>55.9</td>
<td>1.2</td>
<td>42.4</td>
<td>0.6</td>
<td>170</td>
</tr>
</tbody>
</table>

* Can result from text operations of ‘Removal’ or ‘No change’ in response to correct/incorrect error messages

** When students made no changes in response to a ‘False Positive’

A combined look at Figure 5.2 and Table 5.3 shows that the overall successful revision rate using all error types was 54%, with the success rate being defined as the number of correct revisions out of total feedback points received. Avoidance accounted for 28%, while 10.5% of all their revisions were incorrect, leaving 7.5% for retention of originally correct text in response to Criterion’s false positives.

**Revision outcomes by error type.** Table 5.4 gives details about revision success rates for 14 error types with 20 or more occurrences in the data. The table is ordered by the successful revision rates across the error types from the highest to the lowest correct revision rate.

Table 5.4

*Revision Success Rates for Most Frequently Recurring Error Types*
<table>
<thead>
<tr>
<th>Error Type</th>
<th>Count</th>
<th>Correct</th>
<th>Partial</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determiner-noun agreement</td>
<td>55.5</td>
<td>10.4</td>
<td>34.1</td>
<td>0</td>
<td>164</td>
</tr>
<tr>
<td>Ill-formed verbs</td>
<td>52.5</td>
<td>19.2</td>
<td>27.5</td>
<td>0.8</td>
<td>120</td>
</tr>
<tr>
<td>Missing or extra article</td>
<td>46.7</td>
<td>12</td>
<td>32.9</td>
<td>8.4</td>
<td>749</td>
</tr>
<tr>
<td>Confused words</td>
<td>45.7</td>
<td>13.8</td>
<td>33.6</td>
<td>6.9</td>
<td>116</td>
</tr>
<tr>
<td>Extra comma</td>
<td>43.7</td>
<td>9.9</td>
<td>26.8</td>
<td>19.7</td>
<td>71</td>
</tr>
<tr>
<td>Preposition errors</td>
<td>36.4</td>
<td>13.1</td>
<td>33.3</td>
<td>17.2</td>
<td>99</td>
</tr>
<tr>
<td>Missing comma</td>
<td>35.3</td>
<td>6.4</td>
<td>14.4</td>
<td>43.9</td>
<td>187</td>
</tr>
<tr>
<td>Fragments</td>
<td>33.3</td>
<td>11.8</td>
<td>41.2</td>
<td>13.7</td>
<td>51</td>
</tr>
<tr>
<td>Missing final punctuation</td>
<td>30.4</td>
<td>0</td>
<td>69.6</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Proofread this</td>
<td>28.8</td>
<td>42.4</td>
<td>27.1</td>
<td>1.7</td>
<td>59</td>
</tr>
</tbody>
</table>

As can be seen in Table 5.4, revision success rates varied across error types. Students were most successful at correcting errors related to spelling (68.4%), followed by subject-verb agreement (66.4%), garbled sentence (65%), and wrong article (59%). Capitalization, determiner-noun agreement, and ill-formed verbs also had high revision success rates, at above 50%. Interestingly, five out of these seven most successfully remedied error types receive generic feedback from Criterion, suggesting that a highly implicit feedback approach may lead to effective self-revising activities.

On the contrary, if we consider correct revision rates only, students performed worst at revising structures receiving fragments, missing final punctuations, and proofread error codes (33%, 34%, and 28% success rates, respectively). It can be hypothesized that revising fragments or proofread error codes requires textual operations that are beyond the individual word level. Understandably, the two error types of ‘fragments’ and ‘proofread this’ were actually not specific error categories. Instead, the whole clause was coded as containing an error or errors without the specific errors being pointed out. This seems to require more extended mental effort on learners’ part, and thus may have impacted learners’ ability to revise errors in these
categories. What is baffling is that for missing final punctuation, deemed a highly treatable error with a simple correction strategy as pointed out in Criterion’s specific message for this error type, students chose to avoid making the suggested revision most of the time, despite the finding earlier reported in Chapter 4 that Criterion had a satisfactory precision statistic of 87% for this type of error. Due to the treatability of errors in this category, all of the students’ revisions in response to Criterion missing final punctuation error tags were correct.

Worth special mention are error types with high percentages of incorrect revisions and avoidance. Judged by incorrect revision rates, proofread errors were by far the most challenging for students in the present study, with 42.4% revision attempts failing and 27.1% avoidance. Ill-formed verbs and wrong articles were also among the least likely to be successfully remedied using Criterion error feedback as a guide. In addition, the figures also suggest that students were more likely to adopt an avoidance strategy (that is, choosing not to address an error pointed out to them) than to attempt a revision that may be incorrect. This is demonstrated in generally higher avoidance rates compared to incorrect revision rates across most error types. Very high avoidance rates were experienced among mechanics error codes of missing final punctuation (69%), capitalization (42.4%), and extra commas (26.8%), as well as the usage error type of missing/extra articles (32.9%). Grammar error codes were also quite frequently ignored, including determiner-noun agreement, subject-verb agreement, ill-formed verbs (all above 20%).

Furthermore, the findings reveal feedback on comma as a specific area where Criterion generated the highest rate of false positives. Fortunately, in response to these feedback points, students chose to retain their text rather than taking up the system’s suggested change for most of the time. False positives were also found regularly with preposition and fragment errors according to Table 5.3.

**Summary of findings for RQ3.** Data for Research Question 3 reveal interesting results about students’ cognitive engagement as well as revision practices in response to Criterion automated corrective feedback. First of all, students more often perfunctorily than extensively engaged with the feedback, with two thirds of all the coded episodes showing their limited engagement with Criterion ACF. When being extensively engaged with the feedback, however, their most commonly chosen strategies were looking up online resources and referring to stored metalinguistic knowledge to find ways to resolve the errors. Students’ behavioral engagement with the feedback as demonstrated through their textual operations partially reflected their level of cognitive engagement with the feedback. Specifically, the most commonly found textual operations were substitution and no change responses, followed by addition and deletion.
Notably, students’ textual operations took place at the surface level of texts, usually individual lexical items or short phrases. Recorded revision success rate was modest, with slightly more than half of all the revised forms being correct. They also demonstrated high rates of avoidance strategies when faced with difficulties to find a revised form.

**Research Question 4. How does EFL students’ accuracy in L2 writing change after a semester’s access to and use of Criterion ACF?**

Six measures were calculated for the examination of changed accuracy in students’ writing over the studied period. The measures include overall accuracy and five specific accuracy measures of article, third person singular, copula, plural, and comma usage. Besides overall accuracy, the five specific measures were selected based on the most frequently tagged error types by Criterion in the total practice essays written by the participating students. By focusing on the error types which received the highest numbers of Criterion error feedback comments, the current study aims to make possible connections between the use of Criterion ACF for writing and revision practice among the learners with their observed changes in writing accuracy over the five-month period.

In this section, the presentation of findings for each measure starts with descriptive statistics, followed by the graph visualizing the two groups’ performance on that measure, and ends with inferential statistics related to the intervention effect of using Criterion over a semester among the experimental group and how that compares to the comparison group.

Repeated measures ANOVAs were conducted to see if the intervention (the incorporation of Criterion automated corrective feedback in the writing course) over a semester led to any changes in students’ writing accuracy. In each model, the between subject effect was group (with two levels, comparison and experimental), and the within subject effect was time (with three levels, pre-test, post-test, and delayed post-test for measures where the two groups did not differ significantly at pre-test or with two levels, gain score 1 and gain score 2 for measures where there was a significant difference at pre-test for the two groups, including overall accuracy and third person singular use). Where relevant, post-hoc tests were also conducted to shed further lights on where the effects were actually located.

Each of the tables from Table 5.5 to Table 5.10 provides a summary of descriptive statistics of students’ performance for the holistic accuracy measure or analytic measures for the five specific morphemes under study, while the figures are visualizations of the statistics for better understanding of the data.
**General accuracy gains.** General accuracy gains were calculated using the overall accuracy measure which investigated error-free clauses versus clauses with error(s) in learners’ essays from the pre, post, and delayed post-tests using the weighted clause ratio approach (See Appendix C for details). The experimental group’s pre-test overall accuracy score was significantly higher than that of the comparison group, as can be seen in Table 5.5 below which introduces descriptive statistics about the overall accuracy of the two groups.

<table>
<thead>
<tr>
<th>Time</th>
<th>Experimental group</th>
<th>Comparison group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
</tr>
<tr>
<td>Pre</td>
<td>38</td>
<td>0.81</td>
</tr>
<tr>
<td>Post</td>
<td>38</td>
<td>0.83</td>
</tr>
<tr>
<td>Delayed</td>
<td>38</td>
<td>0.81</td>
</tr>
</tbody>
</table>

The descriptive statistics shown in Table 5.5 indicate that the experimental group’s overall accuracy experienced little change across the pre, post, and delayed post-test ($M = 0.81$ for both pre and delayed post-test, and 0.83 for post-test), while that of the comparison group saw some improvement from 0.69 for the pre-test to 0.77 and 0.74 in the post and delayed post-test results.

Figure 5.3 provides a visual illustration of the estimated means from the model of students’ accuracy changes from the pre-test to post and delayed post-tests. It can be seen that the comparison group made improvement from the pre-test to post-test, but there was some regression in accuracy from the post-test to delayed post-test results. On the other hand, the experimental group performed quite consistently from the pre-test to the post and delayed post-tests, with a slightly higher mean accuracy score in the post test.

To examine whether the change over time in the two groups was statistically meaningful, inferential statistics were used. Before inferential statistics were conducted, the two groups’ pre-test scores were tested to see if they were statistically different by using an independent-samples T-test (see Table 3.7, Chapter 3). The result showed that the two groups’ pre-test scores were statistically different for this accuracy measure, $t(73) = 4.56, p < .001$. Therefore, analyses for this measure were adjusted for baseline levels using gain scores instead of raw scores recorded for individual learners.
A repeated measures ANOVA using gain scores was used for this analysis. The repeated measures ANOVA results show that there was no significant interaction effect between time and group on the accuracy scores for the experimental and comparison group, \( F(1, 73) = .208, p = .649, \eta^2 = .003 \). There was also no significant overall main effect of time \( F(1, 73) = 3.605, p = .062, \eta^2 = .047 \), but a significant main effect for group on accuracy gain scores for the two groups was found, \( F(1, 73) = 12.396, p = .005, \eta^2 = .101 \). The significant main effect for group was followed up by an independent-samples t-test to see how and where the groups differed.

![Figure 5.3. Overall accuracy from pre-test to post-test and delayed post-test among two groups.](image)

The independent-samples T-test results showed that there was a significant difference in mean overall gain scores at post-test between the experimental and comparison groups, \( t(62.799)^1 = -2.798, p = .007 \). The average gain score for students in the experimental group was 0.069 lower than the average gain score at post-test for comparison group students. These results suggest that the incorporation of Criterion automated corrective feedback over a semester did not facilitate general accuracy gains for the experimental group, while the comparison group,

---

1 Equal variances not assumed
Despite the lack of access to Criterion ACF, significantly gained more in overall accuracy. Such an intriguing result will be discussed in more details in Chapter 7.

Next, an independent-samples T-test was conducted to compare overall accuracy retention scores (i.e., gain scores at the delayed post-test) in the experimental group and that of the comparison group. There was no significant difference in the experimental group (\( M = -.0161, SD = .084 \)) and comparison group (\( M = -.0262, SD = .108 \)) conditions, \( t(73) = .457, p = .024 \). These results suggest that neither group retained overall accuracy, despite the access to Criterion ACF as an additional resource for the experimental group.

In summary, the above results do not suggest any significant effect of using Criterion ACF in helping learners to gain in overall accuracy. No retention effect was found regarding ACF effect either.

**Article use.** Participating students’ article use was investigated by examining their accuracy in using the indefinite articles *a/an* and definite article *the* in the pre, post, and delayed-post tests. This accuracy measure was calculated using an obligatory occasion analysis approach.

### Table 5.6

Descriptive Statistics of Article Use Accuracy

<table>
<thead>
<tr>
<th>Time</th>
<th><strong>Experimental group</strong></th>
<th><strong>Comparison group</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N )</td>
<td>Mean</td>
</tr>
<tr>
<td>Pre</td>
<td>38</td>
<td>0.74</td>
</tr>
<tr>
<td>Post</td>
<td>38</td>
<td>0.80</td>
</tr>
<tr>
<td>Delayed</td>
<td>35</td>
<td>0.78</td>
</tr>
</tbody>
</table>

The descriptive statistics shown in Table 5.6 indicate that the experimental group gained slightly in article usage from the pre-test (\( M = 0.74, SD = 0.18 \)) to the post-test and delayed post-test (\( M = 0.80, SD = 0.15; M = 0.78, SD = 0.14 \), respectively). This is also true of the comparison group with pre-test mean at 0.70, which increased to 0.74 and 0.73 in post and delayed post-tests, respectively.

Figure 5.4 provides a visual illustration of the mean scores of students’ article use from the pre-test to post and delayed post-tests. It can be seen that the two groups generally made improvement in the use of article from the pre-test to the post-test, but there was some slight
regression from the post-test to the delayed post-test accuracy. The pattern was quite similar for both groups, experimental and comparison.

![Graph](image)

**Figure 5.4.** Changes in article usage from pre-test to post-test and delayed post-test among two groups.

Pre-test scores of the two groups were not statistically different for this accuracy measure, $t(73) = .896, p = .373$. Therefore, a repeated measures ANOVA using pre, post and delayed post-test scores was used for this analysis. The repeated measures ANOVA results show there was no significant interaction effect of time and group on the article use scores for the experimental and comparison groups, $F(2,140) = 0.015, p = .985, \eta^2 = .000$. No significant effect of time was found, $F(2,140) = 1.746, p = .178, \eta^2 = .024$. The main effect of group on article use accuracy scores for the two groups was not significant either, $F(1,70) = 3.027, p = .086, \eta^2 = .041$. The results suggest no significant treatment effect of using *Criterion* ACF on students’ article use accuracy throughout a semester, as well as the lack of retention over time.

**Third person singular.** Participating students’ third person singular use was investigated by examining how accurately they used the third-person singular morpheme in the pre, post, and delayed-post tests. Descriptive statistics shown in Table 5.7 indicate that the experimental group performed worse over time, decreasing from a mean of 0.80 in the pre-test to means of 0.64 and
0.70 in the post and delayed post-tests. On the contrary, the comparison group improved substantially in third person singular accuracy from the pre-test \((M = 0.50, SD = 0.42)\) to the post-test \((M = 0.80, SD = 0.23)\), but their delayed post-test result was lower, \(M = 0.50, SD = 0.42\).

Table 5.7

*Descriptive Statistics of Third Person Singular Use Accuracy*

<table>
<thead>
<tr>
<th>Time</th>
<th>Experimental group</th>
<th>Comparison group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N)</td>
<td>(Mean)</td>
</tr>
<tr>
<td>Pre</td>
<td>36</td>
<td>0.80</td>
</tr>
<tr>
<td>Post</td>
<td>36</td>
<td>0.64</td>
</tr>
<tr>
<td>Delayed</td>
<td>37</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Note.* N-sizes for the two groups vary and can be lower than the total group numbers because not all of the students in either group had errors recorded for specific targeted linguistic features.

Figure 5.5 provides a visual illustration of the means of the students’ third person singular use accuracy changes from the pre-test to post and delayed post-tests, which accentuated the fact that the two groups followed quite opposite patterns in this accuracy measure as earlier shown by the descriptive statistics. While the comparison group students improved substantially from the pre-test to the post test, followed by some regression in the delayed post test results, the experimental students experienced a decrease in accuracy in the post-test but a slight gain in accuracy in the delayed post-test.
Before conducting inferential statistics to examine whether the change over time in the two groups was statistically meaningful, the two groups’ pre-test scores were tested to see if they were statistically different by using an independent-samples T-test (see Table 3.7, Chapter 3). The results showed that the experimental group’s pre-test third-person singular accuracy score was significantly higher than that of the comparison group. Therefore, analyses for this measure adjust for baseline levels using gain scores instead of raw scores recorded for individual learners.

A repeated measures ANOVA using gain scores was conducted. The repeated measures ANOVA results show that there was a significant interaction effect of time and group on the third person singular accuracy gain scores for the experimental and comparison groups, $F(1, 73) = 4.731, p = .033, \eta^2 = .061$. No significant main effect of time was found, $F(1, 73) = .680, p = .062, \eta^2 = .009$. The main effect of group on third person singular accuracy gain scores for the two groups was also significant, $F(1, 73) = 17.493, p < .001, \eta^2 = .193$.

Because of the significant interaction effect and group effect, a follow-up independent-samples T-test was conducted to explore these findings in greater depths through the comparison of third person singular accuracy gain scores in the experimental group and that of the comparison group. The independent-samples T-test results showed that there was a significant difference in mean third-person singular gain scores at post-test between the experimental and

*Figure 5.5. Changes in third person singular usage from pre-test to post-test and delayed post-test among two groups.*
comparison groups, \(t(59.990^2) = -5.267, p < .001\). The average gain score in third-person singular use for students in the experimental group was 0.472 lower than the average gain score at post-test for comparison group students. These results suggest that the incorporation of Criterion automated corrective feedback over a semester does not facilitate accuracy gains for the experimental group, while the comparison group, despite the lack of access to Criterion ACF, significantly gained more in third person singular use accuracy.

An independent-samples T-test was also conducted to compare third person singular accuracy gain scores at the delayed post-test in the experimental group and that of the comparison group. The independent-samples T-test results showed that there was a significant difference in mean third-person singular gain scores at the delayed post-test between the experimental and comparison groups, \(t(73) = 2.175, p = .033\). The average gain score in third-person singular use recorded at the delayed post-test compared to the post-test results for students in the experimental group was 0.235 higher than the average gain score at the delayed post-test for comparison group students. These results suggest that despite some regression in the post-test, some accuracy gain in terms of third person singular use was recorded from the post to delayed post-tests among the experimental students who had Criterion ACF incorporated as part of their writing course for the experimental group, while the comparison group, without access to ACF, did not retain accurate use of third person singular over time.

Overall, the above results do not indicate any significant effect of using Criterion ACF in helping learners to gain in third person singular use accuracy. Moreover, the increased accuracy from the post to delayed post-tests regarding ACF effect for the experimental group should be interpreted with caution due to a previous drop in accuracy gain scores from pre to post tests. This will be further discussed in Chapter 7.

**Copula use.** The copula use under investigation is comprised of both singular (*is, am*) and plural (*are*) forms of the copula *be* in learners’ essays across the pre, post, and delayed-post tests. This accuracy measure was calculated using an obligatory occasion analysis approach.

\[o^2\] Equal variances not assumed
The descriptive statistics shown in Table 5.8 indicate that the experimental group was quite consistent in copula use accuracy from pre-test ($M = 0.89, SD = 0.19$) which stayed the same before increasing slightly to $M = 0.92$ in the delayed post-test. This was also true for their comparison counterparts with means for all three tests being around 0.88.

Figure 5.6 provides a visual illustration of the means of students’ copula use accuracy changes from the pre-test to post and delayed post-tests. It can be seen that both groups were fairly consistent in copula use performance throughout the period, yet with slightly different trends. While the comparison group gained in accuracy from the pre-test to the post-test, their performance in the delayed post-test regressed. On the contrary, despite a decrease in performance from the pre to the post-test, students in the experimental group did better in terms of copula use in their delayed post-test.
Figure 5.6. Changes in copula usage from pre-test to post test and delayed post-test among two groups.

Pre-test scores of the two groups were not statistically different for this accuracy measure, $t(70) = .466, p = .643$. Therefore, a repeated measures ANOVA using pre, post and delayed post-test scores was conducted. The repeated measures ANOVA results show no significant interaction effect of time and group on the copula use accuracy change for the experimental group and comparison group over the five-month period, $F(2,138) = 0.499, p = .608, \eta^2 = .007$. No significant main effect of time was found, $F(1,69) = 0.418, p = .520, \eta^2 = .006$. The main effect of group was not significant either, $F(1,69) = 0.380, p = .539, \eta^2 = .005$. Overall, the results show no significant treatment effect of using Criterion ACF on students’ copula use accuracy throughout a semester, as well as the lack of retention over time.

Plural use. The plural morpheme under investigation is comprised of both its short (suffix *s*) and long (suffix *es*) surface forms in learners’ essays across the pre, post, and delayed-post tests. The accuracy measure for this morpheme was calculated using an obligatory occasion analysis approach.
Table 5.9.
Descriptive Statistics of Plural s/es Accuracy

<table>
<thead>
<tr>
<th>Time</th>
<th>Experimental group</th>
<th></th>
<th></th>
<th>Comparison group</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Pre</td>
<td>36</td>
<td>0.78</td>
<td>0.24</td>
<td>35</td>
<td>0.74</td>
<td>0.25</td>
</tr>
<tr>
<td>Post</td>
<td>36</td>
<td>0.83</td>
<td>0.15</td>
<td>35</td>
<td>0.83</td>
<td>0.17</td>
</tr>
<tr>
<td>Delayed</td>
<td>36</td>
<td>0.87</td>
<td>0.19</td>
<td>35</td>
<td>0.65</td>
<td>0.35</td>
</tr>
</tbody>
</table>

The descriptive statistics shown in Table 5.9 indicate that the experimental group steadily gained in plural use accuracy from the pre-test ($M = 0.78$, $SD = 0.24$) to post and delayed post-tests, while the comparison group’s mean score, starting at $M = 0.74$, $SD = 0.25$ at pre-test, increased quite sharply to 0.83 in the post-test before plummeting in the delayed post-test ($M = 0.65$, $SD = 0.35$).

Figure 5.9 provides a visual illustration of the means of students’ plural use accuracy changes from the pre-test to post and delayed post-tests. Interestingly, without *Criterion* corrective feedback as part of the feedback mechanism, the comparison group made great improvement in using plural forms from the pre-test to the post-test before the accuracy rate of their plural usage plunged in the delayed post-test. On the contrary, the experimental group experienced a gradual increase in accuracy across the three tests.

Pre-test scores of the two groups were not statistically different for this accuracy measure, $t(72) = .813$, $p = .419$. Therefore, a repeated measures ANOVA using pre, post and delayed post-test scores was conducted. The repeated measures ANOVA results show that there was a significant interaction effect of time and group on the plural use accuracy scores for the experimental group and comparison group over the five-month period, $F(2,138) = 5.709$, $p = .004$, $\eta^2 = .076$. However, there was no significant main effect of time, $F(1,69) = 0.000$, $p = .997$, $\eta^2 = .000$, but there was a significant main effect of group on plural use accuracy scores, $F(1,69) = 5.756$, $p = .019$, $\eta^2 = .077$. 

140
Figure 5.7. Changes in plural usage from pre-test to post-test and delayed post-test among two groups.

Because of the significant interaction and group effect, a follow-up independent-samples T-test was conducted to explore these findings in greater depths through the comparison of plural use accuracy scores at post-test in the experimental group and that of the comparison group. The independent samples T-test results showed that there was no significant difference in mean plural use scores at post-test between the experimental and comparison groups, $t(72) = -0.216$, $p = 0.830$.

An independent-samples T-test was also conducted to compare plural use accuracy scores at the delayed post-test in the experimental group and that of the comparison group. The results showed that there was a significant difference in mean plural use scores at the delayed post-test between the two groups, $t(71) = 3.387$, $p = .001$. The average plural use accuracy score for students in the experimental group was .221 higher than that of comparison group students.

Paired samples T-tests were conducted to compare within-group effects for the experimental and comparison groups. The results showed that the experimental students did not make significant gains in plural use accuracy from the pre-test to the post and also from the post to delayed post-tests, $t(69) = 1.22$, $p = .225$, and $t(69) = 1.76$, $p = .084$, respectively. On the other hand, there was a significant gain in plural use accuracy measured at post-test for comparison group students, $t(69) = 2.17$, $p = .032$, and the decrease from the post-test to delayed post-test
was also significant, \( t(71) = 2.15, p = .001 \). This is interpreted in combination with the independent samples T-tests results and the decrease in plural use accuracy scores from the pre to post-tests among comparison group students. All the results suggest no significant treatment effect of using Criterion ACF on students’ plural use accuracy throughout a semester, as well as the lack of retention over time.

**Comma use.** The use of comma in learners’ essays from the pre, post, and delayed post-tests was measured using an obligatory occasion analysis. For the six specific rules that were taken into account for the coding of comma usage in the study, see Appendix D.

<table>
<thead>
<tr>
<th>Table 5.10 Descriptive Statistics of Comma Use Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Pre</td>
</tr>
<tr>
<td>Post</td>
</tr>
<tr>
<td>Delayed</td>
</tr>
</tbody>
</table>

The descriptive statistics shown in Table 5.10 indicate that the experimental group gained slightly in comma use accuracy from the pre-test \( (M = 0.79, SD = 0.13) \) to post-test and delayed post-tests, while the comparison group’s mean score for this same measure, starting at \( M = 0.83, SD = 0.12 \) at pre-test, slightly decreased in the post-test before decreasing in the delayed post-test \( (M = 0.77, SD = 0.12) \).

Figure 5.8 provides a visual illustration of the means of students’ comma use accuracy changes from the pre-test to post and delayed post-tests. Without Criterion corrective feedback as part of the feedback mechanism, the comparison group’s performance from the pre-test to post-test and delayed post-test deteriorated over time. On the contrary, students in the experimental group improved their comma usage slightly but steadily from the first to the final test.

Pre-test scores of the two groups were not statistically different for this accuracy measure, \( t(73) = -1.458, p = .149 \). Therefore, a repeated measures ANOVA using pre, post and delayed post-test scores was conducted. The repeated measures ANOVA results show that there was a significant interaction effect of time and group on comma use accuracy scores for the
experimental and comparison groups, $F(2,146) = 5.501, p = .005, \eta^2 = .070$. However, the main effect of time was not significant, $F(1,73) = 0.529, p = .469, \eta^2 = .007$. The main effect of group was not significant either, $F(1,73) = 0.065, p = .800, \eta^2 = .001$.

However, the main effect of time was not significant, $F(1,73) = 0.529, p = .469, \eta^2 = .007$. The main effect of group was not significant either, $F(1,73) = 0.065, p = .800, \eta^2 = .001$.

Because of the significant interaction effect of group and time, post-hoc tests were conducted to investigate the within-group effects. The results showed that there was no significant change in comma use accuracy measured at post test for students in the comparison group, $t(72) = -.58, p = .564$, and the increase in accuracy from the post to delayed post-test was not significant either, $t(72) = 1.63, p = .109$. On the other hand, students in the experimental group made no significant gain in comma use accuracy from the pre-test to post-test, $t(72) = 0.75, p = .461$, but the gain from the post to delayed post-tests was significant, $t(72) = 4.48, p < .001$. This shows no significant treatment effect of using Criterion ACF on students’ plural use accuracy throughout a semester, although some significant improvement in comma use accuracy from the post to delayed post-tests was recorded.

![Figure 5.8. Changes in comma usage from pre-test to post-test and delayed post-test among two groups](image)

**Summary of findings for RQ4.** The descriptive and inferential statistics procedures conducted on participating learners’ six accuracy measures using either the set of scores from pre/post/delayed post-test essays or gain scores recorded at their post/delayed post-tests provided...
answers to Research Question 4 in the current study. Regarding the treatment group, the incorporation of Criterion ACF in the writing course was found to have no significant treatment effect on the accuracy of the experimental students’ use of the five specific morphemes under study (article, third person singular, copula, plural, and comma) as well as the overall accuracy scores. In addition, there was no significant retention effect recorded for these students in the accuracy measures of article, copula, and plural, while that for comma use showed some statistical retention effect, despite the nonsignificant accuracy gain from the pre-test to post-test. Noticeably, there was increased accuracy in third person singular usage from the post to delayed post-tests among these learners, which is to be interpreted with caution due to their previous decreased accuracy from the pre to post-tests.

Turning to the comparison group, despite the lack of access to Criterion corrective feedback, significant accuracy gains were found for their use of third person singular, plural use, and overall accuracy from the pre-test to the post-test, although no significant retention was recorded across all of the six measures. With the exception of plural use, their significant gains in overall accuracy and third person singular among the comparison group students should be interpreted in combination with the fact that their pre-test accuracy scores for these two measures were significantly lower than those of the experimental students, a point to be further discussed in Chapter 7.

**Research Question 5. What are students’ perceptions of Criterion ACF?**

In the current study, students’ perceptions of automated corrective feedback were investigated mainly through focus group interviews. As the interviews were thematically coded through an emergent approach, this section is structured accordingly based on the themes that emerged from the analyses of students’ group discussions, with some triangulation with the TAP students’ stimulated recall interviews. There are three broad themes discussed below: (a) *general assessment of Criterion ACF*, (b) *automated corrective feedback in comparison to teacher feedback*, and (c) *students’ expectations about feedback explicitness*.

**General assessment of Criterion ACF.** Four focus group interviews were conducted for the current study, with a total of 19 students participating in the focus groups. All of the interviewed students had ambivalent feelings about the automated corrective feedback provided by Criterion, with more positive than negative comments about the software. Their positive evaluation of the feedback is presented with further elaboration on the specific benefits students
mentioned during the interviews. This is followed by the findings on negative evaluation of the feedback and explanatory comments regarding the source of their distrust in the feedback.

**Positive evaluation.** All of the students were generally positive about *Criterion* corrective feedback. Firstly, they perceived the ACF to be helpful for developing their writing skills. Below is a representative comment,

*In general, I think this feedback tool is useful, at least for me, as with its correct error feedback I can write better over time.* (Nhat, Focus Group Interview 2)

Nhi expressed a similar comment below: “For me, the software is pretty useful as it shows me my grammatical errors” (Focus Group Interview 3). The opinion was shared by An when this student discussed her revisions using *Criterion* ACF in the stimulated recall,

*When I try to, it helps me to find out which mistake I have and the final goal is I want to know which mistake I’m getting and it helps me enlarge my knowledge about writing…. All in all, it’s very useful for me, I think.*

Among students’ several comments mentioning their favorable views of the feedback, *Criterion* treatment of errors was considered systematic and well-organized. In addition, the system’s straightforward and simple language of error explanation was perceived as facilitating revisions of errors. For example, all of the eleven students participating in the first two focus groups said that they had no problem understanding the error messages generated by *Criterion*, adding that its vocabulary was easy to follow:

Researcher: *The language used to explain the errors, is it easy or difficult for you?*

Students (answer in chorus): *Pretty easy to follow... no issues with it.* (Focus Group 1)

The students added that the vocabulary was “easy to understand, we found it easy to follow, the vocabulary was quite simple.” (Focus Group 2). Similarly, Nga from Focus Group 4 said, “Generally, it was easy to understand, comprehensible… There were sentences that we understand straight away upon reading them.” Another student from the same focus group further commented that her understanding of the error messages from *Criterion* came with
practice. Once she got more familiar with the feedback, she could easily know what the error was about without much effort.

*On my first practice, I did not quite understand the error explanations. But from the second time, I got used to it, so just a brief look at it was enough for me to understand what it wanted to tell me.* (Thu, Focus Group 4)

According to the students from the four focus groups, *Criterion* ACF also contributes to raising their awareness about frequently made errors. An exemplary comment related to this point was made by Duong, “it’s like we became aware of the errors we often made, so we learn to avoid them the next times” (Focus Group Interview 3). This could be realized thanks to the fact that, as Thu from Focus Group 4 pointed out, “As we write [on *Criterion*], our essays could be corrected many times, so we could draw the lessons to avoid repeating the same errors in later submissions”. She cited her example of usually making spelling errors previously, but with *Criterion* ACF pointing out this error type in her essay, she added that she gradually became more conscious about spelling in later compositions. *Criterion* seemed to have transcended from the treatment-of-errors function to that of error prevention by raising students’ alertness to errors they committed previously, which made writers become more careful about their grammatical or lexical choices:

*Overall, I think the machine’s [Criterion] corrective feedback was useful for me, and with correct feedback, it helps me write better in subsequent assignments.*

Trying to explain the comment that “it helps me write better in subsequent assignments”, Mai said that the lowered numbers of errors she and her peers found in later compositions were not attributable to increased accuracy over a short period of a few weeks. Instead, she believed that this was thanks to her “becoming more careful, as it [Criterion] pointed out certain types of errors to her before”. Regarding this, Nhung made a clearer point by comparing *Criterion* ACF to a gatekeeper while she was learning to write, “When I wrote using the machine [Criterion], it served as a supervisor”.

In summary, *Criterion* was thought to be good at detecting errors related to spelling and grammar. Such comments kept coming up during the four focus group interviews when students reflected on the feedback they received. When asked about their level of trust in the accuracy of the feedback they received, all of the focus group students cited an overall percentage of 70% or
more in the system when it came to its detection of grammar, mechanics, and spelling errors. Among these, 16 out of the 19 students reported high scores of Criterion ACF accuracy, mostly ranging between 8 to 9 on a 1-10 rating scale. As in the following comment, the student made it clear that if content was excluded from her evaluation of Criterion accuracy, she would assign a score of up to 9 out of 10:

If judged solely by feedback on grammar and usage, and simple errors like punctuation, I would give Criterion a 9, and of course content is excluded from this evaluation. (Phuong, Focus Group 4)

Similar to Phuong, nine other students in the focus groups assigned a trust level of 90% or more to the accuracy of Criterion’s ACF on spelling, mechanics, and simple grammatical errors. From the discussions about the feedback, there were two most frequently cited benefits of using Criterion ACF during the interviews. The first is that Criterion’s feedback generation approaches help students to develop a sense of error types they usually made, thus raising their consciousness about grammatical accuracy when writing. The comment made by Nga below illustrates this point:

It gets me into the habit of thinking more consciously about grammar issues when I write, and that helps with revising ideas and stuff, too. It is generally very helpful. (Focus Group 2)

A similar point was made by Duong,

Like, after a while we get the sense of the kinds of errors we usually commit, and learn from that…. For example, if we often make a lot of mistakes in spelling or sentence structures, we will learn from it and do better next time. (Duong, Focus Group 3)

Van (Focus Group 2) made this point more specific by mentioning the benefit of using Criterion ACF for her writing practice,

It mainly helps me avoid spelling errors, spelling and grammar errors. We need to work on content through self-practice ourselves, but the system gives us an overall evaluation of what our most frequent errors are to avoid losing marks on these aspects.
Secondly, the benefit of *Criterion*’s instant and online feedback helping with more effective revisions was mentioned by five students across the four focus group interviews, as in the comment that “The feedback generated by the web page was instant right after I submitted the essay, so I could revise straight away” (Trung, Focus Group 1). Other comments were more specific about why the immediacy of the feedback helps.

*The essay is still fresh in my mind and I can still remember what I wrote very clearly. I know what is not correct. If I wait for the teacher feedback, I would forget what I wrote in my essay, so I will also forget all the mistakes in it.* (Duyen, Focus Group 2)

Besides the easy retrieval of information thanks to the instant automated corrective feedback from *Criterion*, the convenience of access to the feedback as well online reference materials was mentioned in the next comment,

*I think quite good when I do all the searching online so I can use the Internet for for search the vocabulary, some dictionary online. And I think it's quite good when we can write and check this word at the same time.* (Ngan, Focus Group 3)

A particularly interesting comment during a focus group interview provides a wrap-up note on this positive discussion of *Criterion* ACF:

*Traditionally we had to hand-write our essays. In this semester I could go to the computer lab for my writing classes while others still attend the traditional class. This is a really different experience.... My computational skills were improved significantly, my typing and composing essays in English.... Overall, it is a new and cool breeze [laugh]. My way of learning to write, this different way of learning to write in English really excites me.* (Huy, Focus Group 3)

**Negative evaluation.** Despite overall satisfaction with *Criterion* ACF, students occasionally cited issues they experienced when they used the feedback for revisions. The most frequently mentioned criticism of ACF is its limited elaboration on errors and the lack of opportunities to follow up on flagged errors with *Criterion*. This resulted in some limitations of putting the feedback to use during revision processes. The following excerpt from the interviews is quite representative:
Sometimes I made a revision, but the system only highlighted the error without giving any further comments regarding what actually went wrong and why it was wrong. I revised a lot of times but failed to produce a correct revision. At that time, I truly wanted to know what the problem was but had no way to find out. After several failed attempts, I really wanted to see what the correction should be, which was not provided by the system. (Yen, Focus Group 2)

The above comment tended to point to the generic nature of Criterion metalinguistic explanations, which is the focal point in the next citation from Nga (Focus Group 2):

*The machine [Criterion] is just a software. It provides very general comments about an essay which scored a certain band and that’s about it. As it was not able to individualize the feedback, two good essays of the same band score would receive exactly the same messages.*

For these students, genericity seemed to be associated with lack of usefulness as they experienced repeated failures to correct errors using generic feedback. In the focus group discussions, students often highlighted the need for extra and more detailed help, which they simply could not get from Criterion itself,

*Sometimes I tried to correct the errors, but what I received was just the error type. Criterion did not point to the exact location of the error, so my several attempts failed. If I kept failing, I would love to know what the correct form is, which was not provided by Criterion. It just kept saying I got an error.* [Ngan, Focus Group 2]

Even in some cases of specific feedback where Criterion provided a suggested alternative for students’ errors, students also raised the need to follow up if they did not fully understand why they should adopt a suggested change:

*With errors we were still unclear about, we could bring it up with the teacher. With the machine, however, we could not ask it why we should correct the errors that way.*

Furthermore, students complained about the absence of feedback on more complex linguistic features like word collocations and subtle word choices in their writing. The following
citation from Ngan (Focus Group 3), who expressed her lack of trust in Criterion corrective feedback on collocation errors, is quite typical of the comments made by the interviewed students:

\[ I \text{ feel like, it [Criterion] is not good at correcting errors related to collocations. It is not very accurate in addressing lexical issues.} \]

Such comments were similar to an opinion expressed by one of the think-aloud protocol students, An. She criticized Criterion’s failure to understand the content of her essay.

\[ \text{Some mistakes, I don’t like these because the technology don’t [sic] have enough intelligence to understand what I think, so some mistakes, some mistakes Criterion only checks based on the academic study….so it’s not useful in enlarging my content and showing me whether the content is suitable. (An’s stimulated recall interview)} \]

Coming back to the discussion regarding students’ levels of trust in Criterion’s ACF presented earlier, some interviewed students explained the sources of their distrust in Criterion ACF. Seven students mentioned that they found certain corrective feedback from Criterion questionable. The following comment was among the many similar of its type: “I see that some of my correct writing was flagged as wrong by Criterion” (Mai, Focus Group 1).

On the one hand, students questioned the correctness of Criterion error feedback, citing cases when the system was insensitive to varieties of English. For example, Nhi (Focus Group 1) reflected on her distrust in Criterion’s feedback on spelling errors related to British and American English alternatives.

\[ \text{Sometimes I find feedback from Criterion on certain words incorrect. I believed that the spelling of the words I wrote was correct, but I still got error messages from the machine. I mean alternative spellings in British and American English. The system seemed unable to recognize the words.} \]

Also relating to feedback on spelling from Criterion, Nhi from Focus Group 3 voiced her concern about the false positives she received,
There were cases when I got error messages on correct words in my essay, like error flags on words in Vietnamese we used in our essay.

Additionally, some others even challenged the authority of ACF, citing their feelings that Criterion ACF is similar to a peer’s feedback, by which they implied Criterion’s lack of authority and even validity as a resource for reference. This was the point made by Duong below:

It’s like, when I received the feedback from Criterion, I had the feeling that it was a friend’s feedback. When I got its feedback, I asked myself why Criterion provided that correction, but I could not talk to it directly. So I did not have a chance to argue with it, things like this highlighted error is actually not an error or why it didn’t flag that error in my essay. (Focus group 3)

**Criterion ACF in comparison with teacher corrective feedback.** Overall, students tended to be equally favourable about teacher and automated corrective feedback, commenting on the benefits of both sources. This section is divided into two parts to contrast ideas that are (a) in favor of the automated feedback through comments on its superior features, and (b) in favor of teacher feedback citing comments on the characteristics that make Criterion ACF inferior to teacher feedback.

**Preference for automated corrective feedback.** Students perceived the feedback they received from Criterion to be instant, accurate, detailed, systematic, and thus very efficient. During the group discussions, three students from Focus Group 2 explicitly expressed their preference for Criterion’s real-time and easily accessible feedback over the delayed teacher feedback, as cited below:

I found it [Criterion] efficient as after I submitted the essay, I could know straight away what errors I had made, so I found it more effective than waiting for the teacher’s feedback. (Duyen)

The essay was still fresh in my mind, so I could remember it well and knew what was wrong in my writing. If I waited for the teacher feedback, I would have forgot its content. I neither remembered what I had written nor realized the errors I had made. (Van)
Uyen commented on similar features as her peers did but also added that she especially favored automated corrective feedback over teacher feedback for its superior systematicity and accuracy.

I particularly liked Criterion’s instant feedback, because it could instantly provide feedback and suggest ways to correct my errors. It took much more time to receive the teacher’s feedback…. The program provided faster and better feedback. Sometimes, teachers can be wrong, while the machine’s feedback is standard and thus its feedback is correct. Also, the teachers sometimes missed errors in our essays, while the machine pointed out all the errors for us to revise.

Students from Focus Group 3 suggested that writing teachers usually missed a lot of errors in their learners’ essays because of work overload. They cited the advantage of error detection of an automated program like Criterion over the teacher who may be overwhelmed by too many essays.

As human beings, teachers can be overwhelmed by reading too many essays and therefore, they may miss a lot of errors. (Nhi, Focus Group 3)

Building up on Nhi’s point, another student in this group, Ngan, considered Criterion ACF to be more objective with more comprehensive coverage of different error types compared to the spontaneous error correction approach among writing instructors:

The teacher could miss errors in our essays, such as errors in comma usage if there are too many essays to read. On the other hand, the machine is programmed to give systematic feedback. It scans our essay from beginning to the end to detect errors in grammar, organization, vocabulary. Human beings are more easily affected by different things. They can be distracted by different things when they are reading an essay and may accordingly forget about an error they detected.

The comment coming from Duong in Focus Group 3 can sum up all of the arguments in favor of Criterion ACF, “The machine is objective when giving us feedback. It seems like the
system can detect and categorize errors in standard ways, without being interfered by meaning or emotions."

**Preference for teacher corrective feedback.** Compared to teacher feedback, *Criterion* ACF was at times considered too strict and inflexible, as admitted by Duong, despite her preceding comment about the benefits of *Criterion* ACF:

*It cannot get the meaning from our writing as a teacher does. It is like, an error is an error, while a teacher is more flexible.*

Added to this, five focus group students complained that they did not have a chance to seek further elaboration on the feedback or follow up on its comments due to the lack of two-way communication, a point highlighted by the student below:

*I think once we’ve got the feedback showing where the error is, it will be better to have the teacher’s further explanation because after all, the chance to have a conversation with the teacher about the error would help us understand the problem better. On the contrary, when we get the feedback on Criterion, there is always some distance and thus we cannot understand our error very thoroughly. It would be better with the teacher’s follow-up explanations.* (Van, Focus Group 2)

Another shortcoming quite frequently discussed among the students was the absence of automated feedback on higher order skills in writing, which only a writing instructor could provide,

*If I composed my essay on Criterion, it cannot provide me with useful feedback on the content. It only generates form-focused feedback, feedback on structures or grammar use. It does not do a good job at the content of the essays.* (Huy, Focus Group 3)

In students’ perceptions, *Criterion* helps them to correct minor and surface-level errors before they submit their essays to teachers for more comments on higher-level skills. The following comment from Mai emphasized the need for such feedback that could be given by her writing instructor,
The system only corrected spelling and grammatical errors, while the teacher provided comments on how to improve lexical choices, use cohesive devices more effectively ... In general, even after feedback from the system has been incorporated, we still receive comments from the teacher. (Focus Group 1)

A similar argument was made in the next comment from Yen of Focus Group 2, who highlighted the superiority of teacher written feedback on subtle vocabulary choices for idea expressions:

As long as we wrote grammatical sentences, the system wouldn’t...for example, there were equivalent structures which express the same idea, but as long as your chosen structure is grammatically correct, the system would not give you any feedback. But if your teacher is reading your draft, she may offer you some suggestions for better expressions of the same ideas.

The various comments made by learners in favor of teacher over Criterion ACF tended to accentuate the system’s lack of feedback that addressed more complex grammatical and lexical issues to facilitate more fine-grained choices in students’ writing. It is such feedback in combination with the chance for following up on errors that made teacher written corrective feedback superior to Criterion ACF in learners’ perceptions.

Taking the middle ground. From what students discussed in the four focus group interviews so far presented, it can be said that to learners, both automated and teacher written corrective feedback have pros and cons to be considered. According to the students, while teacher feedback could be more flexible and able to address higher-order skills in writing, the teacher can miss a high rate of errors in learner essays. On the contrary, students generally believed that Criterion can do a better job at generating feedback on surface level errors such as grammar, usage, and mechanics, especially spelling; yet its feedback is of a one-size-fit-all type without enough elaboration at times, thus making it not fine-tuned enough to learners’ actual needs. Therefore, the best policy, according to Nhi from Focus Group 3, was a combined assessment using Criterion feedback on grammar, usage, mechanics and content evaluation from the teacher coupled with the additional chances for learners to ask clarification questions, especially in large writing classes,
We can combine the two sources. We can learn from Criterion’s feedback on structures, grammar, while benefiting from content feedback provided by the teacher.

**Expectations about the explicitness of automated feedback.** Interviewed students varied significantly in their expectations about how explicit Criterion ACF should be, which reflects variable needs for scaffolding in revisions among the learners. This ranges from expectations about highly direct feedback, as in the following comment when the student reflected on Criterion ACF about articles and punctuation,

*For me, what Criterion should provide correct revised forms for is its feedback on articles and punctuation, punctuation in particular... The program should do the job of correcting the errors for us, as it takes a lot of time for us to check the essay while we are under time pressure during tests.* (Ngà, Focus Group Interview 2)

Thanh, on the other hand, suggested that Criterion should have more explicit metalinguistic explanations through the provision of specific cues for correction:

*I don’t really expect Criterion to correct the errors for me, but it can give us hints, like more specific cues.* (Thanh, Focus Group Interview 2)

Similar opinions were those expressed by Huong and Tram, who preferred Criterion to offer some alternative corrections from which they could select the one that worked best for their essays,

*I think it [Criterion] can give us clues for corrections, such as an answer key in the box.* (Huong, Focus Group 1)

Tram extended on her peer’s comment by adding that Criterion could be more flexible in offering suggested changes,

*It should not provide one fixed correction. Instead, it should be up to us for the best option for a revised form. It should not dictate a revised form and ask us to do this or that. What it can do is giving us options to choose from.*
In contrast to the above ideas, Nhung and Khanh expressed concern over their loss of independence over time if they depend on *Criterion* ACF for corrections of their errors. Therefore, they prefer implicit over explicit feedback from the software,

*If I receive very specific suggested ways for correction, I will become overdependent. I mean, how we will revise the errors has to come from our own work and thinking. In fact, if we become overdependent on the machine, our skills will become worse over time.* (Nhung, Focus Group 4)

*After some time using the software, if it provides me with direct correction of my errors, I will have trouble writing my own essay when I compose using the paper-and-pencil way, or when I have to think on my own without it.* (Khanh, Focus Group 4)

In summary, what emerged from the interviews were not only learners’ perceptions of the automated corrective feedback they received from *Criterion* but also the revelations about their own needs and expectations from the feedback. Generally, students had similar perceptions about the advantages and shortcomings of *Criterion* ACF. Their arguments were also equally divided between those in favor of teacher feedback and *Criterion* ACF. Despite some concerns about occasional false positives and incorrect feedback they received, students generally trusted *Criterion* for its real-time and satisfactorily accurate feedback on surface-level issues in grammar, usage, and mechanics. Their other concerns rested with *Criterion*’s lack of feedback on more complex issues in writing as well as the absence of chances for elaboration on errors. During the discussions, students expressed varied expectations about the explicitness of *Criterion*’s feedback, ranging from highly explicit error correction to implicit identification. This reflected a major shortcoming currently experienced with *Criterion* ACF, which is the demand for more fine-tuned feedback that could accommodate individualized needs among the learners.

**Chapter Summary**

This chapter focused on issues viewed from learners’ perspectives regarding their engagement, revisions, and perceptions of *Criterion* automated corrective feedback, as well as their observed accuracy changes over the five-month period. Findings on students’ engagement with and subsequent revisions using *Criterion* ACF revealed more perfunctory than substantive processing of the feedback. It was also found that unhelpful feedback tended to initiate more
extensive engagement among the learners. Following their engagement, students performed mostly local textual operations to revise their texts and were modestly successful at correcting the errors Criterion pointed out to them, with 56.2% success rate and 7.1% retention of their original texts in response to false positives. Generally, the surface-level nature of the feedback discussed in Chapter 4 and students’ superficial engagement with it, along with simple and straightforward textual operations presented in this chapter suggest a strong connection between the nature of the feedback, the types of errors addressed and how students responded to it.

Added to the above findings, learners’ changed accuracy was calculated to provide answers for Research Question 4. Over the studied period, there was no significant treatment effect of the use of Criterion ACF on experimental students’ accuracy across the five morphemes for which they received most error tags from Criterion, including article, third person singular, copula, plural, and comma usage. Their overall accuracy did not improve significantly either.

Research Question 5 revolves around the qualitative analyses of the focus group interviews and stimulated recalls which revealed students’ overall positive perceptions about Criterion ACF. Their satisfaction with the feedback, as discussed during the interviews, was thanks to Criterion’s immediacy, systematicity, and comprehensiveness in generating feedback. Students also discussed issues that caused concerns about Criterion ACF, including the lack of elaboration in metalinguistic explanations and absence of feedback on more complex grammatical issues. The findings also pointed out learners’ varied expectations about how explicit the feedback should be and their evaluations of teacher compared to automated corrective feedback. When asked about the use of these two sources of feedback, students expected to use Criterion for its feedback on surface level errors in grammar, usage, and mechanics, while believing that teacher feedback would make up for the other missing yet equally important dimensions of learners’ L2 writing.
Chapter 6
MULTIPLE-CASE STUDY

Overview of the Chapter

Ellis (2010) argued that individual learner factors interact with contextual factors “to mediate between the CF that learners receive and their engagement with the CF and thereby influence learning outcomes” (p. 339). This corroborates the interactionist perspective on second language acquisition which investigates the interaction between the linguistic environment (i.e., the input received by the learner) and his/her internal cognitive processes such as noticing. Motivated by these comments, together with differential accuracy gains recorded for different learners in the main study, this multiple-case study aims to zoom in on the interaction between individual learner factors and their engagement with as well as revision activities using Criterion ACF, and how this interaction relates to their accuracy change over the studied period.

This chapter has four main sections. Starting with the rationale for the choice of a case study approach and its research aims, the second section presents how the cases were selected and how their data were analysed. In the third section, three narratives about the three selected cases are developed, each case being discussed at length with a focus on prominent themes that emerged from the analyses of multiple data sources. The themes for each learner are supported with specific examples and illustrations from their data, sometimes with reference to the other two case-study participants for comparison and contrast. Part four of this chapter includes a summary of findings which combines and synthesizes the three cases’ results before the cross-case analysis is provided.

PART 1. Rationale and Research Aims

Rationale for Using a Case Study Approach

The current case study is inspired by findings in previous research on teacher and automated feedback which indicated that different learners may perceive and benefit differently from teacher feedback (e.g. Ferris, Liu, Sinha, & Senna, 2013; Han & Hyland, 2015; Kim & Bowles, 2019; Qi & Lapkin, 2001; Storch & Wigglesworth, 2010; Zheng & Yu, 2018) and automated feedback (e.g. Zhang, 2017; Zhang & Hyland, 2018), which seems to be corroborated by the findings from the main study. However, a study which includes a few dozen students as in the main study does not allow for detailed information to reveal deep understanding about individual learners. Therefore, there is the need for a case study that can zoom in on individual
learners’ engagement with, response to, and perceptions of automated corrective feedback. Such an approach also allows for cross-case comparison that reveals similarities and differences between learners.

Yin (2018) proposed that a “case study may be part of a larger, mixed-methods study. The main investigation may rely on a survey or other quantitative techniques, and your case study may help to investigate the conditions within one of the entities being surveyed” (p. 64). From all of the findings presented in Chapters 4 and 5, there is a missing yet significant link between what learners actually did (the unknown piece) when coming into contact with Criterion ACF (whose features have been made known in Chapter 4) and their observed accuracy changes in writing across three different time points of the experimental period (a key finding in Chapter 5). Case-study research is characterized by the focus on specific cases to make them understandable with the central role of the researcher being the gatherer of interpretations to make sense of the data and to construct knowledge (Creswell, 2013). Case study research, therefore, can help reveal the missing link mentioned above.

Moreover, to investigate the key issue above, we need more than one case to make comparison between what works and what does not work in terms of the efficacy of Criterion automated corrective feedback for students’ revisions and longer-term acquisition of targeted morphemes. The choice of a multiple-case design follows Yin’s (2018) comment that “the design of multiple-case studies follows an analogous logic where the choice of the cases predicts contrasting results but for anticipatable reasons” (p. 64).

The decision to use a multiple-case design in this study is founded on both practical and theoretical grounds. It promises to provide insights into what happened as learners encountered Criterion feedback: their response, perception, corresponding revisions and revising outcomes, as well as cognitive and affective engagement with the system’s ACF. By using more than one case who demonstrated different learning outcomes despite having equal access to Criterion automated corrective feedback, the multiple case study hopes to contribute critical insights not yet offered by most available studies on automated feedback. It also aims to find some possible causal links between engagement with the feedback and the changes or absence of changes in writing accuracy over time using Criterion automated corrective feedback. Multiple data sources help to build a complete picture and a stronger case for interpretations regarding why automated corrective feedback has or has not facilitated different learners’ improved accuracy in writing.
Research Aims

Findings in the main study showed that there was no significant effect of using Criterion ACF on learners’ acquisition of grammatical forms over time for the whole group; however, individual learners experienced variability in terms of accuracy change after the five-month period. Based on these results, a multiple-case study was developed as an embedded but significant extension which helps provide more in-depth information for Research Questions 3 and 4. In the main study, RQ3 examines all the students’ engagement with Criterion ACF and their subsequent revision practices through the description of their engagement levels and the tabulation of text operations and revision outcomes, as well as the whole group’s accuracy change over the studied period.

For the case study, however, these two research questions are extended to focus on the “how” and “why” aspects that characterize case studies in general (Yin, 2018). The core variables of case study research are individual behaviors over a period of time that lead to a sequence of events such as performance outcomes (Woodside, 2010). Therefore, instead of focusing on the holistic group results, this embedded multiple-case study offers nuanced understanding about individual differences in students’ engagement and response to the automated corrective feedback and thus may provide better insights into possible explanatory factors that may have contributed to observed differential learning outcomes. An in-depth investigation of multiple data sources from selected cases helps reveal the underlying processes and mechanisms that may have impacted the efficacy of learners’ revisions and their subsequent acquisition of the target language forms. In other words, the case studies aim to deepen the understandings gained from RQ3 in the main study by addressing the three sub-questions below.

3a. What are the similarities and differences between the three selected cases in their engagement and revisions using Criterion ACF?

3b. What are the reasons for learners’ failed attempts at certain revisions using Criterion ACF as a resource?

3c. What are the factors that contribute to differential engagement and revision patterns observed among the selected cases?
PART 2. Case Selection

Shaping the Case Study

Seeking to realize the aims presented above, the term that best describes the type of case study approach employed in this research is an explanatory case study (Yin, 2018) with the intent to best understand the observed differential learning outcomes for the participating learners. Adopting Woodside’s (2010) term, the objective of the current case study research is explanation which aims to answer the question “why”. In Stake’s (1995) definition, what Yin called “exploratory” can be termed an instrumental case study which indicates the choice of a case or some cases “to understand a specific issue, problem, or concern” (p. 98). Also, as learners have been observed in this and other studies to differ in their perception, engagement, and revision processes (Han & Hyland, 2015; Qi & Lapkin, 2001; Storch & Wigglesworth, 2010), the focal research problem being pursued is how they are different and how that difference relates to their acquisition of the morphemes most frequently targeted by Criterion and their overall accuracy. This drives the need for cross-case comparisons, and the next question becomes “how many cases?” As noted by Creswell (2013), researchers often select no more than five cases. This decision helps case study researchers address a challenge he pointed out about multiple-case studies: that the depth of analysis is diluted by the choice of more cases.

Selection of cases. The availability of “clearly identifiable cases with boundaries” (Creswell, 2013, p. 100) is a prerequisite that allows a case study to provide in-depth understandings about the cases or to compare more than one case. To address the research aims outlined above, three specific cases were selected following Yin’s (2018) two-phase selection approach.

Phase 1. In this phase, quantitative and qualitative data about all of the participating students in the experimental group were collected and screened, making available 43 prospective candidates. Each of these students satisfied the basic condition of being part of the intervention. They attended all of the three practice sessions on Criterion with at least two of the practice essays having a revised draft. They also needed to have taken the pre, post, and delayed post-tests in weeks 2, 13, and 15 of the semester.

Guided by the research aims, the data available for all the 43 plausible candidates were screened and then these candidates were divided into two sub-groups depending on their different patterns of change using their observed accuracy measures over the studied period (i.e., pre, post,
and delayed posttest accuracy scores). More specifically, baselines at pre-test and their gain scores at post and delayed post-tests (both overall and specific accuracy scores) from the quantitative analyses in the main study were examined to divide students into those who experienced increased accuracy and others who had decreased/unchanged accuracy after five months using Criterion ACF as part of their writing course.

**Phase 2.** In the second phase of case selection, two additional criteria were operationalized to narrow down and finalize the number of case study participants. Firstly, of the 43 prospective cases bounded by the criterion in Phase 1, selected cases had to be among the most productive participants who contributed the richest information about engagement with Criterion ACF. This is to realize the aims set out earlier to seek as deep understandings about individual learners’ engagement and response to Criterion ACF as possible. “Most productive participants” indicate those who had the richest data through the test and practice essays they wrote, the number of revised drafts they submitted on Criterion, as well as their think-aloud protocols and interviews. More specifically, for the shortlisted participants, besides essays from pre, post, and delayed post-tests as well as first and revised drafts from the three practice sessions, they were also involved in producing TAPs where they verbalized a think-aloud protocol to verbalize thoughts as they worked through their revisions to the first drafts during at least one of the practice sessions. Their on-screen interaction with Criterion was also captured. Based on this criterion, fourteen cases were shortlisted. Secondly, for data triangulation, students’ recorded think-aloud protocols had to be followed by stimulated recall interviews to seek clarifications on verbalized thoughts. Three cases were selected based on their ability to satisfy all the criteria set out for the multiple-case study. Brief profiles of these three learners are presented in the next section.

**Brief profiles of cases.** With females forming the vast majority of the sample, all of the three selected participants were female as there were no male participants who satisfied all of the selection criteria. Among these three students, two cases followed a similar trajectory as their accuracy decreased gradually while that of the third one improved over the studied period. Table 6.1 presents the data of three selected cases in terms of their measurement of overall accuracy across pre, post, and delayed post-tests.
Table 6.1

Accuracy Levels of Three Cases

<table>
<thead>
<tr>
<th>Case</th>
<th>Pre-test score (%)</th>
<th>Post-test score (%)</th>
<th>Delayed post-test score (%)</th>
<th>Observed accuracy gains?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>79</td>
<td>83</td>
<td>87</td>
<td>Yes</td>
</tr>
<tr>
<td>Nhien</td>
<td>95</td>
<td>91</td>
<td>84</td>
<td>No</td>
</tr>
<tr>
<td>An</td>
<td>92</td>
<td>89</td>
<td>83</td>
<td>No</td>
</tr>
<tr>
<td>Whole group</td>
<td>81</td>
<td>83</td>
<td>81</td>
<td></td>
</tr>
</tbody>
</table>

All of the three selected cases came from the same experimental group. Therefore, they wrote essays in response to the same pre, post, and delayed post-test prompts for the three test sessions (see Appendix A for the specific writing prompts). According to Table 6.1, what differentiates the three cases was the fact that Trang started off with a lower accuracy rate (79%), while the other two cases had much higher accuracy baselines (above 90%). Over the five-month period, however, Trang followed a significantly different, if not opposite, developmental trajectory compared to those of Nhien and An. While the former improved in accuracy, making a gain score of 4% in each of the two subsequent tests, the two latter cases experienced gradual decreased accuracy over the same studied period.

The data set for each selected case can be found in Table 6.2 below, and the section that follows describes how these multiple data sources were analyzed.

Table 6.2

Data Sources of Three Cases

<table>
<thead>
<tr>
<th>Case</th>
<th>Test essays</th>
<th>Practice essays</th>
<th>Think-aloud protocol</th>
<th>Stimulated recall interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>Pre, post, delayed post-test essays (3)</td>
<td>Practice 1: 3 drafts Practice 2: 3 drafts Practice 3: 1 draft (no revised essay)</td>
<td>1 recording (Second practice session)</td>
<td>1 week after the second practice</td>
</tr>
<tr>
<td>Nhien</td>
<td>Pre, post, delayed post-test essays (3)</td>
<td>Practice 1: 2 drafts Practice 2: 2 drafts Practice 3: 2 drafts</td>
<td>1 recording (First practice session)</td>
<td>1 week after the first practice</td>
</tr>
<tr>
<td>An</td>
<td>Pre, post, delayed post-test essays (3)</td>
<td>Practice 1: 3 drafts Practice 2: 3 drafts Practice 3: 3 drafts</td>
<td>2 recordings (Second and third practice sessions)</td>
<td>1 week after the second practice</td>
</tr>
</tbody>
</table>
**Data analyses.** Besides the analysis procedures in the main study (see Chapter 3), data sources used for the multiple-case study related to the three selected students were examined at greater depth for more insights. Regarding learners’ first and revised drafts in practice essays, each revision episode by case study participants was examined for both affective and cognitive engagement. Following the operationalizations of the concept of engagement in previous research (e.g., Ellis, 2010; Han & Hyland, 2015; Zheng & Yu, 2018), affective engagement refers to students’ attitudes or perceptions, which is demonstrated in their judgement, affect, evaluation, appreciation, or trust/distrust in Criterion ACF. Data about students’ affective engagement came from their stimulated recall interviews as well as occasional observations found in TAP recordings.

Secondly, cognitive engagement is characterized by the cognitive investment in the processing of the feedback (Ellis, 2010). Cognitive engagement is demonstrated in learners’ depth of feedback processing, which is operationalized as extensive or limited in this study depending on their noticing of the feedback (substantive vs. perfunctory) as well as the kind of cognitive strategies employed to revise (e.g. looking up references, guessing correct forms, drawing on stored metalinguistic knowledge, or seeking peer/teacher help). These coding steps had been used in the main study with the data provided by the fourteen TAP students’ think-aloud recordings. However, an extra analysis step taken in the multiple-case study was that each cognitive engagement episode in the think-aloud protocol was cross-examined with the corresponding first and revised texts. This was further cross-checked with any relevant ideas expressed during the stimulated recall interviews.

In order to have a better indication of the case study participants’ writing as well as additional evidence of improvement if any, learners’ test essays were investigated to measure text complexity and fluency. Thus, there were four additional measurements. Firstly, pre, post, and delayed post-test essays of the three cases were submitted to vocabkitchen.com/profiler/awl to extract three specific measures of text complexity. These include the popular indicator of syntactic complexity, ratio of clauses per T-unit (see Table 6.3); the type-token ratio (TTR) statistic for lexical diversity (see Table 6.4); and the percentages of academic words referenced to Coxhead Academic Word List (AWL) for lexical sophistication (see Table 6.5). More specifically, type-token ratio provides information about lexical richness or variety. This is calculated by “the total number of UNIQUE words (types) divided by the total number of words (tokens) in a given segment of language” (“Activity 4: Complexity in oral vs. written language,” n.d.) Coxhead AWL contains frequently found words in academic texts but excludes the most common 2000 words of English in the General Service List. Therefore, the words in AWL.
reflect vocabulary used in academic contexts such as tertiary education levels.

Table 6.3

**Syntactic Complexity through Ratio of Clauses per T-unit**

<table>
<thead>
<tr>
<th>Case</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Delayed post-test</th>
<th>All test essays combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>1.5</td>
<td>1.7</td>
<td>1.7</td>
<td>1.6</td>
</tr>
<tr>
<td>Nhien</td>
<td>1.5</td>
<td>1.7</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>An</td>
<td>1.4</td>
<td>2.2</td>
<td>1.5</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 6.4

**Lexical Diversity through Type-token Ratio (TTR) in Percentages**

<table>
<thead>
<tr>
<th></th>
<th>TTR Pre-test (%)</th>
<th>TTR Post-test (%)</th>
<th>TTR Delayed Post-test (%)</th>
<th>TTR with all test essays combined (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>56</td>
<td>62</td>
<td>49</td>
<td>40.5</td>
</tr>
<tr>
<td>Nhien</td>
<td>43</td>
<td>42</td>
<td>51</td>
<td>35.5</td>
</tr>
<tr>
<td>An</td>
<td>48</td>
<td>51</td>
<td>53</td>
<td>35.8</td>
</tr>
</tbody>
</table>

Table 6.5

**Lexical Sophistication through Percentages of Academic Words Referenced to Coxhead Academic Word List**

<table>
<thead>
<tr>
<th>Case</th>
<th>Pre-test (%)</th>
<th>Post-test (%)</th>
<th>Delayed post-test (%)</th>
<th>All test essays combined (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Nhien</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>An</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

What can be seen from Tables 6.3, 6.4, and 6.5 is that the three students did not differ significantly in terms of both syntactic and lexical complexity. Trang used slightly more diverse vocabulary in her essays with slightly greater vocabulary sophistication judged by TTR and AWL reference. Other than that, Nhien and An were almost identical in their complexity measures of TTR and lexical sophistication.
For fluency, essay length was calculated and presented in Table 6.6 which provides the three case study participants as well as the whole group’s essay lengths at pre, post, and delayed post-tests. With the fluency measure using word counts, Trang generally wrote shorter essays than Nhien and An. Her longest essay was in the delayed post-test, with 375 words, which is just slightly shorter than that of Nhien at the same time point. On the other hand, An tended to write longer essays over time, with 438 words in her delayed post-test composition.

Table 6.6

*Essay Lengths of Three Selected Cases*

<table>
<thead>
<tr>
<th>Case</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Delayed post-test</th>
<th>Averaged essay length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trang</td>
<td>326</td>
<td>302</td>
<td>375</td>
<td>334</td>
</tr>
<tr>
<td>Nhien</td>
<td>279</td>
<td>525</td>
<td>385</td>
<td>397</td>
</tr>
<tr>
<td>An</td>
<td>328</td>
<td>393</td>
<td>438</td>
<td>387</td>
</tr>
<tr>
<td>Whole group</td>
<td>218</td>
<td>293</td>
<td>299</td>
<td>270</td>
</tr>
</tbody>
</table>

As seen in Table 6.1, Trang had a significantly lower accuracy score at her pre-test compared to Nhien and An. However, while the former experienced gradual accuracy gains over time, the two latter cases followed a downward trend in their accuracy over the studied period and their delayed post-test scores were lower than that of Trang. Alongside these accuracy changes, Trang’s lexical diversity decreased from the pre-test to the delayed post-test, which is also the case for her lexical sophistication measure. On the contrary, both Nhien and An had slightly increased lexical diversity and sophistication at the later time points. One common thing of all of the three participants is that they tended to write longer essays over time. These differences in the CAF measures of the three case study participants will also be taken up during the discussion presented in Chapter 7 of the current research.

**PART 3: Individual Cases’ Findings**

This section presents the findings related to individual cases. The first sub-section of each case is a detailed profile which was based on the data from each student’s filled-out forms containing personal information when they took the pre-test, class observation notes, and informal personal interviews. This is followed by quantitative findings about the error codes they
received, their engagement episodes and revision outcomes which are triangulated with qualitative findings from the learner’s think-aloud protocol recordings and stimulated recall interviews to provide a more complete picture about each case.

Case 1. Trang, highly efficient reviser through limited engagement with Criterion ACF

Profile. Class observations suggested that Trang was willing and enthusiastic about responding to teacher-initiated questions. She had learned English for eleven years, primarily during formal classes as a compulsory school subject. She did not have any prior experience using an automated writing evaluation tool. After graduation from the bachelor’s degree, she planned to be a high school English teacher. She had never travelled to an English-speaking country.

Trang had overall accuracy scores of 79%, 83%, and 87% across the pre, post, and delayed post-tests, respectively (see Table 6.1). This indicates some improved accuracy in her writing over the five-month period. In terms of essay length, as seen in Table 6.6, Trang wrote 326, 302, and 375 words in her pre, post, and delayed post-tests. On average, she composed 334 words per essay compared to the whole group’s average 270 words across the three tests. In terms of complexity measures, Tables 6.3, 6.4, and 6.5 show that the syntactic complexity in her essays increased slightly from 1.5 in the pre-test to 1.7 in the post and delayed post-tests, while both her lexical diversity and sophistication measures decreased by 7% and 2% from the pre-test to the delayed post-test, respectively.

Turning to Criterion assessments of Trang’s three first drafts of her practice essays, Trang’s essay lengths and her normed error rates (i.e. the rate of error per hundred words) are provided in Table 6.7 below. As seen from this table, there was an incremental increase in the length of Trang’s practice essays over time. Despite the increase in word counts, Trang’s normed error rates based on Criterion’s error detection system decreased over the same period, with 1.08% at the third practice session in Week 11 of the semester.

Table 6.7
Word Counts and Normed Error Rates in Trang’s Practice Essays

<table>
<thead>
<tr>
<th>Measure</th>
<th>Practice essay 1 (Week 5)</th>
<th>Practice essay 2 (Week 8)</th>
<th>Practice essay 3 (Week 11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word counts</td>
<td>326</td>
<td>413</td>
<td>462</td>
</tr>
<tr>
<td>Normed error rate*</td>
<td>2.76</td>
<td>2.18</td>
<td>1.08</td>
</tr>
</tbody>
</table>

*Note. Normed error rate is the number of errors per hundred words.
Table 6.8 presents *Criterion*-generated tags by error type Trang received on her first drafts in three practice sessions. Trang received a total of 23 error tags, most of which were Missing or Extra Article (nine errors), followed by errors related to comma usage (six tags, five of which were either redundant tags or false positives). Other errors included two confused words (one being a false positive and another about the misuse of *a* and *an*), two fragments (one false positive), one subject-verb agreement, one hyphen, one spelling (false positive), and one determiner-noun agreement error.

In general, Trang’s errors were mostly about usage and mechanics, while few errors in grammar were detected by *Criterion*. It is also worth noting that one third of the total error tags she received were false positives.

In total, Trang conducted two think-aloud protocols for two practice sessions 1 and 2. However, the recording for practice session 1 was not used as data because when she came to the recording room for her TAP procedures, she had finished revising her first draft. Trang’s final recording for this session was just the verbalized thoughts when she was processing *Criterion* feedback on her second draft, making it inappropriate for research purposes. Used as data for the research was Trang’s recorded TAP during the second practice session, which was followed by a stimulated recall interview. Her first and revised drafts for the first two practice sessions were also coded for revision patterns, but her third practice essay had no revised draft, although the number of *Criterion* error tags in it and error counts were tabulated and included for the analyses.

<table>
<thead>
<tr>
<th>Category</th>
<th>Error Total: 23</th>
<th>Error Types</th>
<th>False Positives: 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>3 errors</td>
<td>2 Fragments</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Subject-Verb Agreement</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>12 errors</td>
<td>1 Determiner Noun Agreement</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 Missing or Extra Article</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Confused Words</td>
<td>1</td>
</tr>
<tr>
<td>Mechanics</td>
<td>8 errors</td>
<td>5 Missing Comma</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Extra Comma</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Hyphen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Spelling</td>
<td>1</td>
</tr>
</tbody>
</table>
Engagement episodes and revision patterns. Trang spent a total of 25 minutes revising her essay written during the second practice session. Of these 25 minutes, only about ten minutes were devoted to processing the automated corrective feedback generated by Criterion, and the remaining time was spent on self-initiated revisions on content as well as grammatical and lexical choices in her first draft. The following section is thus divided into two subparts, the first of which relates to the amount of time making Criterion-initiated revisions, and the second subpart presents data on self-initiated revisions.

Criterion-initiated engagement episodes. Tables 6.9 and 6.10 provide a break-down of error tags on Trang’s Practice Essays 1 and 2, respectively. As Practice Essay 2 also had the TAP recorded, more information is available about Trang’s engagement with Criterion ACF in Table 6.10. No data about revisions for Practice Essay 3 are provided because this essay did not have a revised draft.

In addition, in Table 6.10, coded engagement strategies are presented as substantive or perfunctory, the terms borrowed from Qi and Lapkin’s (2001) study, as well as revision outcomes. As earlier defined in Chapter 3, the perfunctory noticing episodes indicate instances when learners noticed but employed superficial strategies to engage with Criterion ACF, or their noticing was simply the mention of the error without any follow-up strategy. The ‘superficial engagement strategies’ as driven by the data of the current study included simple adoption of Criterion’s suggested change, guesses of the correct forms (without further efforts to double-check them), quick revision of the form using Criterion error tag as a reminder (or ‘red flag’ as occasionally used throughout this study), or simply noticing without attending to the error.

In contrast, substantive noticing episodes are instances when a feedback point is noticed (mentioned in the TAP) followed by a reason and/or a follow-up action to find references for checking the revised forms. This includes (a) drawing on one’s stored metalinguistic knowledge; (b) getting cues from Criterion error explanations; (c) looking up online resources; (d) seeking peer/teacher feedback; or (e) translating one’s writing into L1. Details and examples for each of these categories can be found in Appendix F.
<table>
<thead>
<tr>
<th>Category</th>
<th>Error type</th>
<th>Error verification</th>
<th>Revision outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage (6)</td>
<td>1 Determiner Noun Agreement</td>
<td>Correct error code</td>
<td>Correct Revision</td>
</tr>
<tr>
<td></td>
<td>4 Missing or Extra Article</td>
<td>1 Correct error code</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 False positives</td>
<td>3 Retention of Correct Form</td>
</tr>
<tr>
<td></td>
<td>1 Confused Words</td>
<td>False positive</td>
<td>Retention of Correct Form</td>
</tr>
<tr>
<td>Mechanics (3)</td>
<td>1 Spelling</td>
<td>False positive</td>
<td>Retention of Correct Form</td>
</tr>
<tr>
<td></td>
<td>1 Missing Comma</td>
<td>False positive</td>
<td>Retention of Correct Form</td>
</tr>
<tr>
<td></td>
<td>1 Hyphen Error</td>
<td>Correct error code</td>
<td>No change (unjustified)</td>
</tr>
</tbody>
</table>
According to Tables 6.9 and 6.10, in response to the total 18 error tags, Trang made eight correct revisions, eight instances of retention of correct forms when faced with false positive feedback, one unjustified ‘no change’ to a correct error flag, and only one incorrect revision.

For a better understanding of the coding process and how multiple data sources were connected in this process, Table 6.11 presents three exemplary feedback processing episodes on three errors with their accompanying metalinguistic explanations from Criterion, relevant excerpts from Trang’s recorded think-aloud protocol, her engagement in each episode, as well as the revised forms and revision outcomes.
Table 6.11
Trang’s Engagement Episodes and Revision Outcomes Based on Criterion ACF

<table>
<thead>
<tr>
<th>Processed error</th>
<th>Excerpt from recorded think-aloud protocol</th>
<th>Engagement level</th>
<th>Revision outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>First draft: by tricking others with something wrong in progressive technology.</td>
<td>Umm for for the progressive technology. Ah I will change it into just progressive technology, not the. Delete the.</td>
<td>Perfunctory</td>
<td>Incorrect Revision</td>
</tr>
<tr>
<td>Criterion: You may not need to use an article here.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revised: by tricking others with hacking personal banking information in progressive technology.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First draft: They will earn a lot of money which provides you easier and happier life.</td>
<td>Next, umm, they will earn they will earn a lot of money which provides you an easier and happier life. I think it’s not wrong so I won’t change it…. an easier life.</td>
<td>Perfunctory</td>
<td>No Change to False Positive</td>
</tr>
<tr>
<td>Criterion: You may not need to use an article here.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No revised form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First draft: Thus some people firmly believe that it is a measure of success.</td>
<td>I miss some commas. Thus OK thus I will I will put a extra comma here. Thus, comma, some people firmly believe…OK.</td>
<td>Perfunctory</td>
<td>Correct Revision</td>
</tr>
<tr>
<td>Criterion: You may need to use a comma after this word.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revised: Thus, some people firmly believe that it is a measure of success.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Across all of the nine coded engagement episodes, Trang was perfunctorily engaged with the feedback through employing the strategies of adopting Criterion’s suggested changes, quickly revising the form using ACF as a red flag, or simply skipping to the next error after cursorily noticing it. However, as seen from Table 6.11, Trang seemed to be very effective in both responding to false positives from Criterion and revising the errors. In both sessions, she received a total of eight false positives on which she exercised sound judgment before appropriately responding by retaining the original text. Specifically, seven out of eight false positives were appropriately responded to. The remaining one was a redundant change about adding a comma before the last item in a list:

First draft: They might be courage, creativity, pride, **kindness** and sometimes as simple as be able to spend lot of time with family and friends.

Criterion: You may need to use a comma after this word.

Trang decided to adopt Criterion’s suggested change with the insertion of a comma after “kindness”. Interestingly, the stimulated recall interview revealed that despite her amendment, Trang still believed that no change was necessary:

Researcher: So it asks you to use the comma after "kindness" and before "and", do you agree with this idea in this feedback from Criterion?

Trang: Not really because according to my former teacher, he said that with a comma or without a comma at this space is not necessary.

In her interview, Trang explained her resistance to Criterion’s suggested change by citing her teacher’s instruction, implicating higher trust in the instructor than automated feedback. However, the motivation to get a higher Criterion score drove her to correct the error, as she admitted,

Researcher: But why did you make that change, in the end?

Trang: I hate things but I like to to…to correct it in a further, in a further way to get a higher mark.

---

3 For each draft students submitted on Criterion, the system generated an automated holistic score on a scale of 1-6
Comparison of Trang’s think-aloud protocol and her stimulated recall interview also suggests that her superficial engagement through perfunctory noticing and employment of superficial cognitive engagement strategies in processing Criterion ACF was justified. There seemed to be little need for extended engagement with the feedback for arrival at correct revisions. In all of the coded episodes in her TAP, she demonstrated limited engagement with the feedback, but triangulation with data from her interview reveals that these perfunctory episodes were sufficient for making a decision on a revised form, at least for a learner like Trang who seemed to quickly “notice-with-understanding” the errors pointed out to her. The following example of Trang’s response to an error tag on missing额外 article illustrates this.

First draft: As a student who has not paid much attention to money pressure yet depends on family¹, …

Criterion: You may need to use an article before this word.

On the surface, she verbalised in the recorded TAP, “I have missing…missing extra articles. Family, yep depends on attitudes students who have not paid much attention to money pressure yet and depends on the family. OK. I will change it into the the family. Depends on the family.” Then she came up with the correct revised form by adding the before family. Trang’s verbalization during this revision episode gave the impression that there was overdependence on Criterion’s suggested change. However, it actually was a case of well-informed decision based on reference to the learner’s own stored metalinguistic knowledge, something that did not come out from the think-aloud protocol. In the interview, Trang’s reliance on self rather than on the suggestion from Criterion became clear,

Because I mentioned that as a student who has not paid much attention and I, I, I prefer that the family… There is the family of this student I mentioned before [emphasis added], so I use the the before family

Observation of the engagement episodes also showed that the majority of Trang’s decisions were quick adoptions of Criterion-suggested changes. Therefore, during the stimulated recall, Trang was asked whether she felt sure about the revised forms. In response to this question, Trang confirmed that she actually referred to her previously learned metalinguistic knowledge to arrive at revision decisions instead of basing her work on wild guesses or merely
on Criterion’s suggested changes: “After I, after the Criterion helps me, I certainly, I remember the old knowledge and I answer, and I trust it’s the correct way so I try it.”

In summary, with a total of 18 Criterion feedback points, Trang made 16 well-justified responses. She only incorrectly revised an error in the missing/extra article type, a notoriously problematic linguistic feature for English language learners (Han, Chodorow, & Leacock, 2006; Jordan, 2004). On only one occasion, she made no change to an error correctly flagged by Criterion. In response to false positive error tags which accounted for a high proportion in Trang’s data, she retained the original correct forms in response to seven out of eight false positives and made one redundant change for the remaining one. Generally, Trang demonstrated competent handling of the errors through reliance on sound judgement of the feedback and making appropriate changes where needed. In order to do this, she only needed to deploy perfunctory engagement strategies to process the provided error tags and came up with justified decisions.

**Self-initiated revisions.** Trang spent 15 minutes revising aspects other than the errors pointed out by Criterion. Her self-initiated revisions not only attended to higher order skills in writing, including content and language register, but also linguistic aspects that Criterion had failed to address in her writing. In those 15 minutes, Trang made three changes in word choice, three stylistics amendments, two major reformulations of ideas, two grammatical self-initiated revisions, and one language register adaptation in her essay. All of these self-revision episodes are further analyzed in the next parts.

**Higher-order skills.** In her interview, when asked about revision activities she normally conducted when writing in English, Trang commented,

> Because I always check… I mean I am not always satisfied with something ... with the task that I wrote, so I checked it if I want to be more … more formal.

In the above comment, Trang showed her awareness about the formal register of academic writing, as illustrated in the following excerpt from her TAP when she replaced and with as well as:

*It strikes me that the one who has great knowledge and owns a very proficient professional. Umm. Who has great… OK. For and here maybe I will change it into as well*
as for more formal…formal way. OK. As well as. Who has great knowledge as well as owns a very proficient professional is the one who is paid well as they deserve it. OK

In a similar vein, Trang was thinking about stylistics when she commented that the phrase a lot of money was used repetitively in her essay, which resulted in the substitution of a lot of with plenty of in the following excerpt from the think-aloud protocol recording:

There’s no denying the fact that anyone who earns a lot of money is successful. Umm I think a lot of money will repeat many times so I will change it into ahh plenty of… earns plenty of money, maybe, plenty… earns plenty of money is successful.

Of the same nature, in the next excerpt from her TAP, Trang intended to substitute “there is no denying the fact that” with its equivalent structure “It is undeniable that”. However, remembering that she already used it previously in the essay, she ended up keeping “there is no denying” to improve the syntactic variety of the essay.

Next, there is no denying the fact that. There’s no denying ….maybe it is undeniable, ah no. I used it once. There’s no denying the fact that anyone who earns a lot of money is successful.

The next example further illustrates Trang’s stylistics concern when she commented that the sentence may be lengthy:

Since they are educated people, it means that they are conceded to be successful people. And as a result. Oh I think here is so long sentence so I will delete the and and separate it into two sentences, two sentences. Since they are educated people, it means that they are conceded into successful people, dot. As a result, they will earn a lot of money which provides... which provides you an easier…

Linguistic revisions. In the stimulated recall, Trang attributed her routine of making self-initiated revisions to lack of satisfaction with the first draft, deeming revising and editing essential steps towards good writing. Therefore, she also conducted self-initiated revisions on lexico-grammatical features of the essay, especially on those aspects that were absent in
Criterion ACF, with the English verb tense and aspect system being among them. In the following except, verb tense became the major point for revision:

As you can see from Bill Gate, the founder of Microsoft Office, he has an incredible brain which helps him create this modern IT system. This modern IT system that we use every day. Extra that [add that after system] and that as a monopoly position, as a monopoly position in this field ... Uhm ... of course he became one. OK, because fully he is still, he still he is still well-known now, so I will change it from past simple into present simple. Present simple is maybe more appropriate. He becomes one of the richest man in the world.

Through the excerpt, it became clear that the learner was evaluating her own writing with reference to stored metalinguistic knowledge. Using her knowledge about the use of tenses, she adjusted “became” from the past simple to present simple to better express the idea which she believed to still hold true in the present, “he is still well known now”. Despite Trang’s falling short of taking a step further by choosing the more appropriate tense of present perfect in this particular case, her self-initiated evaluation of linguistic features in her own writing reflected a proactive approach to editing and revisions.

In the next example, as Trang was going through her writing again, she self-edited her lexical choices:

As a student who has not paid much attention. Umm...maybe I will change it into care ... care about care about rather than paid much attention to, cared much about. As a student who has not cared much about money pressure. Yet and depends on family I seem to remain my mind neutral. OK.

In the above example, Trang was trying to find a better word to express her idea, considering “care about” as a substitute for “pay attention to”. This consideration has gone beyond the question of being grammatically correct or incorrect. Instead, it focused on the more nuanced choices between alternative lexical expressions for the same meaning. A similar example to this is given below when Trang was considering changing “words” into “sayings” to refer to what Bill Gate said.
Beside that, anything belongs to him such as his words ... uhm ... his words ... I think his sayings will be clearer. His sayings, his way... his way of educating his children, etc. is thought to be successful too.

The next excerpt from Trang’s recorded TAP is probably the clearest exemplification of her self-regulation in conducting revisions as it included not only simple text operations of substitution or deletion. Instead, it was a major reformulation with some significant addition to the original draft through reworking of the idea for better general communication of the content,

Neverthele$$s$$, not anyone who earns much money is successful. Nowadays, people can get huge amount of money by tricking others with something wrong... in progressive technology. Umm... something wrong here. Nowadays people can get a huge amount of money by tricking others with... with... I think it has to be more detailed. By tricking others with... with hacking... with hacking... with hacking... OK. It’s an example, so I will add for example here. Nowadays people can get a huge amount of money, for example by tricking others with hacking with hacking, I mean I mean they will hack to... to steal the personal information. So I will add by tricking others with hacking personal information in progressive technology... banking... with hacking personal banking information. By tricking others with hacking personal banking information in progressive technology.

Overall, Trang’s self-initiated revisions revolved around the question of making ideas clear. Her explicit focus on the meaning of the essay has allowed her to make revisions at more global levels. This perhaps partially accounted for her improved accuracy over the studied period. In addition, the error tags she received from Criterion seemed quite straightforward and thus only need perfunctory noticing to help her revise successfully. She was also competent enough in evaluating the correctness of the error tags and responded appropriately. With a manageable number of error tags, she spent more time making self-initiated revisions besides responding to Criterion feedback, which gives the impression about Trang as a highly competent reviser who demonstrated limited engagement with Criterion ACF.
Case 2. Nhien, extensively engaged and self-motivated reviser

This section provides quantitative data summarizing Nhien’s error tags in three practice sessions, coded feedback processing episodes, her revisions initiated by Criterion ACF, as well as findings about her self-initiated revisions.

Profile. Nhien had learned English for eleven years and she had never travelled abroad. Her study of English was mainly in formal English classes at school as a compulsory subject. From classroom observation, Nhi demonstrated hard work and enthusiasm during the lessons, contributing a lot of ideas to class discussions and occasionally raising relevant questions. Similar to Trang, Nhien was new to learning to write using an automated writing evaluation tool like Criterion. Upon graduation, she wanted to become an English teacher to put what she learned at university to practice, hoping to follow a teaching career path.

Nhien obtained an overall accuracy rate of 88% (see Table 6.1). She had a very high score in the pre-test (at 95%), but it decreased to 91% and 84% in the post and delayed post-tests, respectively. According to Table 6.6, her pre, post, and delayed post-test essays were 279, 525, and 385 words long, making an average test essay length of 379. In terms of complexity measures, Table 6.3 shows that the syntactic complexity in Nhi’s essays increased slightly from 1.5 in the pre-test to 1.7 in the post-test, then decreased to the original level in the delayed post-test. Her lexical diversity, on the contrary, increased from 43% at pre-test to 51% at delayed post-test while lexical sophistication judged by AWL percentage was the same (at 5%) at these two time points (see Tables 6.4 and 6.5).

Turning to her practice essays, Nhien’s essay lengths and her normed error rates are provided in Table 6.12 below. As seen from this table, Nhien’s essay length increased incrementally across the three practice sessions, especially with the significant jump from 321 words in Week 5 to almost six hundred words in Week 8. Despite the increase in word counts, Nhien’s normed error rates based on Criterion’s error detection system decreased over time, with 5.3, 4.7, and 3.4 percent, respectively.

Nhien supplied all the data pieces required for the case study. Her data include pre, post, and delayed post-test essays; two drafts for each practice essay, a recorded TAP during the first practice session, and a stimulated recall interview one week after the TAP procedure.
Table 6.12
Word Counts and Normed Error Rates in Nhien’s Practice Essays

<table>
<thead>
<tr>
<th>Measure</th>
<th>Practice essay 1 (Week 5)</th>
<th>Practice essay 2 (Week 8)</th>
<th>Practice essay 3 (Week 11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word counts</td>
<td>321</td>
<td>592</td>
<td>613</td>
</tr>
<tr>
<td>Normed error rate</td>
<td>5.3</td>
<td>4.7</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 6.13 provides a break-down of error tags for different error types belonging to the three categories of grammar, usage, and mechanics in Nhien’s practice essays. As seen from this table, Nhien received most error tags on mechanics (29 errors). However, the vast majority of her mechanics problems (26) were about spelling. Nhien had a total of 21 error tags in usage across the three practice essays, with missing/extra article and confused words being the most frequently found. In terms of grammar, 12 of Nhien’s total 16 grammar errors were about subject-verb agreement. Twelve of the total 66 error tags were actually false positives spreading across seven error types. Missing or extra article and spelling had most false positives, with four and three counts, respectively.

Unlike other cases, Nhien was the only one who chose to use both Vietnamese and English when she was revising her essay and conducting the think-aloud protocol. Her chosen language for the retrospective interview was also Vietnamese. In this section, both her original Vietnamese verbalizations and the translations are presented in square brackets. Table 6.14 provides details about error tags Criterion generated in Nhien’s first practice session essay, followed by cognitive strategies she adopted to engage with the feedback, and revision outcomes.

Nhien spent a total of 23 minutes revising her essay written during the first practice session. Of these 23 minutes, she spent 17 minutes processing the automated corrective feedback generated by Criterion, and the remaining time was spent on self-initiated revisions related to content, idea organisation as well as lexical choices in her first draft. The following section is divided into two subparts, the first of which presents findings on Nhien’s Criterion-initiated revisions, and the second subpart about her self-initiated revisions.
Table 6.13
Error Tags in Nhien’s Three Practice Essays

<table>
<thead>
<tr>
<th>Category</th>
<th>Error Total: 66</th>
<th>Error Types</th>
<th>False Positives: 12*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>16 errors</td>
<td>12 Subject-Verb Agreement</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Ill-formed Verbs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Fragments</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Possessive Error</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>21 errors</td>
<td>10 Missing or Extra Article</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 Confused Words</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Determiner Noun Agreement</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Preposition Error</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Wrong Article</td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td>29 errors</td>
<td>26 Spelling</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Extra Comma</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Missing Comma</td>
<td></td>
</tr>
</tbody>
</table>

*All of the 12 false positives received ‘no change’ response.

**Criterion-initiated engagement episodes.** Table 6.14 provides details about Nhien’s first practice session essay. In this particular essay when Nhien’s think-aloud protocol was recorded, she received a total of 17 error tags and she spent 23 minutes processing all the revisions in her essay. Regarding Nhien’s response to the feedback, she had nine episodes coded as substantive noticing compared to eight perfunctory noticing cases. For the fragment error, Nhien processed the feedback three times, each time using a different engagement strategy. Overall, in different engagement episodes, Nhien employed different cognitive strategies, including drawing on stored metalinguistic knowledge, looking up online resources, translating into L1 (extensive engagement episodes), along with quick revision using ACF as a red flag, guessing the correct form, or simply skipping to the next error (limited engagement episodes). Turning to revision outcomes, Nhien made 13 correct revisions, one unjustified ‘no change’, one incorrect revision, and two appropriate ‘retention of the correct form’ in response to two false positives.
Table 6.14

Error Codes in Nhien’s First Practice Session Essay

<table>
<thead>
<tr>
<th>Error Types</th>
<th>Total: 17</th>
<th>Engagement Level</th>
<th>Coded Engagement Episodes</th>
<th>Revision Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar Errors (4)</td>
<td>1</td>
<td>Substantive</td>
<td>Skipping to the next error (first processing), Drawing on metalinguistic knowledge (second processing), and Looking up online resources + and Translating into L1 (third processing)</td>
<td>1 Retention of Correct Form</td>
</tr>
<tr>
<td>2 Subject-Verb Agreement</td>
<td>2</td>
<td>Perfunctory</td>
<td>2 Quick revisions using ACF as red flags</td>
<td>2 Correct Revisions</td>
</tr>
<tr>
<td>1 Possessive Error</td>
<td>1</td>
<td>Perfunctory</td>
<td>1 Quick revision using ACF as a red flag</td>
<td>1 Incorrect Revision</td>
</tr>
<tr>
<td>Usage Errors (7)</td>
<td>1</td>
<td>Substantive</td>
<td>1 Looking up online resources</td>
<td>1 Correct Revision</td>
</tr>
<tr>
<td>3 Missing or Extra Article</td>
<td>2</td>
<td>Perfunctory</td>
<td>1 Guessing the correct form 2 Quick revision using ACF as a red flag</td>
<td>3 Correct Revisions</td>
</tr>
<tr>
<td>2 Confused Words</td>
<td></td>
<td>Perfunctory</td>
<td>1 Quick revision using ACF as a red flag</td>
<td>1 Correct Revision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Substantive</td>
<td>1 Looking up online resources</td>
<td>1 No Change (Unjustified)</td>
</tr>
<tr>
<td>1 Preposition Error</td>
<td></td>
<td>Substantive</td>
<td>1 Looking up online resources</td>
<td>1 Correct Revision</td>
</tr>
<tr>
<td>Mechanics Errors (6)</td>
<td>6</td>
<td>Perfunctory</td>
<td>1 Quick revision using ACF as a red flag</td>
<td>1 Retention of Correct Form</td>
</tr>
<tr>
<td></td>
<td>(1 False positive)</td>
<td>Substantive</td>
<td>5 Looking up online resources</td>
<td>5 Correct Revisions</td>
</tr>
</tbody>
</table>
**Effective utilization of resources for correct revisions.** Nhien was successful at most revision attempts after using online references and being extensively engaged with the given feedback. Her main resources were online dictionaries and online forums of grammatical topics. For instance, in the following episode, she expressed uncertainty about whether ‘law’ was a countable or uncountable noun, which explains why she used the incorrect structure “many strict law” in the first draft. The error tag from Criterion gave her the chance to double check this using Oxford online dictionary to confirm “law” as a countable noun, contrary to her previous belief. More specifically below:

**Student text:** Second, having many strict law is another effective way to reduce the shortage of water.

**Error message:** You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.

**TAP excerpt:** Determiner noun agreement. Agreement of nouns. Many, have many strict laws, have many strict law. Many strict law. Yeah. Law is a an uncountable? I will check in the Oxford dictionary… Ah OK. I will check the law to see ah it’s a countable or uncountable noun. OK. It is it is a count noun. So so I use the many here. Second, having many strict law. OK, I will change this into plural noun. Hope it’s correct.

**Revised text:** Second, having many strict laws is another effective way to reduce the shortage of water.

Similar to the example above, Nhien’s processing of the next preposition error is another illustration of her extensive engagement with the feedback when she again referred to the Oxford online dictionary, yet this time taking a step further by utilizing example sentences as a way to confirm her doubts:

**Student text:** The amount of fresh water is decreasing rapidly while the demand on it is increasing day by day.

**Error message:** You may be using the wrong preposition.

**TAP excerpt:** Sai lỗi về giới từ. Uhm…the demand on ahh the amount of fresh water is decreasing rapidly while the demand on… the demand. Minh sẽ check xem demand xem thử có cái gì? Demand. Ahh. Sao đây là ghi là demand for nhi? A series of demands for for…Ahh [reading the example provided

[Translation: Errors on prepositions. Uhm… the demand on ahh the amount of fresh water is decreasing rapidly while the demand on… the demand. I will check demand to see if there is some information. Ahh. Why does it say demand for here? A series of demands for for…Ahh [reading the example provided by Oxford Dictionary]. It expressed a necessity, a request. He’s got enough demands on his time already. A demand for specialists. Uhm, can I try using it? Demand for. Demand for. OK.]

Revised text: The amount of fresh water is decreasing rapidly while the demand for it is increasing day by day.

**Taking the feedback seriously.** On one particular occasion, Nhien experienced difficulty understanding the error message and made no change in response to a false positive on a fragment error. She kept coming back to the error three times (see Table 6.14 for more details about how Nhien processed this fragment error), trying to figure out what the meaning of the error explanation was by googling the meaning of the word “fragment”. After three attempts, she decided that she did not know what was wrong with her sentence:

**Student text:** Indeed, water accounts for 75 percent of humans’ bodies.

**Error message:** This sentence may be a fragment. Proofread the sentence to be sure that it has at least one independent clause with a complete subject and predicate.

**TAP excerpt 1:** Fragments. I don’t know what does fragments mean? This sentence may be a fragment, proof…proofread this sentence to be sure that it has at least one independent clause with a complete subject and predicate. Oh my god. Indeed, water accounts for seven…seventy five percent of humans’ bodies. I don’t know how to correct it. Ahh indeed, water accounts…OK. I don’t know how to correct this sentence. Indeed, water accounts for seventy five percent of humans’ body.
Nhien went back to this false positive fragment error code two more times after having corrected the spelling error from acounts to accounts, yet Criterion still tagged the sentence as a fragment.

TAP excerpt 2: Ahh one fragment here… Indeed, water accounts for seventy percent of the human body. I really don’t know what the error is. Here. Indeed water water, water is not a subject?

TAP excerpt 3: Go back to the first one. Indeed, accounts for…water accounts for…They say that it is a fragment. … Run-on sentences or Garbled Sentences. Garbled sentences. What does garble mean? [Google translate] Câu bị cắt bớt à? Câu bị cắt bớt nghĩa là rằng? OK. Thực sự thì nước chiếm 75% cơ thể người. Sai rồi? Minh không biết mình sai chỗ nào nên mình vẫn để như vậy. [Translation: A truncated sentence? What does it mean by a “truncated sentence”? OK. Actually, water accounts for 75% of humans’ bodies. Is this wrong? I don’t know what is wrong with this sentence, so I will just leave it unchanged.]

With another false positive from Criterion, an error tag on the word “pollutive” as a spelling error, she was reluctant to revise, believing that it was a correct word to use. She double checked by googling the word ‘pollutive’ to confirm her no-change decision.

Student text: The main source of these pollutive things is daily activities of residents living near by water source...

Error message: The word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.

TAP excerpt: Ahh this main source of these pollutive things. Pollutive [Check Oxford dictionary] Pollutive, pollutive. I think it’s correct.

As seen from the above examples, when faced with false positives, Nhien often came back to the error codes more than once to make sure that her decision to resist making changes was well thought out. She tried to evaluate her writing and read the feedback to see if she really needed to make a change. After careful examination of the feedback and her own writing, Nhien was able to respond well to all of the twelve false positives across three practice sessions by making no change. Her dual approach of examining her own writing as well as evaluating the correctness of the feedback helped Nhien make well-informed decisions on her revisions.
Failed attempts to revise. Nhien’s first incorrect revision was about a complex issue at the clausal level. The error being tagged was a “possessive error”, but this was an incorrect error code. Nhien took the error feedback seriously and revised the text in the direction given by Criterion. However, she was unable to get useful information from the Criterion error message. This, in combination with her insufficient linguistic resources to deal with this particular error, resulted in an incorrect revision.

First draft:  
First of all, saving the water by raising people’s aware seems to be the most actical solution.

Error message: You may need to take out the apostrophe to make this word a plural noun.

TAP excerpt: About possessive. People. This is the possessive error. People, people should be… it should not have this form. By raising, raising….aware of people aware…. Of everyone.

Revised draft:  
First of all, saving the water by raising aware of everyone seems to be the most reality solution.

Nhien’s serious effort to revise her errors was also manifest in the two attempts at an error tag of confused words. Similar to the previous episode, Nhien’s failure was again due to the gap in her knowledge. In the first attempt to revise, Nhien cursorily examined the error, which was followed by a quick decision to make no change. This decision was unjustified.

Student text:  
Making all the effort to reduce the affect of this problem is the responsibility of us.

Error message: You have used affect in this sentence. You may need to use effect instead.

TAP excerpt 1: The affect of this problem, OK. To reduce the affect of this problem, OK. Don’t know how to check, so I will make it… will keep it.

However, Nhi came back to the same error with substantive noticing and made a more serious effort to find out the issue being tagged,

TAP excerpt 2: À chữ affect. Affect here. In conclusion, water pollution….to reduce the affect. This is the mistake of confused forms. What’s this problem? Sai cái
Yet, even after the second attempt, Nhien decided to give up by making no revision. She seemed to have missed the correct suggested substitute from Criterion after all. Her insufficient knowledge about the difference between effect and affect was confirmed during the stimulated recall, when she was asked if she understood the metalinguistic message provided on this error.

Researcher: So you were repeating a lot “what’s wrong, what’s wrong” with this one. But you ended up making no change. What’s the reason for that?

Nhien: Yeah, with this error, they say that it’s a confusing word, I don’t know, I really don’t know what it means by confusing word.

Researcher: So you don’t understand the message of the error here?

Nhien: Yeah, I don’t understand so, and I read it again and find it quite quite OK, so I don’t make any change.

In addition to the gap in her knowledge about the two confused words affect and effect, she also had difficulty understanding Criterion’s metalinguistic explanation for this particular error. Therefore, despite her attempt to substantively engage with the feedback, as is the way she dealt with most other error tags from Criterion, she was unable to revise successfully.

In another engagement episode when Nhien failed to arrive at a correct revision, she was struggling with an alternative word for “actical” (tagged as a spelling error). Despite two attempts at both google-translating the word she wanted to use and looking up the thesaurus for a synonym, Nhien came up with “reality” as the alternative for “actical”, although she was still uncertain.
**Student text:**  
Saving the water by raising people’s aware seems to be the most **actical** solution.

**Error message:**  
This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.

**TAP excerpt :**  
Minh muốn viết một chữ gọi là thực tế. Không rõ nó có đúng hay không?  
[Search on Google Translate] …**Reality**. I want to have another word rather than **reality**. Where is it? Where is it? Umm…I will use the Thesaurus [a function available on Criterion] [change from “actical” to “reality”]. **Reality**. …Thesaurus. Umm, it does not have any word I want.  
 **[Translation]:** I want to use a word which means ‘practical’. I am not sure if it is correct. [Search on Google Translate] …**Reality**. I want to have another word rather than **reality**. Where is it? Where is it? Umm…I will use the Thesaurus [a function available on Criterion] [change from “actical” to “reality”]. **Reality**. …Thesaurus. Umm, it does not have any word I want.

In this failed attempt, Nhien made an incorrect revision despite the correct error tag from Criterion which pointed out to her that “actical” was incorrectly spelled. During her stimulated recall interview, Nhien discussed this error extensively.

**Researcher:**  
After you made the change, you used the Thesaurus, Google translate, and finally you have “reality solution”, did you feel satisfied with that?

**Nhien:**  
Yeah, at that time, I felt satisfied with that… I wanted to use an adjective but at that time I think of “practical”, but I don’t know why I can’t remember that word.

**Researcher:**  
At that time?

**Nhien:**  
Yeah. At that time I just remembered “actical” but it’s a wrong word, so I have checked it on Google, but it’s not the “practical” here, it just showed me the “reality” so…

**Researcher:**  
Now if you have the choice, which word would you go for?

**Nhien:**  
Yeah, I want to choose the “practical”.

**Researcher:**  
**OK.** Were you aware that “reality” is a noun then? At that time?

**Nhien:**  
Yeah, no, maybe.
Researcher:  *Now you know. On looking back, what would be your explanation? It was because of the online resources or because of something else?*

Nhien:  *Uhm, I was quite stressed, so I can’t think very clearly at that time.*

Contrary to the two cited examples above, it turned out from Nhien’s explanation during the interview that the word “practical” was in her lexical repertoire. However, being in the “stressful” situation of composing and revising, she was not able to immediately retrieve the word from her long-term memory in response to *Criterion*’s error tag. Nhien attempted to refer to online resources but Google did not provide the correct alternative for her error. Therefore, her revision was incorrect.

Overall, the source of Nhien’s incorrect revisions was different in the different examples above. During the failed attempts, she encountered occasional problems of understanding *Criterion* error message, dealing with incorrect error codes, lacking sufficient stored knowledge to revise correctly identified errors, and her inability to retrieve learned knowledge under the pressure of writing and revising.

**Self-initiated revisions.** Besides the revisions made in response to *Criterion* ACF, after attending to all the errors tagged by the system, Nhien spent six out of the total 23 minutes revision time making changes related to content and the overall organization of ideas in her essay. She re-read her writing from beginning to end, stopping occasionally to see if the ideas were well organised. The main approach to proofreading her essay was translation of her writing into L1, Vietnamese. She especially paid attention to the last paragraph, which she considered not clear enough for readers:

First draft:  *Finally, according to some volunteer organizations, encourage movements could make a big change in the community.*

TAP excerpt: Nhưng cái cuối cùng này khá là mờ hồ. Minh sửa lại According to some volunteer organizations, encourage movement could make a big change in the community. Encouragement. Encourage movement là…[check on google search] encouragement movement encouragement movement...là…Another big change in the community. Another big change in the community. Cái ý này nó không có rõ ràng. Từ change này phải là một sự thay đổi trong cách mọi người sử dụng nước …in the way… people… treat…. This cái nguồn cái nguồn quý giá này. Quý giá là
Finally, according to some volunteer organizations, encourage movements could make a big change in the way people treat this worthless natural resource.

[Translation: These last bits are quite vague. I will revise them.

According to some volunteer organizations, encourage movement could make a big change in the community. Encouragement. Encourage movement...this means... Another big change in the community. Another big change in the community. This idea is not clear. The word change has to refer to a change that takes place in the way people use water...in the way...people...treat....this this valuable resource. What is valuable in English? Valuable is worth worth priceless...is worthless resource, natural resource, natural resource....OK. Finally, according to some volunteer organizations, encourage movements could make a big change in the way people treat this worthless natural resource.]

Revised draft: Finally, according to some volunteer organizations, encourage movements could make a big change in the way people treat this worthless natural resource.

Although her word choice of ‘worthless’ in the revised version was an incorrect one, it shows Nhien’s self-initiatives in trying to improve the clarity of her ideas and overall structure of the essay. She also demonstrated experimentation with new expressions to present her ideas more effectively. This was further explained in her stimulated recall:

Yes, after writing, I always check it again....and sometimes I find it very and some sentences when I writing I write as my thought but when I writing it maybe quite hard. I think it’s quite hard for the reader can understand my sentences, so I change it.

Her main focus during the revising stage, however, was still on linguistic features than the ideas due to lack of confidence in the grammatical accuracy of the essay, which was substantiated by the much longer period of time Nhien spent working on form-focused errors using Criterion feedback (17 of the total 23 minutes). Nhien believed that content should be the first priority during the pre-writing stage when she was making an outline for the essay:
I will focus on the grammar more than the idea.... Because when I’m writing maybe it can be affected much by speaking grammar, so I want to revise it to writing grammar.

To sum up, Nhien went from local to global revisions in her editing approach. She seemed to be a strategic reviser through quite effective utilization of the automated feedback, careful evaluation of the correctness of the error tags before resorting to either online resources or stored metalinguistic knowledge to inform her final revising decision. Her dual focus on local (Criterion-initiated) and global (self-initiated) aspects of her essay with her own experiment with new phrases in revisions of the content demonstrates her habitual systematic attention to revising as a critical stage in the writing process.

Case 3. An, confident and superficially engaged reviser

Profile. An started learning English at a much younger age than Trang and Nhien, with 17 years of language learning experience. She had been to Singapore on a short vacation and that was her only experience being in an English-speaking country. Her plan was to further her education in some English-speaking country upon graduation. Like Nhien, An seemed hard-working and was willing to contribute ideas during classes. During an informal interview before her TAP recording, she also mentioned that she had some experience using Grammarly, a free web-based automated essay evaluation tool, to check her own grammar in English writing courses.

An attained an overall accuracy rates of 92%, 89%, and 83% for the pre, post, and delayed post-tests, respectively (see Table 6.1). Her essay length was 387 on average. She wrote 328 words in her pre-test, which increased to 392 and 438 words in the two subsequent tests, as seen in Table 6.6. Regarding complexity, Tables 6.3, 6.4, and 6.5 indicate that her syntactic use was most complex at the post-test (2.2) but equally low at pre-test (1.4) and delayed post-test (1.5). An’s lexical diversity based on type-token ratio increased over time, and so did her lexical sophistication.

Turning to her practice essays, An’s essay lengths and normed error rates are provided in Table 6.15 below. Similar to the other two cases, An’s essay length increased incrementally from the first practice session in Week 5 to the third practice session in Week 11. An’s normed error rates based on Criterion’s error detection system in the first practice session was 3.9%, followed by an increase to 4.4% in the second practice session. However, her draft in Week 11 (Practice session 3) had a lower error rate of 2.9%.
Table 6.15

*Word Counts and Normed Error Rates in An’s Practice Essays*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Practice essay 1 (Week 5)</th>
<th>Practice essay 2 (Week 8)</th>
<th>Practice essay 3 (Week 11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word counts</td>
<td>389</td>
<td>455</td>
<td>622</td>
</tr>
<tr>
<td>Normed error rate</td>
<td>3.9</td>
<td>4.4</td>
<td>2.9</td>
</tr>
</tbody>
</table>

An completed all the data required for the case study, including pre, post, and delayed post-test essays, three drafts for each practice essay, two recorded think-aloud protocols during the second and third practice sessions, and a stimulated recall interview one week after the third TAP procedure.

**Engagement episodes and revision patterns.** Table 6.16 provides details about the numbers of errors in three categories, specific error types, and false positives in all the first drafts of An’s practice essays.

Table 6.16

*Error Tags in An’s Three Practice Essays*

<table>
<thead>
<tr>
<th>Category</th>
<th>Error Total: 53</th>
<th>Error Types</th>
<th>False Positives: 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>4 errors</td>
<td>2 Fragments</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Ill-formed Verbs</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>24 errors</td>
<td>16 Missing or Extra Article</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Preposition Error</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Determiner-Noun Agreement</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Confused Words</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Wrong Form of Word</td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td>25 errors</td>
<td>20 Spelling</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Missing Initial Capital Letter in a Sentence</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Missing Comma</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Fused Words</td>
<td></td>
</tr>
</tbody>
</table>
As seen from Table 6.16, An received a total of 53 error tags, most of which were of mechanics and usage types (25 and 24 errors, respectively). Only four errors were about grammar, equally divided between fragments and ill-formed verbs.

Across the two essays from Practice sessions 2 and 3, An spent a total of 28 minutes and 40 seconds revising. Unlike Trang and Nhien who had only one recorded think-aloud protocol each, An had two TAP recordings for two different practice sessions. Therefore, for clearer information regarding each think-aloud protocol and their relevant error tags as well as An’s response to them, there will be three tables below. Table 6.17 presents the total numbers of errors by error category in An’s second and third practice sessions when her think-aloud protocols were recorded. This is supplemented by two more tables, 6.18 and 6.19, which give further details about An’s cognitive engagement with the error tags she received and her revision outcomes for each of these two essays.

Table 6.17

<table>
<thead>
<tr>
<th></th>
<th>Grammar</th>
<th>Usage</th>
<th>Mechanics</th>
<th>Total</th>
<th>Total revision time</th>
<th>Number of false positives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second practice</strong></td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>20</td>
<td>16mins 30secs</td>
<td>3</td>
</tr>
<tr>
<td><strong>Third practice</strong></td>
<td>0</td>
<td>13</td>
<td>5</td>
<td>18</td>
<td>12mins 10secs</td>
<td>2</td>
</tr>
</tbody>
</table>

According to Table 6.17, An received quite similar numbers of error tags in these two essays. However, she spent more time processing the error codes in the second practice than those in the third practice essay. In total, she received five false positives (three in the second and two in the third practice session). Further details are discussed in subsequent sections. Tables 6.18 and 6.19 provide more detailed information regarding the specific numbers in each error type, false positives, her engagement levels with the feedback for each error through specific cognitive strategy, and the subsequent revision outcomes for the errors.
Table 6.18

An’s Error Tags on Practice Essay 2

<table>
<thead>
<tr>
<th>Error Types</th>
<th>False Positives</th>
<th>Coded Engagement Episodes</th>
<th>Engagement Level</th>
<th>Revision Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Fragments</td>
<td></td>
<td>2 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td>2 Correct revision</td>
</tr>
<tr>
<td>2 Ill-formed Verbs</td>
<td></td>
<td>2 Refer to stored metalinguistic knowledge</td>
<td>Substantive</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td>1 Incorrect revision</td>
</tr>
<tr>
<td>2 Missing or Extra</td>
<td>1</td>
<td>1 Quickly adopting <em>Criterion</em> ACF</td>
<td>Perfunctory</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td>Article</td>
<td></td>
<td>1 Skipping to the next error (1st processing), and drawing on stored metalinguistic knowledge (2nd processing)</td>
<td>Substantive</td>
<td>1 No change (in response to False Positive)</td>
</tr>
<tr>
<td>2 Preposition Error</td>
<td></td>
<td>1 Guessing the correct form</td>
<td>Perfunctory</td>
<td>2 Correct revisions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Quick revision using ACF as a red flag</td>
<td>Substantive</td>
<td></td>
</tr>
<tr>
<td>8 Spelling</td>
<td>2</td>
<td>2 Drawing on stored metalinguistic knowledge</td>
<td>Substantive</td>
<td>2 Retention of correct form</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Looking up online resources</td>
<td>Substantive</td>
<td>6 Correct Revisions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td></td>
</tr>
<tr>
<td>2 Missing Initial</td>
<td></td>
<td>1 Getting cues from <em>Criterion</em> error message to revise</td>
<td>Substantive</td>
<td>2 No change (Correct structures due to revisions to other error tags)</td>
</tr>
<tr>
<td>Capital Letter in a</td>
<td></td>
<td>1 Skipping to the next error</td>
<td>Perfunctory</td>
<td>1 Incorrect revision (to incorrect error tag)</td>
</tr>
<tr>
<td>Sentence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Missing Comma</td>
<td></td>
<td>1 Getting cues from <em>Criterion</em> error message to revise +</td>
<td>Substantive</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Looking up online resources</td>
<td>Substantive</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Skipping to the next error</td>
<td>Perfunctory</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.19

*An’s Error Tags on Practice Essay 3*

<table>
<thead>
<tr>
<th>Error Types</th>
<th>False Positives</th>
<th>Coded Engagement Episodes</th>
<th>Engagement Level</th>
<th>Revision Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Determiner-Noun</td>
<td>1</td>
<td>1 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td>1 Incorrect revision</td>
</tr>
<tr>
<td>Agreement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Missing or Extra</td>
<td>1</td>
<td>1 Guessing the correct form</td>
<td>Perfunctory</td>
<td>6 Correct revisions</td>
</tr>
<tr>
<td>Article</td>
<td>2</td>
<td>2 Quickly adopting ACF</td>
<td>Perfunctory</td>
<td>1 Incorrect revision</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2 Drawing on stored knowledge</td>
<td>Substantive</td>
<td>1 No change (unjustified)</td>
</tr>
<tr>
<td>3 Confused Words</td>
<td>2</td>
<td>2 Drawing on stored knowledge</td>
<td>Substantive</td>
<td>1 Retention of correct form</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 Quickly adopting ACF</td>
<td>Perfunctory</td>
<td>2 Correct revision (one in response to a false positive)</td>
</tr>
<tr>
<td>1 Wrong Form of Word</td>
<td>1</td>
<td>1 Looking up online resources</td>
<td>Substantive</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td>4 Spelling</td>
<td>1</td>
<td>1 Quick revision using ACF as a red flag</td>
<td>Perfunctory</td>
<td>4 Correct revisions</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2 Looking up online resources</td>
<td>Substantive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1 Guessing the correct form</td>
<td>Perfunctory</td>
<td></td>
</tr>
<tr>
<td>1 Fused Words</td>
<td>1</td>
<td>1 Getting cues from <em>Criterion</em> error message</td>
<td>Substantive</td>
<td>1 Correct revision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ Looking up online resources</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As seen from Tables 6.18 and 6.19, An had a total of 43 engagement episodes with *Criterion* ACF, including instances where she was engaged with some of the feedback points more than once. During these 43 episodes, she demonstrated both extensive and perfunctory engagement, with 21 substantive and 22 perfunctory noticing episodes. Regarding revision outcomes, An made 26 correct revisions and only four incorrect ones. There were four cases of justified ‘no change’ response to false positive error tags, and one correct revision to another false positive due to the acceptable alternative forms. There was only one unjustified ‘no change’ response where An received a correct error code from *Criterion*.

In addition, two errors corrected themselves after An responded to other error tags as *Criterion* double-coded these errors. The following is one of these two cases,

**First draft:**  
*However*\(^1\) \(\; ^2\) *many people have many different satisfaction*.\(^3\)

In this sentence, *Criterion* flagged three places to indicate three errors, as below:

**Criterion error tags:**

\(^1\) *You may need to use a comma after this word.*  
\(^2\) *This sentence may be a fragment. Proofread the sentence to be sure that it has at least one independent clause with a complete subject and predicate.*  
\(^3\) *Remember to capitalize the first word of each sentence.*

After attending to *Criterion*’s error tag on missing comma by replacing the full stop with a comma, the other two error tags were no longer valid. Therefore, she did not respond further to *Criterion*’s messages about fragments and missing initial capital letter in a sentence.

**Revised draft:**  
*However, many people have different satisfaction.*

**Focus on grammatical accuracy.** During her stimulated recall, An explained that she focused on addressing every single error being pointed out by *Criterion*. She said, “*I try to correct all the mistakes when, until it has no [more] mistakes*”. This may provide a partial explanation for her submission of three drafts during each practice session within the time limits, compared to only two submissions from Nhien, or the less consistent numbers of submissions across the three practice sessions for Trang. An’s ultimate goal, as she further commented during the interview, was to “find
out which mistake I have and the final goal I want to know which mistake I’m getting and it helps me enlarge my knowledge about writing”.

An’s main resources for error correction as generally seen from Tables 6.18 and 6.19 were her stored metalinguistic knowledge (nine episodes), followed by reference to online resources (eight episodes). Notably, An used Internet resources mostly for issues in the error types related to vocabulary, the great majority of which were about spelling. She also made use of Criterion’s error tags for revisions such as simply considering it as a reminder for revision (seven episodes). In addition, she adopted Criterion’s suggested changes in four cases, and read Criterion’s error explanation to get clues for revised forms in three cases. In-depth analyses of how An used these engagement strategies for specific errors are presented in the qualitative findings below.

Quick and successful revisions based on stored linguistic resources. Unlike the other two cases, all of An’s revisions were initiated by Criterion error tags in her essays. She did not conduct self-revisions and her total revision time was generally shorter than that of Trang and Nhien, at an average of 14 minutes per session. Overall, An had almost an equal number of substantive and perfunctory noticing episodes coded in her think-aloud protocols. Across both practice sessions when her TAPs were recorded, her revision success rate was 63%, with more correct revisions deriving from perfunctory noticing episodes, at 73%. Incorrect revisions were at a low 15% and 9% for substantive and perfunctory noticing, respectively. With closer investigation of her revised essays, the high revision success rate could be explained by the fact that most of the errors she got feedback on were surface level issues such as spelling. This also explained why An spent relatively brief time periods processing each error. For example,

First draft: The next reason is that how we spend our salaries decide how successful we are.
Error message: This word is not spelled correctly. Use a dictionary or spellchecker when you proofread your work.
TAP excerpt: And the next is successful. OK. Delete one L. Successful.

In the stimulated recall, An reflected on why she very quickly revised the spelling mistake of “successfull” into “successful”. She commented, “Successful is very easy to see, easily to see, always ful is one L, when I see it two Ls, I easily to realize that.”
For other error types such as the one below about *missing or extra article*, An also processed the feedback quickly and effectively, again thanks to her previously acquired linguistic knowledge.

First draft:  
*I prefer to face challenge to renew myself.*

Error message:  
*You may need to use an article before this word.*

On seeing *Criterion’s* highlighting of the word “challenge”, An realized that her use of “challenge” was grammatically incorrect. However, instead of proceeding with reading *Criterion’s* pop-up screen for its metalinguistic explanation which suggested adding an article before the word, An depended on her judgement of the error and edited according to her belief that the plural form of this noun should be used to talk about more than one challenge in general, as she verbalized,

TAP excerpt:  
*Face challenge. OK, there are many challenges. Add s, OK challenges.*

Revised draft:  
*I prefer to face challenges to renew myself.*

Similar to Trang, by perfunctorily noticing the highlighted error, An was successful at revising her text without the need to refer to *Criterion’s* error message. She seemed to be confident in her ability to self-correct the errors.

An’s dependence on her stored metalinguistic knowledge also helped her make informed decision in response to incorrect error codes. In the following example when An processed an error tag on missing or extra article, she engaged with it twice. Her first engagement with this error tag was a perfunctory noticing episode which resulted in her rejection of the feedback.

First draft:  
*Some people donate their money to charity, others use for travelling and hanging out with friends, family for memorable moments.*

Error message:  
*1You may need to use an article before this word.*

TAP excerpt 1:  
*And family? Use an article before this word. Use an article before this word? I don’t think I will need to use an article before this word family.*

After submitting the revised draft and finding the same feedback on the word “family”, she extensively engaged with it and supported her resistance with reference to her stored metalinguistic knowledge.
TAP excerpt 2: *Family*. You may need to use an article before this word. Use an article before this word? No, *friends* or *family* is, is general. I don’t think use an article before this word.

**Retrieval of learned knowledge.** It was not always the case that An’s stored metalinguistic knowledge was immediately retrieved to facilitate correct revisions right at the moment she came into contact with a highlighted error. Instead, on some occasions, it was *Criterion*’s error tag and/or its metalinguistic explanation that triggered the retrieval of previously learned knowledge. The cited example below illustrates this point.

First draft:  

As a result, this method of making [living]¹ can not bring us success and happiness but receiving our society’s disrespect.

TAP excerpt:  

¹And missing or extra article. *Making living?...*

It can be seen from An’s hesitation that she was not sure why the phrase ‘making living’ was flagged as an error. That was why she turned to the error message to read *Criterion*’s further explanation,

Error message:  

*You may need to use an article before this word. Consider using the article a.*

On finishing reading the popped-up error message, An continued with the think-aloud, stating that she would do what was suggested, “A living…. OK I’ll add a after *making* and before *living*. *Making a living. OK. Making a living.*” This resulted in her revised draft below,

Revised draft:  

As a result, this method of making a living can not bring us success and happiness but receiving our society’s disrespect.

Despite the impression that An was simply adopting *Criterion*’s suggested change through her think-aloud protocol excerpt, during the follow-up interview, it became clear that *Criterion*’s error message had actually helped her confirm some knowledge she had gained previously but was not totally sure about,
An: I was reminded, I think “making living” it means “kiểm sống” in Vietnamese and I don’t think … and I use “make living”, it’s simply, and when I use Criterion mistake to check they give me a…

Researcher: A message that it’s wrong?
An: message that it’s wrong and when I click to the “living”, they, I have to use article, so I think I should, I add article “a” before “living”.

The next example further exemplifies her successful revision based on reference to stored metalinguistic knowledge, yet with the help of Criterion’s error identification. This error tag on an ill-formed verb was ignored during An’s first processing episode,

First draft: On the other hand, several earners try to pull down a large income because of fame or social position, which makes them to live meaninglessly.

Error message: You may have used the wrong form of this verb. Match the subject to the verb to decide whether you have used the verb correctly.

TAP excerpt 1: Now the next part about grammar. Ill-formed verbs. The first, make them to….Uhm….what’s wrong? Make them…You may have used the wrong form of this verb. No, I don’t think it is wrong.

In the stimulated recall, An reflected on her decision to disregard Criterion’s error tag at the first time of seeing it by explaining that “I think maybe, at that time, I may be afraid many things so I don’t I don’t have to focus on this.” However, to this same error above about ill-formed verb, An came back to process it for a second time. Seeing that the error was still there with the same Criterion tag after she submitted the revised draft, this time she spent more time reading Criterion’s metalinguistic explanation and processing it through reference to stored metalinguistic knowledge:

TAP Excerpt 2: Ill-formed verbs. Let me see. Which makes them to…used wrong form of this verb. Ah OK. Make somebody. OK. Make somebody plus verb infinitive, not to. OK. I’ll delete to.

Revised form: On the other hand, several earners try to pull down a large income because of fame or social position, which makes them live meaninglessly.

In the stimulated recall interview, she confirmed her previously learned knowledge regarding this structure,
An: I remembered I studied after “make”, “make” plus object and infinitive.

Researcher: Use a bare infinitive. And why doesn’t this come at the beginning? Here, during the second minute of your recording you were looking at this [showing An the part of her processing making living for the first time] and you didn’t make any change.

An: I think maybe, at that time, I may be afraid many things so I don’t I don’t have to focus on this.

Researcher: So all these things are in your grammatical knowledge already?

An: Yes. Everything I have, everything when I was writing I remember “make a living” and “make them live”.

To further explain why she ended up making a revision to this error which she resisted during the first attempt, she said,

After after revising two times and several mistakes, the Criterion shows me that I still have mistakes in this, in “them to” or “making living”, it reminds me, it reminds me to remember what I have studied, I try to correct it.

Overall, had it not been for Criterion’s error flags in the above cases, An would have overlooked the mistakes she made despite the fact that she had learned the structures previously. The automated corrective feedback was able to both initiate An’s revisions and help her resolve uncertainties in her learned knowledge.

Handling unhelpful feedback: Incorrect error codes and false positives. An received a total of five incorrect error codes and five false positives. In the case of incorrect error codes, An’s taking the error message at face value followed by quick adoption of suggested changes was problematic. The following example illustrates an error tag from Criterion which failed to pinpoint the real problem of the error:

First draft: Recently, the youth have said that a lot of money can help them make their big dreams come true because they might not be do success without big researchs which need a huge amount of money.

Error message: You may have used the wrong form of this verb. Match the subject to the verb to decide whether you have used the verb correctly.
Instead of directing An to the correct structure of “copula be + adjective” (i.e., might not be successful) or other alternative and grammatically correct structures like ‘might not succeed’, Criterion focuses locally on the combination of the two verbs ‘be’ and ‘do’ without considering the meaning of the whole clause. In other words, the system seems to fail to capture the connection between these verbs and the noun success that follows. Being provided with insufficient guide to correct this error, An verbalized in her TAP:

Ahh, might not be do success. Might not be do. OK. It’s wrong. Might not do success, not be. So I will delete be. Where where where. OK, here it is, I will delete be, might not do success, because after an auxiliary verb, use verb infinitive.

An’s metalinguistic reflection was correct as she verbalized, “I will delete ‘be’, might not do success, because after an auxiliary verb, use verb infinitive”. However, this metalinguistic knowledge is not sufficient to warrant her change from “might not be do success” to “might not do success” as the revised phrase was neither grammatically correct nor comprehensible.

Revised draft: Recently, the youth have said that a lot of money can help them make their big dreams come true because they might not do success without big researchs which need a huge amount of money.

In the next example, Criterion’s error tag was about missing comma. An’s correction of this error provides further evidence about her failure to benefit from adopting Criterion ACF due to the unhelpful feedback she received,

First draft: As a result, this method of making living can not bring us success and happiness but receiving our society’s disrepect.

Error message: You may need to use a comma after this word.

Although An was extensively engaged with the feedback, she was unsuccessful in revising the error. The first reason that may account for this failed attempt is that Criterion’s error tag was at the very superficial level of suggesting the need for a comma preceding the coordinating conjunction “but”. However, the real problem lies beyond such local boundaries and fixing it requires global revisions across two clauses in this sentence. What An did was to simply make a local revision to address the issue as suggested by Criterion. That is, after extensive engagement with the error message, she decided to adopt Criterion’s unhelpful recommended change:

Revised draft: As a result, this method of making a living can not bring us success and happiness, but receiving our society’s disrespect.

A very similar engagement episode was seen with the determiner-noun agreement error tag below. In this particular case, after engaging with the feedback, she decided to make a quick change.

First draft: Furthermore, staying a jobs1 for longer time makes you become as a part of an organization.

Error message: You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies.

TAP excerpt: Determiner noun agreement. Staying a jobs. OK. Staying a jobs. Yes, let’s begin. About noun agreement. Let me see. Staying a jobs. OK staying a job or staying jobs. I will delete letter a. OK. That’s all.

An’s unsuccessful attempt at revising these structures was not due to a gap in her linguistic knowledge. Instead, probably her inability to evaluate the insufficiency of the error tag she received explained why the revision did not solve the real problem. Added to this, the lack of thorough treatment of the feedback in the case of the determiner-noun agreement error could explain why she was not successful at making a more substantive revision to address the problem.

Turning to false positives, An responded as expected to four out of the total five false positives she received by keeping her text unchanged. Interestingly, An had extensive engagement with these points of false positive feedback from Criterion and she also processed each falsely flagged error more than once. For example, both of the false positives on spelling (Keangnam and meaninglessly) received substantive noticing when An reflected on her previously acquired knowledge to reject Criterion’s feedback. Similarly, in the third false positive about confused
words, An was deeply engaged with it through reference to her stored metalinguistic knowledge, as below:

**First draft:** In addition, you remarkably increase your earning power in another company which appreciates your ability and strengths.

**Error message:** You have used your in this sentence. You may need to use you’re instead.

**TAP excerpt:** Increase…you have used your in this sentence. You may need to use you’re instead. I don’t think, I don’t think I need to change this word because your is an adjective. Earning power is a noun and your is a possessive. I don’t I don’t think I need to change this word.

Her resistance to making a change as suggested by Criterion was justified, as she did with the two other spelling false positives. The only exception was that of the other confused words error tag when An did not read the pop-up explanation, which she admitted forgetting to do during the stimulated recall. Instead, by simply seeing the highlighted “to”, she misread it for an ill-formed verb error and responded accordingly:

**First draft:** In brief, no one can help you to choose which career is suitable for you.

**Error message:** You have used to in this sentence. You may need to use two instead.

**TAP excerpt:** OK, help someone do something, not to. I will delete the to.

**Revised draft:** In brief, no one can help you choose which career is suitable for you.

Although An could have simply ignored the false positive from Criterion instead of spending time working on a revision in this case, her revised draft is another correct structure in place of the original one, given the fact that these two alternatives (i.e., help someone + infinitive vs. help someone + to infinitive) are both grammatically correct. This correction, however, was a diversion from the focus in Criterion’s error tag.

Overall, observations about An’s handling of the unhelpful feedback from Criterion suggest that this student was able to diagnose the redundancy of false positives, but her judgement of the incorrect error codes was insufficient. This also had partial impact on how deeply the feedback was processed as well as the efficacy of her revisions accordingly. Regarding false positives, An was sceptical about the value of Criterion feedback and seemed to realize its false tagging of correct structures in her text. That was why she substantively noticed these error tags by referring to previously acquired linguistic knowledge to reject the feedback. It shows her understanding of the
structures about which she had confidence in her own knowledge base to refuse up-take of *Criterion* ACF. However, the incorrect error codes were more challenging for An as she did not seem to evaluate the feedback to a satisfactory extent. Instead, her main approach was following *Criterion*’s suggestions, which was not helpful for revision due to the insufficient, even detrimental guidance in some of the feedback.

In summary, An possessed sufficient linguistic resources and some competence in feedback evaluation as evidenced from her writing, effective revisions using *Criterion*’s simple error codes, as well as confident reference to stored metalinguistic knowledge. However, probably overwhelmed by the number of error tags received and partly due to her personal focus on getting rid of all surface-level errors rather than looking at higher order thinking skills in writing during the revising stage, she did not fully mobilize her potential to conduct evaluation of error messages generated by *Criterion* to a sufficient depth. Accordingly, she did not process the error codes to a satisfactory extent at times.

**PART 4. Cross-case Analyses**

In this section, quantitative data of the three selected cases are summarized before the qualitative cross-case analyses are presented.

**Summary of Quantitative Findings**

Trang, Nhien, and An were selected based on five criteria set out for the multiple case studies. Among these five, one key criterion that differentiates them was their accuracy change trajectory. Specifically, Trang experienced increased accuracy while Nhien and An shared gradually decreased accuracy after the five-month period. The findings presented in this chapter also reveal differential engagement levels as well as varied revision patterns using *Criterion* automated corrective feedback among these three students. Before proceeding with the cross-case qualitative analyses, it is worth summarizing major quantitative data that characterize the three cases. Table 6.20 includes statistics that reflect each individual case’s performance, ranging from general measures of timed test essay length and practice essay normed error rate to feedback-processing statistics (i.e., engagement time duration, noticing level, and revision outcomes).
### Table 6.20

**Summary of Cases’ Findings**

<table>
<thead>
<tr>
<th></th>
<th><strong>Trang</strong></th>
<th><strong>Nhien</strong></th>
<th><strong>An</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average word length in timed test essays</td>
<td>334</td>
<td>397</td>
<td>387</td>
</tr>
<tr>
<td>Average word length in practice essays</td>
<td>400</td>
<td>509</td>
<td>489</td>
</tr>
<tr>
<td>Adjusted normed error rate for practice</td>
<td>1.9%</td>
<td>2.9%</td>
<td>2.3%</td>
</tr>
<tr>
<td>essays (spelling errors not included)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Averaged revision time per practice essay</td>
<td>25mins</td>
<td>23mins</td>
<td>14mins</td>
</tr>
<tr>
<td>(rounded)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extensive engagement</td>
<td>0%</td>
<td>52%</td>
<td>51%</td>
</tr>
<tr>
<td>Limited engagement</td>
<td>100%</td>
<td>48%</td>
<td>49%</td>
</tr>
<tr>
<td>Revision success rate</td>
<td>83%</td>
<td>88%</td>
<td>84%</td>
</tr>
<tr>
<td><strong>Observed improved fluency?</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Observed increased complexity?</strong></td>
<td>Varied</td>
<td>Varied</td>
<td>Varied</td>
</tr>
<tr>
<td><strong>Observed accuracy gains?</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

In brief, Trang wrote relatively shorter essays and also made fewer errors. Yet, she spent a slightly longer time period revising than Nhien and much longer time than An. The difference in revision time durations among these three learners were explicable as Trang and Nhien attended to local and global revisions through both *Criterion*-initiated and self-initiated revisions, while An focused exclusively on responding to the error tags generated by *Criterion* in her essays and worked at surface level revisions. Interestingly, Trang demonstrated limited engagement with *Criterion* ACF throughout her coded engagement episodes, while the proportion of substantive and perfunctory noticing episodes were more balanced for both Nhien and An. Revision success was not markedly different for the three cases, with Nhien as the one with slightly greater revision success. However, only Trang, being the case with a lower baseline (at 79% overall accuracy at pre-test), gained in accuracy over the studied period while both Nhien and An experienced decreased accuracy. These opposite accuracy change trajectories are discussed more extensively in Chapter 7.

Along with changes observed in accuracy, changes in fluency and complexity among these three learners somewhat varied. Generally, Trang wrote shorter essays judged by average word counts in both test and practice essays, at approximately 20% fewer words than Nhien and An. All of the three cases also showed increasingly greater essay length over time. Linguistic complexity did not follow a clear pattern of change across the three learners, but overall figures suggest that their essays had almost equal syntactic complexity. On the other hand, Trang used a slightly more...
diverse lexical range and higher percentage of academic words in her texts.

**Qualitative Cross-case Analyses**

The three learners selected for this multiple-case study shared a few characteristics as they worked on revisions using *Criterion ACF*. However, they also demonstrated interesting differences in their perception, engagement, and revision patterns that may have had implications on the observed accuracy change trajectories each of them experienced.

**Similarities.** A few similarities emerged from the cross-examination of the three cases’ data. Firstly, all of the learners responded well to false positive feedback. That is, each learner was able to retain their original text irrespective of the feedback and suggested change provided by *Criterion*. This was not the result of skipping the error tag or simply giving up making a revision. Instead, false positive feedback points were noticed and deeply processed which resulted in well-informed decisions to make no change among all of the cases. In other words, the ‘No change’ response as retention of correct forms mostly came from extensive engagement with the automated feedback, which shows the learners’ competence of feedback evaluation, at least in the cases of false positives.

The second similarity concerns the learners’ attitudes towards the automated corrective feedback they received. A common feeling expressed by the learners was that they trusted *Criterion ACF*, finding the automated corrective feedback an extra and helpful resource that facilitated their revision activities. Nhien’s comment cited from her stimulated recall below is exemplary:

> *I think it’s quite good and useful when I want to check some basic errors like grammar, vocabulary, spelling, fragment…. [Compared to the traditional way of writing on paper] I think it’s quite good when I do all the searching online so I can use the Internet for searching the vocabulary, some dictionaries online.*

Turning to the learners’ failed attempts at revisions, there were two common reasons. Firstly, *Criterion*’s automated feedback on certain errors was incorrect. In such cases, learners could find neither sufficient guidance nor helpful information about the cause of their errors. Examples for this are An’s failed attempts at revising the ill-formed verbs error on “might not be do success” and determiner-noun agreement tag about “staying a jobs”. Due to the lack of sufficient scaffolding for correction, learners could not come up with a correctly revised form just by referring to *Criterion ACF* and its metalinguistic explanations. The second reason for learners’ failure was gaps in their
knowledge, which led to unsuccessful revisions even in cases of correct error codes. Illustration for this can be found in the error tag on confused words “effect” and “affect” provided to Nhien or the error tags on missing/extra articles for Trang presented above.

**Differences.** The three selected cases can be differentiated by their perceptions about the revising stage as well as engagement and revision approaches they adopted when processing *Criterion* automated corrective feedback.

**Different perceptions about revisions.** Trang considered revising using *Criterion* ACF as “a further way to get a higher mark” (stimulated recall interview). She believed that by addressing the errors *Criterion* pointed out to her, she would receive a higher score from the system. This explained why she adopted a suggested change against her own belief. She also exercised self-evaluation of her writing style and grammatical accuracy, explaining that she was not always satisfied with all the things she wrote in her first drafts. Therefore, revision activities gave her the opportunity to address these issues.

Unlike Trang, An’s aim was to correct all the errors until her text was free of *Criterion*-generated error tags, and to ultimately find out the kinds of errors she made most frequently. This, according to An, helped her to improve her knowledge about writing. Her focus on grammatical accuracy was further substantiated by the fact that she had previously used Grammarly to check her essays. In addition, in response to the large number of error tags (53) she received from *Criterion* across three practice essays, she submitted two revised drafts each session to make sure that she did not miss responding to any error tag. Her revisions were exclusively form-focused initiated by *Criterion* feedback.

On the other hand, Nhien seemed to consider revising using *Criterion* ACF as a learning opportunity to build up her linguistic repertoire through reference to different online resources. Nhien had a more systematic approach to writing and revision when she commented that during the revising stage, she would focus on grammar more than the content of her essay. To explain her personal approach, Nhien added that, “when I’m writing maybe it can be affected much by speaking grammar, so I want to revise it to writing grammar.” To her, content was focused on during the pre-writing stage when she brainstormed ideas. Therefore, revising to her was a chance to refine the linguistic devices used to express these ideas.

**Differential engagement levels with Criterion ACF.** The learners’ different perceptions of the use of automated feedback and their learning goals determined how each learner chose to
engage with it. For example, when engaging with a feedback point, Nhi substantively noticed the error being detected and processed the feedback with reference to external resources as well as previously learned knowledge, while An was more likely to pay perfunctory attention to the error messages as she was focused on the immediate task of addressing all the tagged errors instead of thinking of longer term learning goals. Similarly, driven by the extrinsic motivation to “satisfy” Criterion and get a higher score, Trang chose perfunctory noticing as a way to process all the given feedback. Therefore, her revising was for the short-term goal of addressing the errors pointed out.

In addition, in coded revision episodes, Nhien and An gave hints about the detrimental impact of the cognitive overload they experienced when processing the large number of error codes. For example, such cognitive overload sometimes discouraged An from having extensive engagement with the feedback as well as making substantive revisions to the essays. In certain cases, she did not attempt to revise the errors the first time she processed them. Instead, after revising several other errors and reducing a considerable number of error tags, she was able to notice the remaining ones more substantively and allocated more cognitive efforts to making revisions. No such issue was recorded for Trang who got much fewer error codes. Instances of cognitive overload can be partially attributed to Criterion’s comprehensive feedback mechanisms which may overwhelm learners with so many error tags at the same time.

**Revision patterns.** All of the three selected cases demonstrated satisfactory revision outcomes with success rates exceeding 80%. However, depending on personal priorities as earlier mentioned, each learner took a different approach to revising their essays, most clearly seen in the two aspects of textual level of revision and time spent on revising.

**Textual levels of revision.** As seen from the previous sections on each individual case’s findings, there seems to be a consistent relationship between the level of textual change made and how that revision was initiated. That is, if students were revising as part of their response to Criterion automated corrective feedback, they tended to make surface-level and local changes to their texts. On the other hand, if the revisions were self-initiated, they were most likely to be substantive and global changes. This pattern in students’ revision practices reflected the fact that Criterion ACF targeted the superficial and local issues in learner essays. What can also be inferred is that Criterion with its form-focused feedback has encouraged not only the editing of local problems pointed out by the system but also self-initiated revisions on more substantive issues in the essays among Trang and Nhien. With that said, Trang made most substantive revisions and devoted more effort to higher-order concerns in writing compared to the quick and superficial
textual changes in An’s essay. The third case, Nhien, tended to take a balanced approach by both correcting linguistic issues and improving textual features to ensure clarity of idea expressions.

_Timing._ Although the number of error tags received by Nhien and An was quite similar, at around 20 for each essay, the time they spent on revisions was different. Nhien adopted a more careful approach by going from one error to the next and sought extra help from online resources to double check her revisions. During the 17.5 minutes devoted to addressing _Criterion_-generated error tags, Nhien usually came back to the errors that she struggled with more than once to reconsider ways to deal with them. The remaining 5.5 minutes for self-initiated revisions was mainly about proofreading the whole essay and making changes where needed. Unlike Nhien, An did not make any revisions other than the issues pointed out by _Criterion_. She spent the total 14 minutes addressing the automated error tags. Her revision choices were based on the metalinguistic explanations and suggestions from _Criterion_ which she examined to make quick decisions on revised forms. Besides that, the other major resource was her stored metalinguistic knowledge. She made occasional references to online dictionaries and resources for correct forms, mainly in response to spelling errors.

Among the three learners, Trang received only about half the number of the error tags received by Nhien and An. Despite this, Trang’s revision time was longest. Interestingly, she chose to engage superficially with the corrective feedback, which explained why the actual time duration Trang allocated to addressing _Criterion_-generated feedback was less than half of her total revision time. She focused more on making self-initiated revisions which ranged from lower level changes (lexical and grammatical features) to higher-order issues of writing such as stylistics and language register. Probably with significantly fewer error tags to attend to compared to Nhien and An, Trang’s focus during the revising stage seemed to rest more with substantive issues that aim at improving the essay’s overall quality.

**Summary**

The cross-case analyses have provided answers to the first two sub-questions which seek information about (RQ3a) how the three selected cases were similar and different in terms of their engagement and revisions using _Criterion_ ACF and (RQ3b) the reasons for these learners’ failed attempts at revisions using _Criterion_ ACF as a resource. The third sub-question (3c) about the factors that contribute to differential engagement and revision patterns observed among the selected cases was addressed by the synthesis of the three cases’ findings. Taking a holistic view of all the data presented above, some factors that contribute to the outcomes of revision and the longer term change
in writing accuracy could be inferred, including (a) learners’ goals, (b) feedback type, (c) error complexity, (d) learners’ competence to conduct evaluation of feedback, and (d) their revising strategy. The next key question to be answered is how these factors interact for successful or unsuccessful revision outcomes, and how that relates to their longer-term acquisition of targeted morphemes as well as overall accuracy. An in-depth discussion regarding this question will be provided in the next chapter.
Chapter 7
DISCUSSION OF FINDINGS

Introduction
This chapter positions all of the findings reported in Chapters 4, 5, and 6 in a structure that holistically evaluates the usefulness of Criterion automated corrective feedback for EFL learners in a tertiary writing classroom context in Vietnam. In order to do this, the discussion starts with reviewing the claims made by Criterion’s developers regarding its use and value for students. This is followed by the development of an argument-based structure that binds all the findings together in a systematic approach to validating the evidence against the claims made about Criterion. Findings are summarized and mapped onto relevant inferences in this argument structure before all the backing and rebuttal evidence is evaluated to make conclusions about this automated corrective feedback tool. Towards the end of the chapter, some initial practical and theoretical implications will be discussed, and areas for future research identified to address certain limitations in the current study.

Developers’ Claims about Criterion
On ETS Criterion official website, a few claims are made about the use of Criterion Online Writing Evaluation Service. The following excerpt has been intentionally taken as it is most relevant to the discussion regarding the value of Criterion automated corrective feedback for students’ learning:

It [Criterion] allows students to improve their writing skills while working independently with immediate, detailed feedback on grammar, spelling, mechanics, usage, and organization and development. Students get more writing practice without adding to instructor workload, and instructors can concentrate on the content and style of students' work and teach higher level writing skills. (ETS Criterion official website, https://www.ets.org/criterion, 2019)

From the above quotation, four expected learning outcomes among students thanks to the use of Criterion automated feedback can be inferred:

• Improved writing skills
• Greater self-regulation of one’s writing and revisions
• Greater learner autonomy
• More regular writing practice
Using these learning outcomes as a starting point, this chapter will put together the findings earlier presented in Chapters 4, 5, and 6 in a systematic argument that aims to examine whether Criterion has fulfilled the claims made by its developer. More specifically, the discussion of the findings adopts an argument-based approach proposed by Chapelle, Cotos, and Lee (2015) to validate Criterion automated corrective feedback as a formative assessment and learning tool in the EFL writing classroom.

**Developing an Interpretative Argument**

The current research was not originally designed to be a validation study. However, with the findings presented in Chapters 4 and 5 regarding the nature of Criterion automated corrective feedback, EFL students’ engagement, perceptions, and use of this feedback, as well as in-depth observations about individual learners’ differential engagement and response to the feedback in Chapter 6, it is helpful to frame the discussion within an argument-based approach to validation. The expected learning outcomes inferred from ETS claims above relate to two inferences from Chapelle, Cotos, and Lee’s (2015) argument-based validation of formative assessment using AWE. The first inference is utilization with the underlying warrant that “Diagnostic results on the quality of academic writing obtained from Criterion are useful for students to make decisions on revisions” (p. 387). The second relevant inference is ramification which is justified by the warrant that the “use of the system is beneficial for learning” (Chapelle et al., 2015, p. 387).

In Chapelle et al.’s (2015) utilization inference, the focus is exclusively on the usefulness of AWE for students’ decision-making processes to arrive at revisions in their essays. However, as Ranalli et al. (2017) pointed out, usefulness as operationalized in other classroom-based research into AWE includes how clear, specific, and relevant the feedback is, and the range of features addressed by such feedback. Taking this broader view of usefulness, the current study examines the accuracy and coverage (i.e., the linguistic areas that are addressed by the feedback) of Criterion ACF using what Ranalli et al. (2017) called “user-centric standards” which imply a coherent interwoven relationship between accuracy and usefulness. This aims to address a criticism levelled against previous research into automated feedback that AWE accuracy and usefulness have mostly been investigated as separate entities and that developer-centric standards dominated the field where the “question of how accuracy problems may affect usefulness” (p. 9) has been neglected.

Therefore, before validating Criterion ACF using the utilization inference, it is essential that the accuracy, coverage, and appropriateness of Criterion automated corrective feedback are evaluated. Thus, the evaluation and explanation inferences are added. The evaluation inference is justified by the warrant that “Criterion feedback provides students with accurate information to
target relevant areas for revision and learning” (Chapelle et al., 2015, p. 387), while the warrant for the explanation inference is adapted from Chapelle et al.’s (2015) as “diagnostic results given by Criterion reflect the aspects of a construct of academic writing ability” targeted by the writing course in the current research, which focused on writing academic essays. In ETS Criterion claims, the mention of “immediate, detailed feedback on grammar, spelling, mechanics, usage, and organization and development” highlights the developers’ beliefs in the strengths of Criterion’s automated feedback: immediacy and elaborateness. Despite the absence of any mention about accuracy or coverage from ETS in the above excerpt, this research considers accuracy and coverage of relevant linguistic issues as a prerequisite for any feedback to be useful for learners.

Regarding the next inference relevant to ETS claims, ramification, Criterion developers seem to place great weight on the benefits of the system for learners’ long-term writing development. By claiming that Criterion “allows students to improve their writing skills while working independently with immediate, detailed feedback on grammar, spelling, mechanics, usage, and organization and development”, they allude to both improved writing skills and greater learner autonomy in writing and revisions. In addition, Criterion, with its “immediate, detailed feedback on grammar, spelling, mechanics, usage, and organization and development”, allows more regular self-practice for learners, and thus greater space for writing instructors to focus on more substantive areas of students’ writing.

The discussion addresses only four of the seven inferences in Chapelle et al.’s (2015) argument-based validation framework due to the exclusive focus on automated corrective feedback in the current study. The fundamental difference between the current study and that conducted by Chapelle et al. (2015) is the purpose and scope of using Criterion. In the current study, students were given access to Criterion for additional feedback on linguistic features as the system’s feedback functions on style and organization/development were turned off. Students had occasional and official writing practice sessions on Criterion to receive its form-focused feedback to conduct revisions, while still receiving teacher written feedback on their essays. Although being put to use for quite similar purposes in Chapelle et al.’s (2015) study, students were encouraged to use Criterion as a self-study tool that provided around-the-clock evaluation of their writing. In addition, students in the latter study used Criterion in its totality with none of its functions or capacities suppressed.

As the domain, generalization, and extrapolation inferences deal with broader issues in the use of Criterion, they become irrelevant in the current study which more narrowly focuses on the nature of the corrective feedback of the system and how EFL learners engage with it to make revisions. In other words, this study isolates the form-focused feedback from Criterion for more in-
depth investigation of its value using the triad of product-process-perception perspectives. At the
centre of this triad is the learner, and the stimulus to be responded to comes in the form of the
automated corrective feedback from *Criterion*. Against this backdrop, what emerge as central issues
are the evaluation of the nature of the stimulus, then the utilization of this stimulus for revisions,
and its longer-term impact on accuracy changes.

The interpretive argument in the current research is, therefore, comprised of four inferences:
Evaluation, Explanation, Utilization, and Ramification. The adoption of an argument-based
validation of *Criterion* ACF examining these four inferences provides the chain that connects
different sources of validity evidence in the form of findings presented in Chapters 4, 5, and 6.
Warrants for each inference have been adapted, focusing exclusively on the corrective feedback
generated by *Criterion*. Underlying assumptions for each warrant have also been accordingly added
and adapted, with the incorporation of two assumptions for the utilization inference from Ranalli et
al.’s (2017) study. Table 7.1 summarizes the complete validity argument for this discussion.

Table 7.1

<table>
<thead>
<tr>
<th>Inference</th>
<th>Warrant that justifies the inference</th>
<th>Relevant assumptions underlying the warrant</th>
</tr>
</thead>
</table>
| Evaluation | *Criterion* ACF provides students with accurate and appropriate information to target relevant areas for revision and learning. | 1. *Criterion* feedback is precise.  
2. *Criterion* feedback is clear and specific. |
| Explanation | Diagnostic results given by *Criterion* reflect the aspects of the construct of academic writing ability targeted by the writing course. | 1. *Criterion* ACF sufficiently covers the range of linguistic areas relevant for the EFL learners in the writing course. |
| Utilization | Automated corrective feedback on grammar, usage, and mechanics errors obtained from *Criterion* are useful for students to make decisions on revisions. | 1. The meaning of *Criterion* feedback is clearly interpretable by students.  
2. Students engage with *Criterion* ACF in meaningful ways.  
3. *Criterion* feedback provides sufficient support for learners to correct the errors identified by the system. |
| Ramification | Use of Criterion ACF is beneficial for learning. | 1. Students gain mastery of grammatical features and increase grammar awareness.  
2. Students perceive the feedback positively.  
3. Frequent access to feedback maximizes learning opportunities.  

**Mapping Findings onto the Validity Argument**

In view of ETS claims about Criterion, the current discussion attempts to fit the available empirical data to an interpretative argument on the use of Criterion ACF as a formative assessment tool in an EFL tertiary writing course. The five research questions of the current study examined the nature of the feedback, EFL learners’ engagement with and perceptions of it, triangulated with the efficacy of their subsequent revision practices and long-term acquisition of targeted language forms. In a parallel manner, the four most relevant inferences from Chapelle et al.’s (2015) argument-based approach to the validation of formative assessment using AWE are examined in the current study: Evaluation, Explanation, Utilization, and Ramification inferences. The Evaluation inference relates to the accuracy and appropriateness of the feedback received by students as realized in RQs 1, 2 and relevant information from RQ5. Explanation focuses on the feedback coverage of relevant linguistic areas, found in the results of RQ1. Utilization is about the usefulness of the feedback for students’ decision-making process in reaching revision operations, as operationalized in RQ3 and data from the case studies. Finally, Ramification deals with longer term effects of the feedback on students’ L2 learning and acquisition, answered by RQ4, RQ5, and the multiple-case study. Tables 7.2, 7.3, 7.4, and 7.5 present the summaries of findings which have been mapped onto each inference in the validity argument developed for this study. The discussion also relates the findings to relevant perspectives in second language acquisition used in the study of corrective feedback as well as previous research on automated feedback to further reveal how this new form of feedback can be situated within the larger picture of corrective feedback research, and what it can do to enhance the expected learning processes and outcomes among learners, an ultimate goal pursued by all pedagogical choices.
**Evaluation Inference**

The gold standard for corrective feedback provided by CALL technology is to achieve a level of accuracy close to that of a learner-teacher interaction independent of whether the tools are used during teacher-directed classroom instruction or for self-study outside the classroom. (Heift and Hegelheimer, 2017, p. 61)

Evaluation is the inference that examines the accuracy, clarity, and specificity of Criterion ACF for the targeted learners in the current research: EFL tertiary learners of English in Vietnam. The inclusion of these characteristics of the feedback is based on user-centric standards as the current research was classroom-based with participants being end-users of this automated writing evaluation and instructional tool. Evidence for this inference is drawn from multiple data sources, including the verification of Criterion ACF, analyses of Criterion feedback generation mechanisms, the error types addressed by the system, as well as content analyses of the metalinguistic explanations it provided. Table 7.2 presents the summary of findings which have been mapped onto this inference.
Table 7.2
Warrant, Relevant Assumptions, and Summarized Evidence to Support or Reject the Warrant in the Evaluation Inference

<table>
<thead>
<tr>
<th>Warrant that justifies the inference</th>
<th>Relevant assumptions underlying the warrant</th>
<th>Summarized findings as backing/rebuttal for the assumptions</th>
</tr>
</thead>
</table>
| **Criterion ACF** provides students with **accurate** and **appropriate** information to target relevant areas for revision and learning | **1. Criterion feedback is accurate.** | - 82.2% overall correct error codes, 11.7% false positives, 6.1% incorrect error codes  
- Error types that passed the 80% precision threshold include *Duplicates, Faulty Comparisons, Garbled Sentences, Negation Errors, Wrong Article, Proofread this, Spelling, Determiner-noun Agreement, Subject-verb Agreement, Compound Words, Capitalization, Missing Final Punctuation, Pronoun Errors, Ill-formed Verbs, Run-on Sentences, and Missing or Extra Article.*  
- Error types that failed the 80% precision threshold: *fragments, confused words, preposition errors, extra commas, hyphen errors, possessive errors, missing comma, missing question mark.*  
- Students’ evaluation of the accuracy of *Criterion ACF* was mostly positive.  
- *Spelling and comma use* are the error types that had largest numbers of false positives. |
| **2. Criterion feedback is clear and specific.** | | - Grammar errors often received generic feedback, while mechanics and usage errors had much more specific than generic feedback.  
- The error explanation messages were repetitive and formulaic.  
- Metalinguistic explanations were generic or specific depending on error type instead of being based on the treatability or severity of the errors.  
- Students perceived the language used in *Criterion* error explanations to be clear and easy to understand.  
- *Criterion* feedback generation approach was perceived to be systematic, consistent, and easy to follow. |
Accuracy of Criterion ACF. Research Questions 1 and 2 sought answers about the nature and accuracy of Criterion’s ACF. Generally, Criterion ACF was found to pass the 80% precision threshold set by its developers (Quinlan, Higgins, & Wolff, 2009) as a prerequisite for corrective feedback from an AWE system to be useful for learners, with slightly higher overall precision (82.2%) compared to previous research (e.g., Lavolette et al., 2015; Ranalli, Link, & Chukharev-Hudilainen, 2017). The rate of false positives is found to closely match that reported in Lavolette et al.’s (2015) study.

Also similar to previous studies, varied precision across error types ranging from well under 50% to high percentages of over 95% was found. Being among the most problematic issues for EFL students and also the two most well-researched error types, articles and preposition errors were satisfactorily detected by Criterion, a result that reiterates previous research with around 70-80% precision (Lavolette et al.’s, 2012; Li, Lee, Lee, Karakaya, & Hegelheimer, 2011; Tetreault & Chodorow, 2008).

However, not all specific error categories have similar precision results across empirical studies. For example, Lavolette et al.’s (2015) findings indicated that Criterion’s detection of run-on sentences, wrong article, and spelling errors were below 50% precise, while these error types were satisfactorily identified by the system in the current study (80% or more precision). On the other hand, missing comma, the second least precise error type for Criterion in this research (28.7%), was found to be more than 85% successfully detected by Lavolette et al.’s (2015). While such differences are intriguing, they can also be indicative of changes that may have taken place in Criterion’s updated error detection algorithms over time, considering the four years’ gap between the two studies. In addition, the lack of consistent approaches to verify error codes across studies can also contribute to the different findings. Specifically, the current research considers comma usage to be a fluid area and that, as Lutovich and Chan (2000) commented, nothing is straightforward about it. Therefore, coding decisions were made to allow for some flexibility in evaluating error codes in this category, while a more prescriptive set of rules for comma use was adopted by Lavolette et al. (2015).

The inconsistent findings across studies in terms of overall precision and precision for individual error types should also be interpreted by taking into consideration different learner populations used as research samples. With the same AWE pedagogical tools such as Criterion, different learner groups with their levels of proficiency and L1 backgrounds are likely to significantly differ in the types of errors they make as well as the numbers of errors in each particular type. Considering Criterion’s formulaic approaches to error detection, besides being consistently successful at certain errors, it can also generate systematically flawed tags in certain
error types. For example, if a certain group of learners tends to make a substantial number of comma use errors, an error type *Criterion* was found to generate most false positives in the current study, then the precision level for this error type as well as overall precision will suffer.

For the sampled population of tertiary Vietnamese EFL learners in the current study, with precision being the measure, *Criterion* did a good job at detecting error types related to grammar and usage, although the system’s performance on a few error types in mechanics was poor due to its rigid treatment of writing conventions in academic English. With what has been found, *Criterion* can be a potentially useful formative assessment tool for EFL learners. This can offer some specific level of support for the assumption about feedback precision in the validity argument.

**Clarity of Criterion ACF.** In this study, *Criterion* was found to generate comprehensive automated corrective feedback on a range of 29 error types, categorized into three broad sections of grammar, usage, and mechanics. Besides highlighting the location of the error (a word, phrase, or whole clause), a metalinguistic explanation of the error was presented in a small pop-up screen. The way *Criterion* divided corrective feedback into three categories which were subdivided into individual error types was generally perceived as systematic and easy to follow among learners. Whenever students selected an error type, only error tags in that type were highlighted for revisions. This means that *Criterion* was designed to present errors in each type one at a time, which was perceived helpful for students to keep track of the errors they made most frequently and to improve awareness of their linguistic accuracy.

Yet, *Criterion*’s comprehensive approach to feedback generation is potentially problematic as it fails to focus learners’ attention on specific areas that deserve more attention due to their major impact on the readability and the overall comprehensibility of the essays. In addition, the system’s indiscriminate approach to feedback generation across all error types seemed to send learners the message that all errors should be treated equally. Less strategic learners may tediously work from one error to another, a process that could be discouraging and overwhelming, especially for those receiving a great number of error tags at a time. As seen from the findings, some writers could receive up to 66 error tags from *Criterion* in one draft. With this number of error tags to process in an essay of less than 500 words, some students would become demotivated through being overwhelmed by the amount of negative input they received, as commented by one of the focus group learners. Talking about comprehensive feedback in written error corrective more generally, Bitchener and Ferris (2012) warned that unfocused feedback may lead to cognitive overload among low proficiency learners if these learners try to attend to all the flagged errors.
Specificity of Criterion ACF. As mentioned earlier, each Criterion error tag both pinpoints the location of the error and provides a pop-up metalinguistic explanation about that error. Regarding the metalinguistic explanations, Mehrabi-Yazdi (2018) commented that, “The level of elaboration can vary, and that is the point where effective metalinguistic explanations can be distinguished from ineffective ones” (p. 94). Elaboration in Criterion ACF messages relates to a widely used term in written feedback, “explicitness”, whose impact on revision outcomes has been quite extensively researched in written corrective feedback (e.g., Bitchener & Knoch, 2009; Bitchener, Young, & Cameron, 2005; Ferris & Roberts, 2001; Sach & Polo, 2007; Sheen, 2007), but not much has been done on AWE corrective feedback. Different levels of explicitness closely relate to how well students can respond to the feedback and the level of attention they allocate to a certain error.

Against the backdrop of scarce prior research on how the explicitness in automated corrective feedback impacts learners’ revisions and acquisitions of linguistic forms, the current study adopted a binary terminology used by Ranalli et al. (2017), generic and specific feedback, to indicate two levels of explicitness in Criterion’s metalinguistic explanations that follow its error tags. For example, recent research conducted by Ranalli (2018) pointed out that specific feedback was nearly five times more likely to lead to successful revisions. In addition, specific feedback was four times more likely to be perceived as clear and three times more likely to be considered useful by the learners.

Findings presented in Chapter 4 showed that the metalinguistic explanations by Criterion can be generic or specific depending on error types. For example, in terms of grammar errors, most of the explanations are generic formulaic statements, while usage and mechanics error types receive mostly specific messages. With two thirds of error types receiving specific feedback, learners may potentially find Criterion ACF helpful for their revisions thanks to the specific suggestions for how to correct the errors. The remaining one third of all error types, including sentence structure errors (fragments, garbled sentences, run-on sentences, proofread this), subject-verb agreement, determiner-noun agreement, ill-formed verbs, wrong article, pronoun errors, negation, and spelling errors, receive generic error messages with a general indication of where and what the error is.

Taken at face value, it seems that as the feedback gets more specific, it is more conducive to effective revisions. On closer examination, however, the allocation of specific or generic feedback to different error types on Criterion was not that well thought out. In the first place, despite being consistently generated depending on the type of errors being tagged, the metalinguistic explanations were not sensitive to how treatable the errors are. Instead, the error tags and concomitant metalinguistic explanations were defined by “the set of algorithms to detect errors and generate
fixed metalinguistic feedback every time a particular error is indicated.” (Mehrabi-Yazdi, 2018, p. 93). Regarding this, Ranalli (2018) further pointed out two issues that may undermine the value of ACF for classroom use. Firstly, technological capacities currently available in AWE tools like Criterion rather than pedagogical considerations determine the number of informational clues available across different error types; and secondly, certain inaccuracies are considered part of the status quo in automated corrective feedback at this current stage. Due to technological limitations that come in scoring engines such as e-rater, the treatability of error types remains a minor consideration. For example, treatments of rule-based errors such as subject-verb agreement or determiner-noun agreement are suitably designed as generic, yet this generic approach was also unreasonably applied to less treatable errors related to sentence structures, including fragments, garbled sentences and suggested action proofread this.

The second reason for the inappropriate level of explicitness in Criterion ACF was that the severity of error types was not a determining factor in the design of its feedback. Let us continue the discussion about sentence structure errors such as garbled sentences which seem to cause major comprehension issues for readers. Considering its complex nature with serious impacts on the comprehensibility for readers, the choice of a generic explanation like “this sentence is confusing because it contains several grammatical or spelling errors” seems unhelpful for effective revisions. In contrast, the use of articles is rule-based and errors are less likely to impede comprehension. Yet, this error type receives specific messages. The message “You may need to use an article before this word. Consider using the article a” provides the writer with an explicit editing suggestion. The implementation of inserting the article ‘a’ as suggested seems quite straightforward with little need for cognitive processing. If feedback of a much more explicit nature as such was generated for complex error types related to sentence structures, it could more substantially support revisions among students.

Summary of Validity Evidence for the Evaluation Inference

This part of the study has discussed validity evidence related to the Evaluation Inference which is based on the warrant that Criterion feedback provides students with accurate and appropriate information to target relevant areas for revision and learning. Taken together, the results regarding the accuracy, clarity, and specificity of Criterion ACF, Criterion seems to be able to address the needs of EFL learners to some extent through its detection of surface-level errors. A mixture of supporting and rebuttal data for the assumptions underlying the evaluation inference has been discussed above. On the one hand, backing for the assumptions in this inference was found, including Criterion’s desirable level of feedback explicitness for a large number of error types that
prepare learners for revisions; clear, systematic, and consistent feedback generation approaches; and satisfactory precision levels on about two thirds of all the error types addressed.

On the other hand, rebuttal evidence relates to Criterion’s comprehensive and potentially overwhelming feedback, lack of pedagogically sound feedback decisions in terms of explicitness and scaffolding that take into consideration the treatability and severity of the errors. Some expected adaptations in Criterion’s technological capacities can help address these concerns. With regards to the overburden created by having to handle a large number of indiscriminately treated error tags, Criterion and similar AWE tools can be further equipped with functions to “toggle flagging for particular error types on and off” (Ranalli, 2018, p. 671). This allows writing instructors to place appropriate focus on certain linguistic aspects that are relevant for their learners at different stages in the writing proficiencies. Along with this, at the current stage in the AWE developmental history, some human intervention such as further teacher feedback and elaborations on errors tagged by Criterion can partially address the shortcoming of lack of pedagogical considerations in the system’s automated corrective feedback. In conclusion, although these potential rebuttals can undermine the validity argument about using Criterion as a formative assessment tool in EFL writing classroom contexts, these shortcomings are hopefully surmountable with ongoing efforts put into improving the technological capacities in AWE systems like Criterion.

**Explanation Inference**

*Both AWE tools [i.e., Criterion and MY Access!] failed to provide adequate feedback on some common errors made by learners. (Feng & Saricaoglu, 2017, p. 356)*

*Explanation* is the inference that examines the relevant linguistic areas covered by Criterion ACF for the targeted EFL learners in the writing course under study. Evidence for this inference is drawn from the tabulation of all the tagged errors in the corpus of essays written by the two writing classes who had access to Criterion over the studied period, content analyses of Criterion metalinguistic explanations, the error types addressed by the system, as well as comparison of this range with the common error types reported in previous research on both AWE and teacher written corrective feedback. Table 7.3 presents the summaries of findings which have been mapped onto this inference.
Table 7.3

Warrant, Relevant Assumptions, and Summarized Evidence to Support or Reject the Warrant in the Explanation Inference

<table>
<thead>
<tr>
<th>Warrant that justifies the inference</th>
<th>Relevant assumptions underlying the warrant</th>
<th>Summarized findings as backing/rebuttal for the assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic results given by Criterion reflect the aspects of a construct of academic writing ability targeted by the writing course.</td>
<td>1. Criterion ACF sufficiently covers the range of linguistic areas relevant for the EFL learners in the writing course.</td>
<td>- Criterion generated feedback on 29 error types in learner essays: Mechanics (48%, most spelling errors), Usage (42%, most missing/extra article errors), Grammar (10%, most subject-verb agreement errors) - The range of errors detected by Criterion generally match those found to be frequent error types among English language learners, such as spelling, preposition, subject-verb agreement. - Criterion did not address three major issues found for ELLs in previous studies: word choice, word order, and verb tenses.</td>
</tr>
</tbody>
</table>

Range of error types addressed and second language learners’ major issues. The range of errors detected by Criterion generally matches those found to be frequent errors among English language learners, such as spelling, preposition, or subject-verb agreement as found in previous studies (e.g., Barzanji, 2016; Phuket & Othman, 2015, Sumalinog, 2018). In addition, among the most challenging structures for ESL learners (Chodorow, Gamon, and Tetreault, 2010; Han, Chodorow, & Leacock, 2006; Jordan, 2004) are articles. Article errors were also the second most frequently detected error type in the current study, to be preceded only by spelling errors. Criterion identified the most prominent problems encountered by the EFL learners participating in this research, which is representative of the larger English language learner population as reported in earlier studies (e.g., Chen, Chiu, & Liao, 2009; Ranalli et al., 2017).

Specifically, the range and frequency of errors reported in the current research partially match what has been reported by other studies. For example, Chen, Chiu, and Liao (2009) found that the five most frequently identified error types by Criterion in a small corpus written by EFL Taiwanese university students were spelling, missing or extra article, subject-verb agreement, fragments, and...
run-on sentences. Except for the first three error types, fragments and run-on sentence errors in the current study were not as commonly detected. Ranalli et al. (2017), who studied Criterion error tags in ESL learners’ essays in a mid-western US university setting, provided a very similar list of frequent errors to that found in this study. For example, Spelling and Missing/extra articles received the largest numbers of error tags from the system, followed by other error types in the highest frequency group including comma usage, preposition, fragments, subject-verb agreement, ill-formed verbs, and determiner-noun agreement.

However, the relevance of Criterion ACF for English language learners is compromised by its failure to address some major challenges experienced by these learners. The major concerns found in the writing of English language learners in previous studies include word choice (e.g., Barzanji, 2016; Phuket & Othman, 2015), verb tenses (e.g., Lavolette, Polio, & Kahng, 2015; Phuket & Othman, 2015; Sumalinog, 2018), and word order (e.g., Dikli & Bleyle, 2014). The following discussion examines each of these concerns in greater detail.

**Word choice:** Meaning is still a missing component in Criterion’s feedback. Word choice was frequently mentioned by students in the focus group interviews as an area on which writing instructors’ feedback was much more useful. Such findings reiterate Dikli and Bleyle’s (2014) comment on Criterion’s unreliability in detecting several word choice errors made by the ESL learners in their study, adding that it is a challenge for a computer software to understand the semantic relationships between words.

Understandably, word choice is more about the meaning of texts than form-focused considerations. Therefore, it poses a challenge for the technical capacities that make up AWE systems. Feedback on vocabulary from Criterion currently deals with simple issues with four relevant error types: wrong form of word (e.g. transportation vs. transport), confused words (e.g. affect vs. effect), wrong or missing word (e.g. they vs. the), and wrong part of speech (e.g. adjective vs. adverb as in effective way instead of effectively way). Although these error types have touched on certain characteristics of diction, most of the tagged errors are still grammatical in the sense that Criterion based its detection on algorithms that work out the fixed collocations and syntactic structures that determine the combination of words in a phrase or sentence. Additionally, these error types are mainly about the right and wrong use of a lexical item in relation to adjacent items. However, more subtle word choice issues require careful consideration of meaning and of how that meaning can be effectively communicated through alternative vocabulary choices, which goes beyond the question of being right or wrong.
Verb tenses: Limited textual levels are targeted. Criterion primarily focuses on superficial error types at lexical or phrasal levels. It fails to deal with errors at inter-clausal or inter-sentential levels. This is most salient in Criterion’s total failure to address errors related to the tense-aspect (TA) system in English, a finding that corroborates Dikli and Bleyle’s (2014) and Lavolette et al.’s (2015) finding. For example,

Student text:  
All of the members in my family were very proud of the effort which she (I) make.

Error messages:  
(1) The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb.

In the error tag of subject-verb agreement above, what is suggested is a change from “make” to “makes” to ensure agreement between the third person singular subject she and its verb. What was overlooked by Criterion is the relationship between the tense usage in the first and second clauses, which should instead be consistent as indicative of past events, and thus, “make” should be “made” accordingly. A similar case is evident in the sentence below:

Student text:  
To me, when I finished my university, I will live in a different city.

No error message was generated in response to this sentence. The system’s failure to read texts at the inter-clausal level resulted in an oversight of tense consistency between the two clauses. Verb tense decisions require contextual clues at sentential or inter-sentential levels where meaning becomes a key component that determines the grammatical forms. This omission is made worse by the fact that there exists a great distance between the English tense-aspect system and that of Vietnamese, the participants’ first language in the current research. Interference from L1 such as Vietnamese where past tense is not an inflectional morpheme has resulted in frequent errors when Vietnamese EFL learners write in English (Dam, 2001; Dao, 2018). Unlike third person singular, which is rule based and can be treated by matching the subject with the verb in a clause, verb tense errors are less treatable due to the much wider textual and contextual clues that they take to be remedied. This probably explains Lavolette et al.’s (2015) observation that Criterion misinterprets errors related to tense usage as subject-verb agreement errors.

Word order: Criterion’s failure to fully capture syntactic structures in English also resulted in the total absence of error tags on faulty word order in learner essays, a finding that reinforced
previous research (e.g. Dikli & Blyle, 2014). The following example shows one of Criterion’s incorrect error code when it provided a wrong suggestion for revision:

Student text: This is very useful and meaningful activity.
Criterion error message: You may not need to use an article before this word.

The most straightforward way to revise the sentence is inverting the positions of very and the article a so that the word order of the nominal phrase a very useful and meaningful activity is grammatically correct. However, Criterion’s error detection algorithms seem to work at lower textual levels than this long phrase, resulting in its simplistic recognition of an adjective being modified by an adverb, and this leads to its incorrect suggestion to delete the article a.

Similarly, the next example illustrates another of Criterion’s incorrect suggestions due to its inability to detect the wrong word order in the learner’s sentence:

Student text: Air has role important in our life.
Criterion error message: You may need to use an article before this word. Consider using the article a.

If word order was part of Criterion’s error detection mechanisms, the correct suggestion for this error would be a two-step remedy: (1) inverting the noun and its modifying adjective, then (2) adding the article an to make it an appropriate noun phrase in English. Based on several incorrect error codes similar to the examples above, it is reasonable to say that Criterion error detection remains at a low textual level of two-word combination or certain fixed phrases. Higher syntactic structures at phrasal, clausal, and sentential levels were beyond the system’s capacities.

Summary of Validity Evidence for the Explanation Inference

This part of the study has discussed validity evidence related to the Explanation Inference which is based on the warrant that diagnostic results given by Criterion reflect the aspects of a construct of academic writing ability targeted by the writing course. The results regarding the sufficiency and relevance of Criterion ACF suggest that Criterion seems to be able to address the needs of EFL learners to some extent through its detection of a range of error types commonly found in English language learners’ essays. However, significant rebuttal evidence for this inference relates to the absence of automated corrective feedback on the key error types of verb tense, word choice, and word order. This calls for EFL writing instructors’ attention to make up for these gaps
in *Criterion* ACF to meet students’ needs at the current stage in the development of AWE systems like *Criterion*. Such human intervention still proves to be a critical part of feedback provision practices in the EFL classroom where the focus on form is as important as idea and content developments.

**Utilization Inference**

If AWE is to realize its potential as an instructional tool in classroom writing instruction it needs to continue its transition from being a sophisticated text editor that assists students in correcting their errors to a tutoring system that assists students in developing the skills and strategies needed to effectively independently revise their texts. (Stevenson, 2016, p. 12)

Utilization is the inference that examines the usefulness of *Criterion* ACF as demonstrated in students’ uptake of the feedback; therefore, evidence is drawn from students’ cognitive engagement with the feedback, the subsequent revising strategies, text operations and revision outcomes. Such evidence comes from the findings for Research Question 3 which investigated the short-term impact of using automated corrective feedback from *Criterion* during the revising stage among students. In this section, the interaction between key factors from the review of the literature: error type, feedback type, and learner goals during learners’ engagement episodes with *Criterion* ACF and their subsequent revisions are discussed to evaluate the usefulness of the feedback. This also serves as validity evidence to rebut or confirm the assumptions that underlie the warrant for the Utilization Inference. Table 7.3 presents the summaries of findings which have been mapped onto this inference.
Table 7.4
Warrant, Relevant Assumptions, and Summarized Evidence to Support or Reject the Warrant in the **Utilization** Inference

<table>
<thead>
<tr>
<th>Warrant that justifies the inference</th>
<th>Relevant assumptions underlying the warrant</th>
<th>Backing/Rebuttal for the assumptions (Summarized findings)</th>
</tr>
</thead>
</table>
| Automated corrective feedback on grammar, usage, and mechanics errors obtained from **Criterion** is **useful** for students to make decisions on revisions | 1. The meaning of **Criterion** feedback is clearly interpretable by students. 2. Students engaged with **Criterion** ACF in meaningful ways. 3. **Criterion** feedback provides sufficient support for learners to correct the errors identified by the system. | - Students reported minor issues in understanding the metalinguistic explanations given by **Criterion**, but overall comprehension of the messages was satisfactory.  
- Students deployed a variety of strategies in responding to **Criterion** ACF and revising their essays, but the strategies showed more limited than extensive engagement with the feedback.  
- Text operations are mainly the substitutions of single words, or in a few cases of short phrases. It is rarely a substitution of longer phrases or whole clauses. This is also true with the addition or deletion text operations, which are at the levels of single lexical items or short phrases.  
- **Criterion** corrective feedback dealt with local and surface-level errors. Therefore, major revisions were not likely to be initiated by **Criterion**.  
- When diagnosing more global errors such as *proofread errors or fragments*, **Criterion** provided generic messages which normally did not help students see the nature of the errors they made, and thus no useful support was given to aid effective revision activities. |
4. The need to differentiate between accurate and inaccurate feedback does not overburden learners’ cognitive-processing capacities, which could affect their willingness to use the feedback.

- Revision outcomes: Correct revision 54%, incorrect revision 10.5%, avoidance to address the flagged errors 28%, retention of correct forms 7.5%.
- Successful revision outcomes varied across error types, with most successfully revised errors being *Spelling, Subject-verb agreement, Garbled sentences, Capitalization, Determiner-noun agreement*, and *Ill-formed verbs*.
- Students performed worst at revising structures receiving *fragments, missing final punctuations*, and *proofread this* error codes.

- The false positives and incorrect error codes tended to trigger substantive noticing and deeper processing.
- Certain students were overwhelmed by the number of error tags they received from *Criterion*. This, however, did not prevent them from processing and responding to the feedback.
- Students responded appropriately to false positives, but their revisions based on incorrect error codes were mostly incorrect.
The comprehensibility of *Criterion* metalinguistic explanations. Although students generally perceived the metalinguistic explanations from *Criterion* to be clear and easy to understand, there seemed to be a discrepancy between their understanding of the error messages and subsequent capacities to implement revisions. On the surface, the words and sentence structures *Criterion* used to explain errors were perceived to be simple and comprehensible among learners. However, comprehensibility should not be viewed in isolation from a number of different factors which interact and determine how well a learner understands the nature of the error and how that understanding contributes to the revision outcome.

Firstly, the genericity and brevity of the information provided in *Criterion*’s metalinguistic explanations caused difficulties for students to sufficiently understand the nature of the errors they made. A student’s comment reflects this concern among learners, “I think the explanation is quite short. For some errors, we did not know exactly what was wrong about them, we had difficulty correcting the errors.” The idea was echoed in Zhang and Hyland’s (2018) study on an AWE program called *Pigai* developed in China, “When comments were too generic or formulaic, she [the participant] felt disgruntled and disappointed” (p. 322). In the end, what matters most to learners is not only the language used to explain the errors but also the sufficiency of information as well as its explicitness about that error.

The next question to be answered is how much elaboration would be enough for certain errors. To effectively address this question, we can look at the interaction between error types and feedback types that have emerged in the automated corrective feedback literature as significant factors which determine the revision activities among learners (Ranalli, 2018). Depending on the error types, some rule-based errors can be easily addressed by highlighting the error followed by a coherent explanation of the rules. On the other hand, error types which are case-specific (e.g. word choice, sentence structure, collocations) will take more than an indication and elaboration in *Criterion* metalinguistic explanations. Ideally, for most effective scaffolding to take place for these errors, clarification of the writer’s intended meaning is needed before suggested corrections can be provided. Yet, with the automatic algorithms and one-way communication characteristic of an online system like *Criterion*, there is no way to conduct such reciprocal exchanges before a resolution to an error can be reached. It is the absence of pedagogically sound decisions on feedback type-error type correspondence and lack of chances to follow up on an error that detrimentally impact the comprehensibility of the metalinguistic explanations from *Criterion*.

The interaction between error types and feedback types becomes more salient when we look at how the students responded to metalinguistic explanations from *Criterion*. For a fair number of surface-level error types detected by *Criterion*, superficial engagement with the feedback proved to
be sufficient for subsequent revisions. Most exemplary are errors in mechanics such as *spelling* or *missing final punctuation*. Because of the straightforward nature of these errors, a simple explanation is enough to draw learners’ attention to the error and to prompt an appropriate textual operation. However, for other error types like *articles* or *comma* usage, understanding the words in metalinguistic messages alone does not guarantee resolved errors. In such cases, it is reasonable to say that *Criterion* metalinguistic explanations are not comprehensible enough to be useful for L2 learners’ revisions, a conclusion reached by Zhang and Hyland (2018, p. 322) as well who found that “there was not much specific information for her [the learner] to take on board and act on”.

In appraising findings about *Criterion* ACF’s comprehensibility as an underlying assumption for its usefulness for students’ revisions, it seems that the system has done well for the majority of error types which address surface-level and local issues in learner essays. However, more rebuttal than support can be found for complex error types that need richer feedback and careful guidance for corrections.

**Meaningful engagement with *Criterion* ACF.** Participants in the current study deployed a variety of strategies to engage with *Criterion* ACF. Generally, their engagement strategies tended to be perfunctory rather than extensive. Perfunctory engagement strategies include *noticing without attending to the feedback* or *quick adoption of *Criterion*’s suggested change* without further actions, while extensive engagement was manifest through more substantive cognitive strategies, with the two most common being *looking up online resources* and *drawing on metalinguistic knowledge*.

Findings from the current research suggest that a number of factors interacted to determine how students chose to engage with the feedback. While the previous discussion pointed out the two factors of error type and feedback type, the case studies accentuate learner goals as another mediating factor that significantly shapes their engagement and textual changes.

**Learner goals.** The in-depth analyses of the three cases suggest that for a learning goal-oriented student, each error code is likely to be perceived as a skill and knowledge enhancement opportunity. In response to each error tag, students pursuing learning goals would examine its explanation and the suggested correction if given. They may also take a step further to either reject or confirm revised forms through reference to different resources. Exemplifying this engagement pattern is the case of Nhien who at times expressed her uncertainties about her knowledge. Nhien was most explicit about her self-perceived limited grammatical knowledge when writing in English, yet this low self-confidence did not discourage her from working diligently on correcting the errors with extended efforts to confirm the revised forms. This showed her pursuit of learning both self-
editing skills and knowledge through the utilization of more sophisticated and diverse strategies (e.g. looking up reference materials, seeking teacher advice, or translating her writing into L1) than the other two cases, despite the surface-level nature of the errors or the feedback provided by Criterion in several cases.

On the other hand, Trang was inclined towards performance goals as she focused on the grammatical accuracy of the essay with the aim to get a higher score from Criterion. Thus, this learner tended to pass up the opportunities to extend learning with more substantive engagement strategies. She only perfunctorily noticed the feedback and did not usually go further than addressing the error at hand. Somewhere in the middle of this goal orientation continuum is An whose ultimate aim was to improve her writing skills through correcting all the errors and getting a sense of the frequent errors she made. This explained her mixture of substantive and perfunctory engagement episodes with Criterion ACF. However, An’s perceived confidence in her skill level was higher than Nhien, and this probably accounted for the fact that her engagement strategies were less diverse with the main resource being her metalinguistic knowledge.

Overall, learner goals can be a good starting point to predict a learner’s engagement patterns with the corrective feedback provided by AWE systems like Criterion, as was found in earlier research on teacher written feedback (e.g. Han & Hyland, 2015). However, it is not a sufficient factor in itself as how a learner engages with the feedback is also mediated by the nature of the feedback, which is the focus of the next discussion.

**The nature of Criterion corrective feedback.** A few characteristics of Criterion ACF compromised students’ meaningful engagement with it. Cross-examination of Criterion metalinguistic explanations, student think-aloud protocols, their revisions, and submissions on Criterion revealed that the distribution of Criterion metalinguistic explanations was not meaningfully graduated in response to learners’ levels of L2 development. More specifically, Criterion-generated formulaic metalinguistic explanations treat errors of the same category exactly the same. Being machine-based, automated corrective feedback from Criterion is neither graduated nor tailored to individual needs. For example, if a student makes no revision in response to an error pointed out and tries revising the second time, Criterion does not adapt the error message to suit his/her actual need for guidance. Similarly, if a student provides an incorrect revision the first time, Criterion cannot provide a follow-up response to the same error based on the previous attempt. Instead, the incorrect form is treated as a new error.

Using Mehrabi-Yazdi’s (2018) terms, the text generated as written feedback to learner essays by Criterion “lacks cohesion and coherence for two reasons: first it is not staged and purposeful,
and second the rationale is not regulated and delivered with regard to what feedback has already been provided to the student in previous submissions” (p. 904). AWE systems like Criterion have not reached a stage when a learner’s developmental level becomes an integral part of its feedback generation mechanism to ensure that the feedback is noticed with understanding. Therefore, variable noticing and engagement levels among different learners are not taken into account, which may negatively impact uptake and longer-term acquisition. To make automated corrective feedback more useful for learners, AWE systems should aim for graduated assistance to expand the benefits of their multiple drafting function. Only a small number of participants in the current research submitted more than two drafts for their practice essays, but once the feedback gets more “graduated”, there are likely to be more submissions per essay among the learners.

An additional issue with Criterion corrective feedback is its lack of interactive communication. Obviously, for any corrective feedback to be facilitative of learners’ linguistic developments, it has to be attended to (Ellis, 2010). Appropriate support for revisions and learning, viewed from a socio-cultural perspective, is a prerequisite for feedback to be noticed and meaningfully engaged with, as Bitchener and Storch (2016) pointed out that the lack of properly staged and graduated feedback could negatively impact L2 development. With Criterion’s feedback generation approaches, however, there are no chances of extending the human-machine communication. It is a challenge to have really meaningful engagement with the feedback when Criterion still lacks the option for students to seek elaboration/clarification in case they need further information or rationale for the feedback received. Criterion ACF, therefore, seems to lack the key characteristics of useful assistance for L2 learners delineated by Bitchener and Storch (2016), which are graduated, contingent, and dialogic.

The cognitive load of evaluating feedback accuracy. An interesting finding from the current study showed that false positives and incorrect error codes were more likely to trigger substantive engagement among learners during the revision stage. The deeper engagement with these two feedback types both added to the time students spent on a revision episode and exerted an extra cognitive load on learners’ mental efforts. Researchers have been divided on the effects of the need to evaluate feedback accuracy, especially incorrect error codes, on learners. For example, Ranalli (2018) found that the extra mental efforts put into evaluating feedback increased learners’ cognitive load, which he hypothesized to be possibly connected with the low uptake rate among the participating students. On the other hand, Steinhart (2001, as cited in Stevenson, 2016) argued that flaws in the feedback provide students with the opportunity to evaluate how appropriate AWE feedback is.
What has been found in the current research showed that different learners demonstrated differential capacity and willingness to incorporate accuracy evaluation as part of their broader engagement process with Criterion ACF. More specifically, learners varied in both their allocated efforts to evaluate the feedback accuracy and the additional steps they took to process the error code they believed to be incorrect. For example, as illustrated in the case studies, Nhien, the student who tended to pursue learning goals, would first spend time reading the metalinguistic explanations from Criterion to evaluate both her writing and the feedback. This prepared her for the next step in choosing a cognitive strategy to help resolve the error (if the feedback was precise) or reject the feedback as inaccurate. On the other hand, An showed more confidence about her judgement of the feedback and quickly decided that she would not need to do anything further to address the error pointed out if she believed that an error tag was incorrect, an approach also adopted by Trang who demonstrated no effort to extensively engage with the feedback whatsoever. Instead, she tended to promptly see a feedback point as correct/incorrect and make quick revisions accordingly.

These learners’ different approaches to evaluating incorrect feedback can be better interpreted by looking at their revision outcomes in response to the incorrect feedback. Firstly, all of them were successful at recognizing false positives from Criterion. For example, Trang detected all of the eight false positives she received from Criterion. Despite the lack of explicit statements from this student about her perceived confidence in her writing / grammatical knowledge, we can conclude based on her successful response to all of these false positives that Trang was highly efficient in evaluating the accuracy of the feedback. Her engagement as well as revising decisions seemed quick and effortless, suggesting that her feedback accuracy judgment may be subsumed in the process of making sense of the feedback and of finding ways to revise her writing accordingly (Ranalli et al., 2017).

On the other hand, these learners were much more confused by incorrect error codes. Compared to false positives which take a one-step judgement followed by refusal to make any change to the highlighted text, incorrect error codes pose a much greater challenge for learners. This is because of the fact that incorrect error codes at least correctly highlight a grammatical inaccuracy in learners’ essays. This may have misled learners to initially trust Criterion’s error tags rather than to reject it outright as in the case of false positives. However, to effectively evaluate incorrect error flags, students need to process the feedback more deeply to realize their flaws. Some codes correctly identified the errors but the accompanying explanations and/or suggested changes were incorrect (e.g. Criterion’s suggestion of “Consider using the article a” in response to the flagged clause “It is convenient for you to work with old colleague”), while others mislabelled the error type. In response to all of the three incorrect error codes in the multiple-case study, none was
challenged as faulty by the students. Both Nhien and An took these error codes seriously and tried to revise their errors through extensive engagement. This showed their failure to realize the problem of the feedback they received.

All in all, evaluation of feedback seemed to have triggered some form of extensive engagement and evaluation of both the text as well as the error tags. Firstly, within unhelpful feedback, false positives were much more likely to be identified by the learners, while incorrect error codes posed a greater challenge for feedback evaluation. Secondly, individual learners’ perceptions about revision and writing also contributed to the learners’ engagement levels with the error codes received. For some learners, especially those driven by learning goals, Steinhart’s (2001, as cited in Stevenson, 2016) point may hold true when the extra efforts put into feedback evaluation become learning opportunities. Besides the knowledge gained from their reference to external resources, they also get the chance to reflect on their own errors as well as correct structures in their texts, which adds richer input for the revision activities. Other learners, however, may be overwhelmed by the number of error tags to be processed or their linguistic knowledge base is insufficient to evaluate feedback accuracy.

**Use of Criterion ACF for revisions.** With the rare exception of Zhang and Hyland’s (2018) investigation of two learners’ engagement with Pigai automated feedback using a case-study approach, previous research has not investigated learner engagement with automated corrective feedback. Instead, the predominant way to determine students’ use of such feedback was comparing changes in the first and revised drafts of the same essay, which provides an incomplete picture about how students actually used and responded to the feedback.

Therefore, the judgement of students’ use or non-use of automated corrective feedback should be founded on more substantive evidence including both the first/revised drafts as well as qualitative data about their engagement with the feedback. Extending on previous research that found high rates of error codes being disregarded by learners (Attali, 2004; Chapelle et al., 2015; Lavolette et al., 2015, Ranalli et al., 2017), the question that needs to be answered is, “What is the reason for no observed revised form?” Is it simply because students ignored the error tag (i.e., non-use of the automated feedback) or does it follow from extensive engagement that led to resistance to make a change (as in the case of false positives)? Other explanations can be the students’ decision to give up making a change due to their inability to understand the feedback or to find an alternative form, all of which can take place after students have noticed and processed the feedback.

It has been reported in previous research on learners’ revisions using Criterion ACF that error tags from this system had a high underuse rate among the English language learners (e.g., Chapelle
et al., 2015; Lavolette et al., 2015). While it is true that not all of the corrective feedback provided by *Criterion* was followed by revisions in this study, this absence of a revised form is not necessarily evidence of non-use of the feedback. It is also true that the intended use of a feedback point was to encourage the attempt to correct an error. However, the picture gets more complicated in terms of actual feedback use if we take into consideration the false positives generated in student essays, and to which we would prefer to see no revised forms. In this study, feedback use is interpreted in a broader sense which indicates that as long as the provision of the feedback has contributed in some way to the learner’s editing process, this is taken as evidence for its use.

The findings in Chapter 5 indicate that almost a third of all error tags did not lead to a revised form. This included a 7.5% of no-change responses to false positives. The quantitative findings regarding textual operations for revisions also suggested that students were more likely to adopt an avoidance strategy by choosing not to address an error pointed out to them than to venture a revision that may be incorrect. This was demonstrated in generally higher avoidance rates compared to incorrect revision rates across most error types. As an example, the data showed feedback on *missing comma* as a specific area where *Criterion* generated the highest rate of false positives (66%). In response to this, students more often chose to retain their text than take up the system’s suggested alternative, with an incorrect revision rate of about 6.4% as opposed to 14.4% avoidance rate.

The think-aloud protocols recorded for 14 learners indicated a mixture of both well-processed error codes resulting in no revisions as well as cursorily noticed errors which were then skipped. In addition, about three quarters of all error codes led to some revisions. Rather than supporting the reported high underuse rates in previous research, the above findings show some form of engagement with the feedback, either limited or extensive, and the subsequent textual operations conducted as a response to the feedback they received. The findings have contributed to the view that *Criterion* has at least drawn learners’ attention to certain gaps in their interlanguage development, and thus it potentially facilitates L2 acquisition (Heift & Hegelheimer, 2017).

**Revision outcomes: The error type-feedback type interaction.** Whether student engagement with the feedback led to effective revisions depends to a large extent on their sufficient understanding of the errors by reading the highlighted parts in their essays and the metalinguistic explanations from *Criterion*. Quite frequently, students commented that just by looking at the highlighted words or phrases, they immediately realized the errors and proceeded with an alternative form without the need to seek further help elsewhere. The error types being dealt with this way were normally rule-based such as *subject-verb agreement* or *determiner-noun agreement.*
Their remedies were straightforward through simple insertions/deletions of a suffix (for subject-verb agreement or determiner-noun agreement) or substitution of a word with another word (for determiner-noun agreement or spelling errors). On the other hand, feedback type (i.e., generic vs. specific, correct vs. incorrect) also greatly influenced the amount of time students spent implementing the revisions. For example, idiosyncratic errors such as confused words or wrong or missing word often received very specific feedback followed by suggested correct forms. In response to such feedback types, students tended to quickly and successfully revise their texts. Yet, for other errors like garbled sentences or proofread this, generic feedback from Criterion fell short of expectations as little information could be elicited from it to facilitate revisions.

However, not all rule-based errors were easily fixed. Two of the most frequently found errors among the EFL learners in the current study were article and comma usage. While article errors are not highly treatable as they require a high level of acquired grammatical knowledge to be dealt with, comma usage was equally challenging because even native speakers experience comma issues in their writing (Mehrabi-Yazdi, 2018). Therefore, it is reasonable for Criterion to provide specific feedback in response to articles and comma usage, with clearly suggested changes in most metalinguistic explanations for these error types. As this feedback type seems to work for less treatable errors, it is expected that such feedback is generated in response to other challenging categories related to sentence structures such as garbled or run-on sentences as mentioned above.

An intriguing finding from this study is that among the most successfully revised error types were those receiving generic feedback from Criterion. This finding seems to contradict the earlier discussion regarding the need for feedback to be explicit for effective revisions. However, it is also worth stressing that the previous discussion highlights the interaction between different factors, with learner goals, error type, and feedback type being the most significant. It is also notable that because Criterion feedback is mostly superficial on simple errors, extensive engagement is usually not required for correct revisions. Error types which received generic feedback but were highly successfully remedied by students include spelling, subject-verb agreement, determiner-noun agreement, and ill-formed verbs. Apart from spelling errors whose correction is straightforward with reference to a dictionary (as done by all the students participating in the think-aloud protocols), the remaining types are local errors with textual clues being sufficient for making corrections.

On the other hand, least likely to be corrected were fragment and errors that received the code proofread this. Following the point above, students’ frequent failures to revise errors of these two types can be accounted for by the fact that revising these error codes requires textual operations that go beyond the individual word or phrasal level. In addition, these two error types are in fact not specific error categories. Instead, the whole clause is coded and highlighted without specific errors.
being pointed out, which demands much more mental efforts and cognitive processing from learners.

Generally, with 54% of total revisions made based on *Criterion* error tags being correct and an additional 7.5% appropriate response to false positives (i.e., retention of correct forms), the current research shows some discrepancy with previous findings reported by Chapelle et al. (2015) or Lee, Li, and Hegelheimer (2012), both with around 70% successful revision rates using *Criterion ACF*. In appraising evidence for the utilization inference regarding the use of *Criterion* for formative assessment, Chapelle et al. (2015) suggested that with 70% *Criterion*-initiated correct revisions among the learners, this system has positively influenced students’ revision process.

The result in this study is closer to Ranalli et al.’s (2017) 55-60% successful revision rate found for 82 ESL learners in a midwestern university in the US. The authors took a more cautious approach to interpreting this statistic, suggesting that “in the absence of an established standard, it is impossible to say whether this middling finding is an acceptable return on investment of time in using *Criterion*” (p. 22). In evaluating how useful *Criterion* corrective feedback is for their participants, the authors suggested making the distinction between *learning to write* (LW) and *writing to learn* (WL) (Manchón, 2011) as baseline for assessing the level of support or rebuttal of the assumptions underlying the utilization inference. In this distinction, a 55-60% correct revision rate could be considered insufficient support for revising practices among learners if the goal was LW which stresses writing skill development and better written products. On the contrary, the same rate can be acceptable if a writing activity is driven by WL purposes.

In the current study, *Criterion ACF* was used for classroom writing practice with a focus on L2 accuracy and writing skills development among EFL learners. Therefore, a rate of 54% correct revisions and the additional needs to evaluate and respond to 11.7% false positives as well as 6.1% incorrect error codes from *Criterion* do not strongly support a case for the use of *Criterion ACF* as a self-assessment tool among these EFL learners. However, data in the multiple-case study showed that all of the three learners were able to correct their errors based on *Criterion ACF* more than 80% of the time. This high success rate is partially accounted for by these learners’ remarkable ability to detect and respond to false positives and their higher performance compared to the group means using accuracy and fluency measures. Taken together, the group and individual cases’ revision outcomes suggest that the usefulness of *Criterion ACF* for revisions should be interpreted on a case-by-case basis where individual differences are taken into account.
Summary of Validity Evidence for the Utilization Inference

This part of the study has discussed validity evidence related to the Utilization Inference which is justified by the warrant that “Automated corrective feedback on grammar, usage, and mechanics errors obtained from Criterion is useful for students to make decisions on revisions”. Overall, Criterion ACF was found to effectively draw learners’ attention to targeted linguistic forms in their essays with all the feedback being noticed either perfunctorily or substantively. This has led to interestingly differential revision patterns and outcomes among the participating learners. Significant rebuttals come from Criterion’s failure to base its feedback types on pedagogically sound considerations about error types and individual learner factors. Because of this, learners’ engagement with the feedback was not as meaningful as expected, especially with more complex error types. Also challenging the assumption regarding sufficient support for learners to correct the errors from Criterion is a middling revision success rate which can be partially attributed to a rather high proportion of unhelpful feedback. However, there is also ample backing for the assumptions underlying this inference, including the system’s clear metalinguistic explanations, its capacity to trigger noticing of linguistic forms among learners, and facilitation of successful revisions for a wide range of error types. Unhelpful error codes in the form of false positives also added to the support for this inference as these error codes tended to substantively engage learners and encourage them to conduct feedback evaluation, which enriched the input during the revising stage.

Ramification Inference

“Overall, it is a new and cool breeze… This different way of learning to write in English really excites me.” [Focus group interview]

Research Question 4 investigates the changed accuracy among EFL learners after the treatment using Criterion ACF as a supplement to teacher feedback throughout a whole semester, and Research Question 5 is about the learners’ perceptions of this feedback. In addition to this, the multiple-case study provides insights into the interaction between learner factors, feedback quality, and student engagement, and how that interaction impacts learning over time. Data from RQ4, RQ5 and the multiple-case study contribute to the following discussion that appraises evidence for the Ramification Inference in the current research. This discussion also refers to relevant tenets in the cognitive-interactionist perspectives as well as goal orientation and self-regulated learning theories as a way to situate the automated corrective feedback in the broader scholarship on written feedback.
Table 7.4 summarizes the warrant, assumptions, and relevant evidence to support or reject the assumptions for Ramification Inference. Summarized findings serve as the backing or rebuttal for the assumptions under investigation.
### Table 7.5

<table>
<thead>
<tr>
<th>Warrant that justifies the inference</th>
<th>Relevant assumptions underlying the warrant</th>
<th>Backing/Rebuttal for the assumptions (Summarized findings)</th>
</tr>
</thead>
</table>
| Use of *Criterion ACF* is beneficial for learning, including: | 1. Students acquire linguistic features and increase awareness of linguistic forms. | 1. Changes in accuracy from pre to post and delayed post-tests:  
  - **Experimental group**: There was neither significant treatment nor retention effect for the incorporation of automated corrective feedback during one semester on learners’ overall accuracy and all of the five morphemes under study.  
  - **Comparison group**: Despite lack of access to *Criterion ACF*, significant accuracy gains were found for third person singular, plural use, and overall accuracy. For third-person singular and overall accuracy measures, the experimental group had a significantly higher score at pre-test. |
| 2. Students perceive the feedback positively. | 2. Despite ambivalent feelings, students generally expressed positive perceptions of *Criterion ACF* for its systematic and well-organized feedback; simple and straightforward metalinguistic explanations; and instant feedback generation.  
  - Students sometimes challenged the accuracy and authority of *Criterion ACF*.  
  - Students tended to take the middle grounds in comparing teacher CF and *Criterion ACF*. | |
| 3. Frequent access to the feedback maximizes the learning opportunities. | 3. Use of feedback as a learning opportunity varied across learners and was dependent on learner goals, cognitive processing load, error type, and feedback quality.  
  - Processing incorrect feedback and false positives triggered substantive noticing of forms. | |
4. The use of *Criterion ACF* facilitates self-regulatory revision strategies.

3-4. Using *Criterion ACF* gradually raised students’ awareness about the errors they frequently made. Depending on learner goals, different learners tended to vary in how much they made use of the feedback through engagement and processing of the feedback. Nhien deployed a wider range of cognitive (e.g. strategies to understand and remember information such as looking up online resources) and metacognitive strategies (e.g. evaluating information and their revisions). Instead, An and Trang were more dependent on their stored metalinguistic knowledge for the major source of reference.

- Instant feedback generation allowed for more effective revision and writing skill enhancement
**Increased grammar awareness and mastery of linguistic features.** To evaluate the validity evidence for the first assumption in the ramification inference, this part of the discussion focuses on how learners’ mastery of linguistic forms as well as general grammatical awareness have or have not improved as a result of using *Criterion* automated feedback.

**Is increased grammar awareness conducive to mastery of linguistic features?** Findings from the interviews indicate that learners became more aware of the errors they made most frequently thanks to *Criterion*’s corrective feedback. Along with this, *Criterion* helped raise their consciousness about grammatical accuracy when composing and revising their essays. Judged by these observations, *Criterion* automated feedback seems to have gradually instilled in learners a more careful approach to selecting appropriate linguistic forms to express their ideas. These results reinforce the widely held consensus about “the capacity of AWE to raise meta-awareness” (Stevenson, 2016, p. 8).

Despite this, there was a mismatch between learners’ perceptions of their increased grammatical awareness thanks to *Criterion* ACF with their actual accuracy gains after using it for a whole semester, a finding that reiterates results from Koizumi, In’nami, Asano and Agawa’s (2016) study on *Criterion* use among Japanese EFL tertiary learners. Results from the pre-post-delayed post tests showed no intervention or retention effects for the use of *Criterion* ACF on learners’ accuracy over time. Interestingly, the comparison group, with overall accuracy at pre-test being significantly lower than that of the experimental group, demonstrated significant overall accuracy improvement after a semester. This is also true of the accuracy measure for third person singular where the experimental group scored statistically higher than the comparison group at pre-test but only the comparison group students gained in third person singular use over time.

The result that experimental learners in the current research did not gain in accuracy in any of the studied measures finds partial explanation in the middling revision success rates discussed earlier. In addition, with these learners’ cautious and not so high uptake as evidenced in their greater likelihood to keep texts unchanged rather than venture a revision they were not confident about, it is reasonable to believe that students did not seem to fully benefit from the corrective feedback in the first place, and therefore, they did not internalize the relevant linguistic forms after a five-month period.

**The ceiling effect hypothesis.** Additional explanation for learners’ failure to acquire the linguistic forms targeted by *Criterion* is based on the multiple-case study insights and the comparison of the two groups’ accuracy changes in the two measures of overall accuracy and third
person singular where the experimental group had significantly higher pre-test scores than their comparison group counterparts. This high baseline (at 0.81 and 0.8 for overall accuracy and third person singular use, respectively) may have left little room for improvement among the experimental students. In other words, the incorporation of *Criterion* ACF as part of the writing course was unlikely to have any significant effect on accuracy gains on the intervention group due to the “ceiling effect”. On the contrary, starting at a low 0.69 for overall accuracy and 0.50 for third person singular use, the comparison group had much larger space to improve their accuracy. This ceiling effect hypothesis is further substantiated by the multiple-case study results. Specifically, Nhien and An had very high pre-test scores (0.95 and 0.92, respectively). Despite having a mixture of substantive and perfunctory engagement episodes with all of the error codes received, both of these learners experienced decreased accuracy over the studied period. Conversely, Trang, starting at a baseline slightly below the group’s mean (0.79), gradually gained in general accuracy from the pre-test to post and delayed post-tests (0.83 and 0.87, respectively).

These observations suggest that learners’ level of proficiency plays a critical role in determining whether they would benefit from the corrective feedback provided by *Criterion*. More specifically, lower proficiency students are likely to acquire more linguistic accuracy over time with access to and use of *Criterion* ACF. Conversely, the automated feedback seems to have little impact on improving higher proficiency learners’ accuracy. This explanation is consistent with the earlier discussion about the nature of *Criterion* ACF showing that it targets simple and surface-level errors which low proficiency learners are likely to make frequently. Proficient learners have a larger linguistic repertoire and tend to make more complex errors that *Criterion* fails to address. Therefore, little meaningful and essential feedback is received by these learners to help them gain in accuracy by using *Criterion* over a period of time.

**Counterarguments to the ceiling effect hypothesis.** Some will argue that within a five-month period, it would be hard for *Criterion* ACF to make a big difference to learners’ linguistic accuracy, irrespective of their proficiency levels. Using Trang’s exclusively perfunctory noticing of the feedback and superficial processing of the errors *Criterion* pointed out to her, it can be argued that such an engagement pattern does not substantiate the gain in accuracy this learner experienced. Interactionist approaches in second language acquisition state that “acquired may mean a number of gradual and non-linear changes in the linguistic (and, in some theories, metalinguistic) behaviour that characterize the developmental course of L2 acquisition, based on construct interpretations such as: (i) a form has ‘emerged,’ has been ‘detected,’ ‘noticed,’ ‘attempted,’ or ‘restructured’; (ii) a learner is ‘aware’ of a form or a form-related pattern” (Norris & Ortega, 2003, p. 727). Viewed
from interactionist perspectives, Trang’s depth of feedback processing does not suggest that this learner sufficiently noticed and understood targeted forms to guarantee linguistic internalization and long-term acquisition. Additionally, it is evident from Trang’s writing fluency and complexity measures across the three time points as well as competent self-initiated revisions that she was not less proficient than the other two cases as it appeared from her pre-test accuracy scores. The ceiling effect hypothesis which places critical weight on learner proficiency level does not seem to hold true for Trang’s gradual accuracy gains, although it can partially account for Nhien and An’s decreased accuracy after a five-month period.

Therefore, we need an alternative explanation for Trang’s low recorded accuracy score at pre-test and for her increased accuracy trajectory over the studied period. Judged by the superficial nature of the errors being tagged in her essay, it is highly probable that this learner made several slips rather than errors coming from gaps in her linguistic knowledge in the pre-test essay. After three practice sessions on Criterion, the automated corrective feedback drew her attention to these easily remedied issues and also raised her awareness about some frequent errors. Thanks to this, Trang was able to be more careful in later compositions and she was also able to gain in accuracy scores as a result. This interpretation reflects a common idea expressed by the focus group students who attributed their more careful spelling and use of punctuations in practice essays over time due to Criterion’s error tags as well as increased consciousness about grammar when they composed their essays.

**Positive perceptions aboutCriterion ACF.** Adding to previous literature on learners’ perception of automated feedback, findings in the current study show that having access to Criterion ACF was considered a novel learning experience among students. Interviewed participants were motivated by the additional source of feedback from Criterion to complement their writing instructors’ feedback, a similar result to findings from other classroom-based studies by Dikli and Bleyle (2014) or Chen and Cheng (2006). Such positive perceptions of Criterion ACF paved the way for more willingness to engage with and process the given feedback, which potentially facilitates error correction (Li, Link, & Hegelheimer, 2015). In addition, learners also expressed positive feelings about Criterion’s systematic feedback generation approach as well as its satisfactory precision for form-focused feedback, which serves as prerequisite for greater mastery of linguistic forms through regular and timely exposure to this additional source of input.

However, interviewed students occasionally challenged the authority of Criterion feedback and questioned its accuracy for certain errors they made. These perceptions reflect students’ awareness about the limitations of the automated corrective feedback Criterion provided, especially
the system’s false positives and incorrect error codes. This may explain the rate of justified and unjustified no-change responses to noticed errors among the learners. Students’ awareness about Criterion’s shortcomings was also manifest in their comparison of teacher and automated feedback. They frequently compared the advantages and disadvantages of either source before concluding that Criterion ACF can be a valuable supplement for the usually delayed instructor feedback, especially with the former’s focus on surface level features of writing.

Generally, while overall positive perceptions and a satisfactory level of trust in Criterion ACF can be seen as supporting evidence, the fact that a number of learners still exercised caution when using the feedback for revisions may be taken as rebuttal evidence for the first assumption underlying the ramification inference.

Maximized learning opportunities. The use of Criterion automated corrective feedback was found to help promote noticing of gaps in learners’ interlanguage development, with all the error tags being noticed among the 14 TAP learners. Although noticing is not considered a guarantee of acquisition, the interactionist approaches hypothesize that “attention is a major component”, and that “some type of focus on form is necessary for certain types of learning” (Polio, 2012, p. 383). This theoretical assumption matches three benefits perceived by learners in the current research: (i) Criterion’s systematic corrective feedback helped with increasing awareness about the errors each learner frequently made; (ii) students could consolidate knowledge about certain linguistic forms after responding to Criterion’s formulaic but consistent feedback; and (iii) the instant and online feedback facilitated more effective revisions.

On elaborating their positive feelings about Criterion ACF, many participants perceived the system as a gatekeeper and proof-reader who provided almost immediate response to their writing. From an interactionist perspective, one of the key features for feedback to be effective is its immediacy (Polio, 2012) as it helps with more effective reflection on one’s own writing. Added to this, with the opportunities to revise as much as possible through multiple drafting features, Criterion and its automated corrective feedback create a space for regular writing practice and feedback provision.

Further insights from the multiple-case study show that learners demonstrated variability in their engagement and use of the corrective feedback, which was probably the reason why they benefited from the feedback at differential levels. For example, Nhien pursued learning goals, and thus found Criterion ACF a highly useful source for developing cognitive and metacognitive strategies in dealing with a range of error codes. For such learners, the highlighting of errors was just a starting point when they noticed a gap in their output. It was probably this noticing that
kicked off a chain of metacognitive strategies from planning, monitoring, and evaluating their revisions.

On the other hand, learners driven by performance goals like Trang and to some extent An did not step up their efforts at engaging with the feedback to extend the learning opportunities the feedback provided. They were more focused on the immediate task of correcting the error pointed out through quick rejection of feedback or adoption of what was given. Beside Criterion ACF, their main resource in confirming a revision was acquired metalinguistic knowledge, which was made explicit in An’s think-aloud protocols but remained implicit in Trang’s case until she had a chance to clarify her error correction approach in the stimulated recall.

What can be drawn from the three selected cases is that whether learners can benefit substantially from the engagement with an automated error tag much depends very much on their personal approaches to revisions and perceptions of learning as a whole. This includes both their learning goals and subsequent strategy choices to conduct revisions in response to the automated feedback. With that said, there is no conclusive evidence regarding support or rebuttal of the assumption under investigation, but human intervention in the form of instructors’ orientation and strategy training may tremendously help learners benefit most from the process of engaging with and responding to error tags from AWE systems like Criterion.

**Use of Criterion ACF facilitates self-regulatory revision strategies.** This part of the discussion focuses on how the use of Criterion ACF facilitates self-regulatory revision strategies. Regarding revisions initiated by Criterion ACF, the multiple-case study findings suggest that depending on learner goals, different students varied in their revision strategies. For example, Nhien deployed a wider range of cognitive (e.g. strategies to understand and remember information such as looking up online resources) and metacognitive strategies (e.g. evaluating the input from reference sources and her revisions). Instead, An and Trang were more dependent on their stored metalinguistic knowledge for the major source of reference. A point worth noting is that while the observation of Trang’s think-aloud protocols suggests this learner’s limited engagement with the feedback, her stimulated recall interview reveals that her extensive engagement was implicit rather than non-existent. This is because Trang actually based her revisions on reference to previously learned knowledge and she was confident enough about the error that no further action was needed for a correct revision.

It is also worth noting that all of the selected cases were among the higher performing group of learners, judged by their accuracy scores, essay length, and complexity measures. As earlier discussed, Criterion corrective feedback was a mixture of generic and specific metalinguistic
explanations. However, overall elaboration on errors was perceived insufficient, which was made worse with the absence of chances to follow up on errors. Where the feedback could be very specific with messages like “You may need to use an article before this word. Consider using the article a”, learners may adopt Criterion’s suggestion but still wonder why they should do that. Certain students, especially lower proficiency ones, may get too little assistance for areas of language in which they are still struggling, because “if learners have insufficient knowledge in store, they may not notice-with-understanding the gap between their output and that which is being provided in the feedback” (Bitchener, 2017, p. 138).

More strategic learners, however, could make use of the error tags and Criterion revision platforms to enhance self-regulation, as in-built features on this system and similar AWE tools were found to generally promote learner autonomy (Stevenson, 2016). In the current study, learners referred to websites for example sentences containing certain words or looked up grammatical rules related to an error being processed. These learners could be seen as exemplary for the positive effects of Criterion ACF on the development of self-regulatory revision strategies. In such cases, the corrective feedback codes generated to a range of errors in learner essays serve as triggers for cognitive processes of noticing and reflection on one’s own writing. Extensive engagement with feedback may follow and is likely to draw learners’ attention to linguistic features. This helps writers not only build up their metalinguistic repertoire but also refine strategy use and improve their self-editing skills.

Criterion’s potential benefit of instilling self-regulated revision strategies is most salient with unhelpful error codes. Findings from this research have shown that incorrect feedback was more likely to trigger substantive noticing, and thus extensive engagement strategies, than correct feedback. The three selected cases in this research were characterized as being able to successfully conduct feedback evaluation. In particular, false positives were mostly successfully detected. Through feedback evaluation, students also reflected on their stored knowledge or use strategies such as reference to external resources for confirming doubts about the codes. These were especially great opportunities for advanced learners whose metalinguistic knowledge was sufficient to refine their strategy use as well as self-editing skills (Polio, 2012).

In the two cases of Trang and Nhien, Criterion ACF was noticed and responded to, but this was only part of the larger picture of these two learners’ revision practices. These learners conducted their self-initiated revisions after finishing with Criterion error tags. One of the possible explanations for the extensive self-revisions made by Trang is that the automated revision interface on Criterion itself with the split screens allows for more substantive changes relating to content and organization of the essay. Compared to traditional paper-based compositions, typing on Criterion
makes it much easier to edit texts by using the cut, copy, and paste commands as well as rearrange large portions of texts. Even in the case of Nhien who made explicit in her stimulated recall that she had the routine of proofreading and refining linguistic devices after she finished writing the essay, her habitual self-initiated revisions were made easier with access to Criterion ACF and its revision interface.

**Summary of Validity Evidence for the Ramification Inference**

This part of the study has discussed validity evidence related to the Ramification Inference which is justified by the warrant that “Use of Criterion ACF is beneficial for learning”. Significant supporting evidence comes from students’ positive perceptions about the corrective feedback they received from Criterion, its facilitation of effective self-regulatory revision strategies thanks to the instant, systematic, and online revision interface, and the promotion of grammatical awareness about one’s frequent error types that need improving. The discussion also stressed individual variability in benefiting from Criterion ACF by looking at the critical roles of learner proficiency and learning goals. On the other hand, main rebuttal evidence includes students’ occasional distrust of the feedback and their lack of gains in overall as well as specific accuracy measures after using Criterion ACF for a whole semester. The discussion also proposed explanatory hypotheses for differential accuracy change trajectories observed for the three selected cases, which provided nuanced interpretations of the validity evidence for the four assumptions in the ramification inference. Zhang and Hyland (2018) made a comment that could summarize the results related to the Ramification Inference in this research. They said that the provision of feedback from Criterion itself does not guarantee writing improvement, either in the short term (as uptake) or long term (acquisition). Instead, the benefits from automated corrective feedback may be more about promoting noticing, encouraging independent editing (Lavolette, Polio, & Kahng, 2015), and training students in using cognitive and metacognitive strategies to improve writing skills.
Chapter 8
CONCLUSIONS

Restatement of Research Aims and Key Findings

The current research set out to seek insights into the impact of automated corrective feedback on EFL students’ learning to write in English. In order to provide detailed answers to this overall aim, five research questions about the nature and accuracy of Criterion ACF (Questions 1 & 2); students’ engagement and subsequent revisions (Question 3); their observed changed accuracy in writing (Question 4); and their perceptions of the feedback (Question 5) were answered. Research question 3 was extended in the multiple-case study where individual learners’ differences were focused on.

The impact of Criterion ACF on EFL learners’ L2 writing practices and outcomes was evaluated based on the findings for all of the research questions above, which were mapped on to a validity argument that examines the four inferences of evaluation, explanation, utilization, and ramification. The examination of the evidence for the warrants underlying evaluation and explanation inferences which focus on the accuracy, nature, and relevance of Criterion ACF provides more support than rebuttal. That is, Criterion was able to satisfactorily address the needs of EFL learners for surface-level error types, while it still lacked coverage of some major issues in the students’ L2 writing.

Turning to the third inference of utilization which emphasizes the usefulness of Criterion ACF as demonstrated in students’ uptake of the feedback, their cognitive and behavioral engagement with Criterion ACF was analyzed, which provided a mixture of support and rebuttal for the warrant underlying this inference. Firstly, Criterion can be praised for its effective facilitation of revising strategies as well as for successfully triggering noticing among learners, even with its false positives and incorrect error codes. Such a finding challenged earlier research (e.g., Chapelle et al., 2015; Ranalli et al., 2017)) citing high non-use rates of automated feedback which based their results on the comparison of first and revised drafts of the same compositions. Despite this, Criterion resulted in only a modest revision success rate among the participating students. Also, the lack of a pedagogically based approach to feedback generation in Criterion has detrimentally impacted student meaningful engagement with the feedback, which may partially account for the modest revision outcomes.

Examination of the ramification inference enriched our understanding about the influence of Criterion ACF on the EFL learners in the current study. This inference stresses the benefits of the automated feedback for students’ longer-term learning outcomes, as demonstrated in their
acquisition of linguistic features and the development of cognitive and metacognitive strategies in writing and revisions. Generally, backing for the warrant of this inference comes from students’ positive perceptions of the feedback they received from Criterion as well as the system’s capacity to facilitate self-regulatory revision and learning strategies. However, the warrant is undermined by the major finding that no significant intervention or retention effects of the use of Criterion ACF on the learners’ accuracy gains were recorded for the studied period. The examination of this inference also highlights the critical roles of individual learner differences including proficiency and learning goals as mediating factors.

Some conclusions can be drawn from the findings to answer the research aims set out at the beginning regarding the impacts of Criterion ACF on EFL learners’ L2 writing. Firstly, the automated corrective feedback generated by Criterion can partially address the need for instant and round-the-clock feedback among EFL learners who have limited access to other feedback sources when learning to write, and when teacher feedback is usually delayed or is not very frequently provided in large writing classes. Secondly, students can benefit from the chance to extend their learning using the automated corrective error codes Criterion provides. The metalinguistic explanations accompanying each error code can serve as a starting point for further references to consolidate previously acquired knowledge or gain new information. Learners can also learn to develop self-regulatory writing and revising strategies by using the feedback on a regular basis.

Although Criterion ACF still suffers from some major shortcomings as the findings in the current research have shown, different learners can still benefit from such feedback if writing instructors, AWE developers, and other stakeholders collaborate. The next section will elaborate on some implications for different stakeholders so that the shortcomings of the current version of Criterion can be overcome or at least downplayed.

Implications

Guided by the findings in the main research as well as the embedded multiple-case study, theoretical implications and practical recommendations are presented below. The practical implications are subdivided into those applicable to writing instructors and others expected of Criterion ACF developers.

**Theoretical implications.** The current study has added a missing piece in the research on EFL learners’ use of Criterion ACF, which further completes the picture regarding the value of AWE corrective feedback for these learners. The inclusion of qualitative data from think-aloud protocols and stimulated recalls unveils the underlying mechanisms that moderate learners’ final
revising decisions as observed through the comparison of their first and revised drafts. The qualitative data have somewhat invalidated certain findings in earlier research assuming that the absence of any change to a flagged error indicates the learner’s non-use of the feedback (e.g., Chapelle et al., 2015; Ranalli et al., 2017) or that the feedback was ignored (Lavolette et al., 2015). In actual fact, learners’ think-aloud protocols and retrospective interviews point to a few other possibilities that resulted in no observed revisions in their texts, and most of these possibilities did mean some engagement with the feedback, ranging from perfunctory to highly extensive engagement episodes.

In addition, this research was able to shed light on the potentials of Criterion ACF which merit further investigation. Earlier we discussed the interaction between different learner factors, internal as well as external, and how that interaction contributes to the subsequent engagement and revision activities. Criterion ACF may lack key features of effective feedback expected from a socio-cultural perspective. However, given the fact that writing online entails access to various reference resources, the extent to which ELLs can make use of the feedback was dependent on individual learners’ learning strategies, cognitive capacities and affective factors. It is the reference to these external resources to support the learning process itself that makes Criterion ACF beneficial for revisions if learners are able to put them to good use.

It would be reasonable to say that how learners conducted revisions was case specific, being dependent on both the learner internal factors as well as external factors such as feedback type and error categories. For instance, Nhi, the self-motivated student driven by learning goals in the current study, provided a clear example of not only using Criterion ACF as a means to reflect on her grammatical and lexical choices but also seeing it as a trigger for self-initiated revisions at global levels. On the other hand, Anh chose to superficially engage with the feedback and thus revisions were quickly done. Given the complexities related to students’ engagement with and use of automated corrective feedback, it is worth investigating the feedback from multiple theoretical perspectives as is the case with written corrective feedback. However, the analyses in the current study suggest that most relevant theoretical tenets for such an investigation come from the cognitive-interactionist perspectives. The operationalization of one of the central cognitive-interactionist constructs, noticing, together with the different levels of awareness in Schmidt’s (1990, 1993) noticing hypothesis, has enabled systematic examination of students’ engagement with the automated corrective feedback in the current study. Through doing this, this study has managed to apply an SLA construct to a novel context related to the use of automated corrective feedback in an EFL writing classroom.
Pedagogical implications. One of the main conclusions mentioned earlier is that the automated corrective feedback from Criterion and similar AWE programs can serve the dual functions of a gatekeeper for trivial/local errors as well as an initiator of more global textual revisions. Different learners benefit differently from the automated corrective feedback, depending on relevant personal traits such as motivation, learning styles as well as their proficiency levels. Therefore, what writing instructors can do is making sound pedagogical decisions that help learners make the best of the available technological capacities offered by AWE systems like Criterion. While still waiting for improved functions in Criterion which address major concerns about its suitability when put to use in an EFL formative assessment context (e.g., the absence of feedback on some critical linguistic issues for English language learners), writing instructors and relevant educational bodies can think of ways to minimize the disadvantages while maximizing the benefits of this automated feedback tool. For example, Criterion ACF can be incorporated alongside teacher feedback to alleviate teacher workload for mechanical and surface level errors or it can also work well as a self-study tool for extra practice and homework assignments, especially among lower proficiency level students.

Specifically, the multiple-case study reveals that there were three factors leading to failed revision attempts, including insufficient guidance from the feedback, students’ gap in linguistic and metalinguistic knowledge, and their inability to retrieve information from previously acquired knowledge. Writing instructors can help address these issues in their writing classes. For example, oral feedback sessions can be provided during meeting hours so that learners can bring up clarification questions after they have engaged with Criterion ACF. The gatekeeping function played by automated corrective feedback can only be fully realized if learners are able to process the feedback they receive properly. This means that they could understand the explanatory language used in the feedback well and are able to evaluate the correctness of these error codes, considering the fact that Criterion as well as other programs still sometimes generate false positives and incorrect error codes. If taken at face value, something that lower-proficiency learners are more inclined to do, unhelpful feedback is sure to do more harm than good to the development of second language learners’ writing accuracy.

Furthermore, strategy training sessions can be embedded in meeting hours during which learners learn cognitive/metacognitive strategies for revisions or share reference sources they find most helpful to seek further information for the error codes received. For different in-class sessions, the instructor can also discuss what error types to attend to with the learners so that they can properly focus attention on the most relevant issues and reduce the cognitive load. This solution helps to address a current problem created by the comprehensive feedback approach in Criterion
and other automated writing evaluation programs. Cognitive overload is of special concern not only for lower-proficiency learners who normally get overwhelmed by the number of error tags generated in their essays but it is also true for higher proficiency students in the multiple-case study, two of whom were mostly superficially engaged with the feedback as they tended to focus on the number of errors they worked on rather than the quality of their processing or the amount of meaningful attention paid to each error code to ensure correct outcomes.

Besides making explicit the ways to approach the feedback or the error types to focus on, the instructor should also be open about the current shortcomings of Criterion ACF. By sharing his/her opinions about what Criterion ACF is currently not able to do well, teachers leave students to their own devices about what to do with the feedback. For instance, a large number of students in the current research expressed their doubts about certain error codes they received. While some of them were confident about the false positives they encountered, a few others were wondering if their evaluation of the feedback was correct. The writing instructor can help remove such self-doubts and help students move forward with appropriate strategies to double check the errors as well as the feedback they receive.

The mostly hedging nature of Criterion error messages raises another issue regarding ACF usefulness, especially for low-proficiency learners who may need definitive answers to their errors. Elsewhere the system has been suggested to work best for novice writers who have limited linguistic knowledge (e.g., Ranalli et al., 2017) as Criterion and other similar automated feedback tools are good at detecting surface level errors than more complex errors and higher order aspects of writing. While the current research tends to support this suggestion due to the simple nature of the errors detected by Criterion, a lot can still be done to boost the value of Criterion ACF for learners. One of the first steps is to raise students’ awareness about the nature of Criterion ACF and train them in using suitable strategies to deal with different types of tagged errors. With appropriate support from the writing instructors, lower-proficiency learners can overcome initial challenges in exercising feedback evaluation and gradually develop a sense of how to deal with certain error tags depending on the error types.

**Implications for Criterion developers.** Criterion benefited learners of different proficiency levels in different ways. This points to the need for a more flexible and adaptable feedback generation approach, as suggested in previous literature. Being adaptable means certain feedback functions can be turned on or off to adjust to learners’ proficiency and other individual traits or preferences and stages of the learning process. Although this is not a new recommendation, it reiterates previous researchers’ call for a more selective approach to feedback (Mehrabi-Zahdi,
and the ability to toggle flagging for particular error types (Ranalli, 2018). Low proficiency learners are likely to receive meaningful support to address surface level errors detected by Criterion. However, they tend to fail to learn from Criterion ACF for more complex errors. On the other hand, for potentially longer-term impact on learning and acquisition, one of the findings in Chapter 5 suggests that implicit feedback can bring about more learning opportunities. This is especially true for high proficiency learners who can effectively engage with the feedback and seek out further references to process the errors. Such a process presents the students with more opportunities to consolidate their knowledge and learning strategies. Therefore, making Criterion feedback more flexible in terms of toggling between its error types or feedback options may help to maximize the benefits for a larger number of learners.

The findings also highlight the absence of graduated feedback generated by Criterion, which detrimentally impacted not only students’ meaningful engagement with it but also their long-term acquisition of targeted structures in English. One way to partially overcome this is inputting optional levels of assistance in the system (i.e., explicit or implicit feedback options) for learners to choose from. Otherwise, technical capacities can be developed to allow for systematic treatment of errors during first and later submissions for the same essay. Specifically, more implicit feedback is provided during the first submission while explicit feedback can be delayed for second submission onwards.

Another aspect that calls for more flexibility in Criterion mechanisms is its feedback on certain mechanics issues in student writing. For example, the large number of Criterion redundant error tags in comma usage found in the current research points to the system’s insensitivity to the dynamic nature and changing realities of writing. This is an aspect that Criterion developers can work on to help second language learners benefit more from its automated corrective feedback and to reduce the number of false positives generated. While the technical challenge of incorporating meaningful algorithms to detect higher textual level error types or providing content-specific feedback in Criterion is much harder to achieve at the current stage, the optional turning on/off of certain functions and more flexible treatment of writing mechanics issues seems plausible.

All things considered, end-users including both learners and writing instructors still need to wait for new technical capacities that allow Criterion error detection mechanisms to work beyond the textual level of two-word combinations or certain fixed phrases to address higher syntactic structures at phrasal, clausal, and sentential levels. This seems to be the only way to avoid a large number of both false positives and incorrect error codes (e.g., missing/extra article errors) currently found in Criterion ACF. What is worth special mention is the need for feedback on verb tenses and aspects in English which remain one of the central concerns for SLA research (Fuchs & Werner,
2018) and a salient feature that influences “the meaning-making potential of second language writers” (Dikli & Bleyle, 2014, p. 9). By finding ways to address these gaps in its error detection mechanisms, Criterion can better fulfil its claimed functions as an automated instructional and assessment tool in English as a second or foreign language contexts.

**Limitations of the Current Research**

Despite the best efforts to conduct the current study, this research suffers from some limitations that may threaten the validity of certain conclusions and recommendations it has forwarded. This section goes into details about these limitations before suggesting recommendations that can address them in future research.

Firstly, the focus of the study is on linguistic accuracy as demonstrated in students’ use of the corrective feedback generated by Criterion. Therefore, the system’s feedback on organisation, style, and development was suppressed. The decision to turn off these feedback functions was informed by previous research which produced unfavourable views about content, organisation, style, and development feedback from Criterion (e.g., Dikli, 2006; Li et al., 2015; Warschauer & Grimes, 2008) but learners’ engagement with and response to the automated feedback may be biased due to this study design. We can argue against this decision by raising the point that student engagement and revision practices may have been at more global levels if access to other feedback aspects had been allowed.

In the current study, other aspects related to writing performance including linguistic complexity or fluency have not been thoroughly investigated alongside accuracy. Therefore, any observed changes recorded for the participating students may not have fully captured their developmental trajectories over a five-month period. For instance, we may have missed important information about the improved complexity in a learner’s essay, which might have been offset by decreased accuracy as the learner experimented with more complicated syntactic structures.

Added to this, although designed to be longitudinal, the current research had three official practice sessions during which students composed and revised essays on Criterion, suggesting only three snapshots of their performance using the automated corrective feedback. Given the fact that Criterion ACF is unfocused in nature, the number of error tags each individual learner received from the system may be insufficient to result in any observed gains in global accuracy or the accuracy of the targeted linguistic features. In other words, this small number of feedback sessions is probably not enough to fully capture the effects of Criterion ACF, and thus findings related to accuracy changes based on the pre-post test design can be questioned to some extent.
Another factor that may have impacted the reported results is the lack of formal evaluation of how proficient students were in using *Criterion*, a novel learning platform for the writing skills in their context. The researcher made great efforts in offering help with the use of *Criterion* throughout the studied period, especially during the early stages when learners needed time to familiarize themselves with this learning tool. This not only partially addressed participants’ needs but also gave an idea of how well they could navigate the system during practice sessions.

In terms of methodology, there was a significant mismatch between Trang’s think-aloud protocol verbalizations and her expressed ideas during the stimulated recall. Specifically, while her TAP tended to portray her cursory treatment of the feedback and errors, her stimulated recall suggested deep yet implicit engagement with the feedback. This reveals a shortcoming in the data collection tool of think-aloud protocol where the value of the actual data was partially compromised by personal traits such as the fact that some people are more expressive than others. The triangulation of data through the use of stimulated recall interviews, however, has helped address this shortcoming to some extent.

The next limitation relates to the lack of randomized sampling for this research: The groups (experimental and comparison) were intact writing classes. Although it is quite common practice in the applied linguistics field to use available classes which provide the naturalistic setting for research, this convenience sampling approach resulted in lack of randomization. Because of this, participants in the two groups were not equally proficient in two out of the total six measures at the outset, causing some challenges for data analyses and data interpretations later on. The researcher has managed to find appropriate ways to deal with this unequal group levels through the choice of suitable statistical procedures, and caution was exercised when the results were interpreted.

About the multiple-case study, with a rather limited number of cases to choose from, the study selected three students, two of whom started at above 90% accuracy in the overall measure. The remaining case had an 81% accuracy rate at pre-test. Therefore, the three selected cases differed at pre-test. In addition, judged by the six accuracy measures as well as their complexity and fluency statistics used in the current study, all of the selected cases were also very proficient prior to the intervention. These two issues have made comparisons across cases and interpretations of their changed accuracy over the studied period a bit difficult. Better controls over proficiency levels and other individual criteria can be exercised in future research to enhance the validity of any arguments or conclusions.

The above limitations lead us on to some possible future research directions.
**Recommendations for Future Research**

Considering the critical impact of the accuracy of ACF on learners’ use and perceptions, future research can seek to answer the question about how accurate the automated corrective feedback from *Criterion* is using both statistics of *precision* and *recall*. This would provide a more complete picture of the value of the automated corrective feedback, making it easier to compare different AWE systems in case teachers or administrators need to make a choice between different AWE programs. It is, however, also worth mentioning the challenge of operationalizing *recall* and of coding errors in learner essays to extract this statistic. An example to refer to is Lavolette et al (2015) study to see how *recall* was operationalized.

Similar to written corrective feedback research, future studies on AWE feedback can also trial different modes of automated feedback provision on specific learner populations to provide empirical evidence that informs pedagogical implementations of AWE systems. For example, intermediate students can be placed in different experimental groups. Each group can have total access to *Criterion* feedback (including *grammar*, *usage*, *mechanics*, *style*, *organisation* and *development*), only form-focused feedback (*grammar*, *usage*, and *mechanics*), or only one of the targeted feedback aspects mentioned above. Comparisons of the effects of these conditions can reveal more nuanced information about different ways to implement the use of automated feedback for targeted groups of learners.

Students’ linguistic accuracy can be studied over longer periods with more practice sessions to reveal richer evidence for any recorded changes. When more ample data is provided through multiple feedback sessions, not only global accuracy but also specific measures related to targeted features can be investigated more effectively. It would be interesting and useful to examine the effectiveness of automated corrective feedback, either focused or unfocused, based on this expansion of the database. In addition, students’ writing development due to the use of automated feedback can be studied in greater depths from multiple perspectives with the addition of other measures besides accuracy. These may include syntactic and lexical complexity, fluency, or idea development in essays. The interrelationship between different measures of writing development (e.g., accuracy, complexity, and fluency) is worth deeper investigation to more accurately capture the changes in students’ writing performance.

Finally, insights about individual differences need more investigation when researching engagement with automated feedback. This can be achieved by using a case-study approach or a carefully planned experimental design comparing learners of differential proficiency levels and revision success rates. This would provide a more complete picture of learners’ engagement and response to ACF to zoom in on its impacts and suitability for different targeted learner groups. In
order to do this, while randomization of sampling is normally hard to achieve, future research can
learn from previous studies to use a stricter set of criteria when selecting participants or grouping
samples. This aims to control for confounding factors that tend to emerge later on in the research
process to ensure the validity of arguments and conclusions.

In conclusion, the current research has managed to enrich understandings of EFL learners’
engagement, use, and perceptions of the automated corrective feedback provided by one of the most
popular AWE systems, Criterion. The empirical evidence based on both quantitative and qualitative
data analyses has confirmed previous findings in the literature regarding students’ perceptions, and
to some extent, engagement with the feedback. At the same time, this study provides new insights
regarding EFL learners’ cognitive and behavioural engagement with Criterion ACF through
revealing deeper layers of the engagement episodes. This has helped to fill some gaps in the
literature about the use of automated corrective feedback as part of classroom feedback provision
practices. Hopefully, these findings can contribute to more sound pedagogical implementations of
these new tools in the second language writing classrooms.
REFERENCES

Activity 4: Complexity in oral vs. written language. (n.d.). Retrieved from https://carla.umn.edu/learnerlanguage/spn/comp/activity4.html?fbclid=IwAR1LSZQAsiNWCVo2YgO8hShMcDH1vw-hrpzwaaxsXcnrj-W4kHgtHvX7MZk


262


Appendix A
Pre-Post Test Papers

A. Sample test paper

Student’s data
Full name: .................................................................
Gender: (Please circle)  Male  Female
Age: .................................................................
Years of learning English: ........................................
Have you been to an English-speaking country? (Please circle)  Yes  No
If Yes, please provide details:  Which country/countries?......................................................
For how long?........................................................................................................
Self-rating of computer typing skill (Please circle the number that applies to you):

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
</tr>
</tbody>
</table>

Writing Test Prompt
Write an essay of about 250 to 300 words in response to the following question. You have 45 minutes to complete this task.

Resources Disappearing
Many parts of the world are losing important natural resources, such as forests, animals or clean water. Choose one resource that is disappearing and explain what can be done to save this source of natural resource. Use specific reasons and examples to support your opinion.

........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
B. Test prompts used in pre, post, and delayed post tests

Test prompt: **Resources Disappearing**

*Many parts of the world are losing important natural resources, such as forests, animals or clean water. Choose one resource that is disappearing and explain what can be done to save this source of natural resource. Use specific reasons and examples to support your opinion.*

Test prompt: **Stay in Hometown or Move**

*After they complete their university studies, some students live in their hometowns. Others live in different towns or cities. Discuss the advantages of each choice - living in your hometown or living in a different town or city? Which do you prefer? Give reasons and examples for your answer.*

Test prompt: **Learn from Mistakes**

*Do you agree or disagree with the following statement? People always learn from their mistakes. Use specific reasons and details to support your answer.*
Appendix B
Guidelines for Coding Clauses

Basically, a clause to be coded is comprised of an overt subject and a finite verb. An exception to this general rule is imperative clauses which do not require a subject to be considered clauses.

Ex: Don’t either kill or hunt animals or use the products from hunters.

1 clause / 1 clause

I. Independent clauses

1. When a coordinator is connecting two or more verb phrases, the subject and all the verb phrases are counted as a clause.

Ex: It encourages people to work hard and challenges them to try new things.

1 clause

* When three or more independent clauses are connected by the coordinators (and, but, or, so, nor, for, and yet), with the coordinator appearing only before the last independent clause, each clause is counted as one separate clause.

Ex: She was watching the news, her husband was reading a book, and their son was playing with his toy cars.

1 clause / 1 clause / 1 clause

2. Run-on sentences and comma splices are counted as independent clauses, with the error being counted in the first clause.

Ex: My school was in Saudi Arabia, it was the best school there.

1 clause / 1 clause

1 error error-free

1. If several comma splices occur in a row, count only the last as error free.

3. For sentence fragments

(a) If the verb or copula is missing, count the sentence as one clause.

Ex: My school best in the district.

1 clause

1 error

(b) If a noun phrase (NP) is standing alone, attach it to the preceding or following clause as appropriate and count it as a clause with an error.

Ex: My university in Vietnam, I liked it very much.

1 clause
1 error

(c) If the subject of a clause is missing, (excluding relative clauses where the relative pronouns are missing), count it as a clause with an error.

Ex: If can be punished by strict regulations, these people will be less likely to repeat their offences.

1 clause / 1 clause
1 error error-free

(d) If a subordinate clause is standing alone, count it as a separate clause with an error.

Ex: Because I want to be independent. I want to move to a different city.

1 clause / 1 clause
1 error error-free

4. Do not count tag-questions as separate clauses.

5. Direct quotes should be counted as an independent clause followed by a subordinate clause.

Ex: He said, “You should not be afraid of making mistakes”.

1 clause / 1 clause

6. Count clauses in parentheses as individual clauses.

Ex: I believe that she works hard (she gets good grades).

1 clause / 1 clause / 1 clause

II. Subordinate clauses

1. Adverbial and relative clauses are counted as separate clauses, except the following cases:

(a) For a reduced adverbial clause in the form of a present participle (Verb + ing) following adverbial conjunctions (after, before, when, on, prior to …), attach it to the preceding/following clause as appropriate and count it as a clause.

Ex: After completing their university studies, some students live in their hometowns.

1 clause
We finished all the tasks before going home last night.
1 clause

(b) Similarly, a reduced relative clause using a present participle (Verb + ing) or a past participle, attach it to the preceding clause as appropriate and count it as a clause.

Ex: The man driving the car is a friend of mine.

1 clause
We read the book written by our colleague.

1 clause

2. Nominal clauses

When a dependent clause is connected to the main clause with a nominal conjunction (whether (if), that, wh-words, etc.), it is counted as a clause.

Ex: *I think that we should deal with this problem using a more practical approach.*

1 clause / 1 clause

3. Comparative clauses

When a dependent clause is connected to the main clause with a comparative conjunction (as, than, the more., the more…), it is counted as a clause.

Ex: *She can sing more beautifully / than his mother does.*

1 clause / 1 clause

- When coding subordinate clauses, include complementizers used to introduce them in these clauses, not the main clauses. Complementizers may include:

(a) Subordinate conjunctions

Ex: *We should prevent deforestation / because forests play such important roles in human life.*

1 clause / 1 clause

(b) Relative pronouns

Ex: *We have to face serious environmental issues / that threaten our well-being.*

1 clause / 1 clause

(c) Relative adverbs

Ex: *Da Nang is the city / where I grew up.*

1 clause / 1 clause

(d) Complementizers introducing nominal clauses

Ex: *This also means / that it will make it easier for you to travel and find something new there.*

1 clause. / 1 clause

*I wonder / if she would like to join us.*

1 clause / 1 clause

*They know / what they need.*

1 clause / 1 clause
Appendix C
Coding Errors: General Accuracy Using Weighted Clause Ratio

“For most research purposes, global measures based on a syntactic unit can provide a better evaluation of accuracy in L2 performance than can local measures, and the best tool in the measurement kit is currently the error-free clause because it combines a reliably defined and valid unit with a finer-grained analysis than offered by a whole T-unit or AS-unit analysis”. (Foster & Wigglesworth, 2015, p.104)

The use of error-free clauses (EFC) serves as an index of learners’ overall accuracy demonstrated in their written texts. Following this statistic:

1. Clauses are coded as error free (entirely correct) or incorrect.

2. For incorrect clauses, code them as being at level 1, 2, or 3 depending on the gravity of the errors they have. In other words, based on the errors’ impact on the comprehensibility of the clauses, the following levels apply:
   * Entirely accurate (A): The clause is accurately constructed
   * Level 1 (1): the clause has only minor errors that do not compromise meaning.
     Some error types that can be considered minor are missing plurals (e.g. three car), third-person singular (e.g. He say), verb form (e.g. He working in the lab), or spelling. These errors tend to affect the text at local (lexical or sometimes phrasal) levels. Clauses at this level usually has only one or two errors at most.
     • Level 2 (2): The clause contains serious errors, but the meaning is recoverable, though not always obvious.
       Some errors that can be considered serious are those related to verb tense, inappropriate word choice, word form, or word order which tend to influence comprehension of the text at phrasal to clausal levels. This level also includes clauses that have two or three errors but the meaning is still mostly recoverable.
       • Level 3 (3): the clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.
       Some error types that can be considered very serious are totally wrong word choices, wrong syntactical structures or phrasal errors which blur meaning. Clauses rated at level 3 also usually
have several errors in a row, which makes meaning only partially recoverable (multiple errors) or comprehension is almost impossible due to convoluted sentence structures.

3. A clause is categorized according to its worst-level error, whatever that might be.

For a better understanding of the levels, below are examples of coded clauses using the above guidelines (Taken and adapted from Foster & Wigglesworth, 2016):

<table>
<thead>
<tr>
<th>Clause</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 As we all know,</td>
<td>accurate</td>
</tr>
<tr>
<td>2 Exams are an important part of education in almost all schools and</td>
<td>accurate</td>
</tr>
<tr>
<td>universities around the world.</td>
<td></td>
</tr>
<tr>
<td>3 However, recently, there are some kinds of widespread concerns over</td>
<td>1</td>
</tr>
<tr>
<td>the issue of the role of exams in education.</td>
<td></td>
</tr>
<tr>
<td>4 Firstly, exams can give the teachers chance to evaluate their students.</td>
<td>1</td>
</tr>
<tr>
<td>5 For most time, the test paper can reflect.</td>
<td>1</td>
</tr>
<tr>
<td>6 The second advantage of the exam systems lies in its function of</td>
<td>accurate</td>
</tr>
<tr>
<td>compelling students to study harder.</td>
<td></td>
</tr>
<tr>
<td>7 Other time than for get their own idea of putting the knowledge to practice.</td>
<td>2</td>
</tr>
<tr>
<td>8 And when he putting out of for the glass.</td>
<td>3</td>
</tr>
<tr>
<td>9 There is no wine in the bottle.</td>
<td>accurate</td>
</tr>
<tr>
<td>10 The cork is going out to the wall.</td>
<td>3</td>
</tr>
</tbody>
</table>

4. After the decision on the level that an incorrect clause fits, each coded clause is weighted using the scores proposed by Foster and Wigglesworth, as follows:
   • Entirely accurate clause: 1.00
   • Inaccurate clause at Level 1: 0.80
• Inaccurate clause at Level 2: 0.50
• Inaccurate clause at Level 3: 0.10

The scores assigned to all the clauses in a transcript are summed up and averaged for a mean accuracy rating for the transcript, with 1.00 being representative of a completely accurate L2 performance. For the total data in the current study, total clauses in all the participants’ transcripts are taken together to reach an averaged accuracy scores for each group, experimental and comparison, for both cross-sectional and longitudinal comparisons.
Appendix D

Coding Errors: Local Measures Using Obligatory Occasion Analysis

For the local measure which aims to investigate the effectiveness of a treatment (in the form of *Criterion* automated corrective feedback) on the development of target language, the target-like use (TLU) is conducted. The five grammatical forms for this analysis include *spelling, articles, subject-verb agreement, determiner-noun agreement*, and *comma usage* (i.e., *missing and extra commas*). This focused analysis of learner written accuracy takes into consideration over-suppliance of the grammatical forms in non-obligatory contexts, and thus it can capture a more complete picture of learners’ interlanguage development in terms of the targeted forms.

**Targeted Morphemes for Obligatory Occasion Analysis**

The error types of *articles, subject-verb agreement*, and *determiner-noun agreement* are operationalized in the following list of targeted morphemes:

- Article *a/an*
- Article *the*
- Third person singular *s/es* (excluding the use of auxiliary *do/does* in negative or interrogative constructions)
- Singular *copula* in both past and simple tenses
- *Plural copula in both past and simple tenses*
- *Plural s/es* with pre-modifying quantifiers
- *Plural s/es* without pre-modifying quantifiers

**Obligatory Occasion Analysis Calculation Formulas**

Scoring of TLU of the above targeted morphemes in learners’ essays across the pre, post, and delayed post-tests was conducted using instructions and procedures from Dulay and Burt (1974) as well as further elaborations from Pica’s (1982) seminal paper in studying adult acquisition of English as a second language. According to these studies, the first step in calculating TLU percentage of certain morphemes is the calculation of the percentage of *suppliance of the morphemes in obligatory contexts* (or SOC for short).

1. The formula to produce SOC is:

\[
SOC = \frac{(n \text{ correct suppliance in obligatory context} \times 2) + (n \text{ misformations in obligatory context} \times 1)}{\text{Total obligatory contexts} \times 2}
\]
2. Percentage of target-like use (or TLU analysis): For the above seven morphemes, a common TLU analysis formula is applied:

\[
\text{TLU} = \frac{n \text{ correct suppliance in contexts}}{n \text{ obligatory contexts} + n \text{ suppliance in non-obligatory contexts}} \times 100
\]

In order to take into account students’ efforts to produce the targeted morphemes instead of totally omitting them despite their failure to effect a correct form, misformations of targeted morphemes are included in the TLU formula. Therefore, the TLU formula used in the current study is adapted to:

\[
\text{TLU} = \frac{(n \text{ correct suppliance in obligatory context} \times 2) + (n \text{ misformations in obligatory context} \times 1)}{(n \text{ obligatory contexts} + n \text{ suppliance in non-obligatory contexts}) \times 2}
\]

**Procedures to Score the Above Morphemes in Obligatory Context**

To quantify morphemes for the TLU analysis using the above formulas, the following definitions apply:

- **Obligatory context**: adapting Dulay and Burt’s (1974) definition of obligatory context in spoken discourse among English as a second language children, “When a child speaks a language he is still learning, he will create obligatory context occasions for functors [i.e., morphemes] in his utterances, but he may not furnish the required forms. He may omit them, as in ‘he like hamburgers,’ where the 3rd person present indicative is missing, or he may misform them, as in ‘They do hungry’, where something was supplied for the copula, but it wasn’t quite the right thing” (pp. 43-44). Obligatory context in the current study refers to the occasions in student writing which requires a certain morpheme to fulfil the grammatical function of a structure in the target language (i.e., English).
- **Correct suppliance in obligatory context**: correct suppliance of each morpheme is counted as the total cases where the correct morpheme is provided in obligatory contexts for that particular morpheme.
- **Misformations in obligatory context** include:
- Substitution of a morpheme in an obligatory context with another morpheme inappropriate to that context. For example: (a) I am one of those who learning from my mistakes; (b) She reaches her achievement.

- Omission of the required morpheme in that context: Learning from mistakes help people to grow up.

• Suppliance in non-obligatory context is defined as the use of the morpheme under study in contexts where it is not required, as in the overuse of the plural s in “I have a cars” or “Their emotions associated with failure creates motivation for them to retrieve”.

Further Notes:

- For occasions where there is an error in learner writing such as “Mistake teach us to take responsibility”, the case should be coded as omission of the plural s and thus, no coding related to third person singular takes place. All in all, when there are errors at certain sections of the text, the correction that incurs the most minimal change to the text will take priority and coding should be done for that particular morpheme over other morphemes, as the above example shows.

- Regarding subject-verb agreement, cases of missing either the subject or the verb are not scored.

- Determiners are used with nouns to clarify the noun. Determiners targeted in this analysis, called pre-modifying quantifiers, include demonstratives (e.g., these, those, other,…), quantifiers (many, some,…), numbers (two, three,…) which take plural nouns.

Remaining Error Types

For the remaining error types of spelling and comma usage, the forms of analysis are as follows:

Spelling. Obligatory analysis is adapted and calculated as the number of misspelled words (in tokens) by the total words in an essay.

Comma usage. Due to the fact that rules related to the use of commas in English are volatile, flexible and sometimes debatable, the coding of comma usage errors only takes into account most relevant occasions for learner language where the use of commas are required, including six following rules⁴:

---

⁴ Adapted from https://webapps.towson.edu/ows/modulecomma.htm
1. *Use a comma before any coordinating conjunction (and, but, for, or, nor, so, yet) that links two independent clauses.*

- Ex: Cleo is a good singer, but she’s an even better dancer
  
  **Note:** If students chose **not** to use a comma in this situation, their written text is acceptable and considered error free. This is similar to simple sentences with compound elements where no comma is required.

- Ex: Marion read the paper but didn’t agree with the author’s idea.

2. *Use a comma after a dependent clause that starts a sentence.*

- Ex: Because the smell of maple syrup makes me queasy, I avoid going into pancake houses.

3. *Use commas to offset appositives from the rest of the sentence.*

- Ex: Godfrey, an amateur taxidermist, dreams of opening a cafe.

4. *Use commas to separate items in a series.*

- Ex: Mary needs bread, milk, and butter at the grocery store or Mary needs bread, milk and butter at the grocery store.
  
  **Note:** The serial comma before “and” is optional, but the comma between bread and milk is compulsory.

- Ex: People defined their mistakes, changed it and got many useful things.

5. *Use a comma after introductory adverbs or a long introductory phrase*

- Ex: Finally, I went running.

- Ex: During the hot summer of 1984, the temperatures set records.

6. *Use a comma when attributing quotes.*

- Ex: The runner said, “I saw a duck.”

**How to score comma usage?**

- **Obligatory context:** Obligatory contexts related to comma usage refers to the occasions as outlined in the six provided rules above in student writing which require the use of a comma to fulfil the grammatical function of a structure in English.

- **Correct suppliance in obligatory context:** Correct suppliance of a comma where there is an obligatory context, restricted to the six occasions described in the above rules.

- **Misformations in obligatory context:** Substitution of a comma in an obligatory context with another punctuation mark inappropriate to that context.

- **Omission in obligatory context:** Omission of the comma in an obligatory context.
- Ex: Therefore I think that people always learn anything from their mistakes
- *Suppliance in non-obligatory context (Over-suppliance)* is defined as the use of a comma in contexts (1) where it is not required or (2) where another punctuation mark is required.
- Ex: There are people, who are always careful. (1)
- Ex: Life has many challenges, people differ in their potentials of facing problems. (2)
Appendix E
Guidelines for Coding Revision Practices

A. Verification of error codes generated by Criterion

This part of coding was done in combination with the evaluation of Criterion feedback using the Verification Annotation Approach. Evaluation of Criterion automated corrective feedback was done using three categories below:

- **Correct error code** (CC): to indicate that an error was correctly identified by Criterion

  *Example student text 1:* Another benefits of staying at the same job for a long time is that you will have strong work relationship

  *Error message:* You may have used the wrong determiner. Proofread the sentence to make sure that the determiner agrees with the word it modifies. (Determiner-noun agreement)

  *Example student text 2:* It may be difficult for they to form strong relationship that endure after they stop working.

  *Error message:* You may have used the wrong pronoun. (Pronoun error)

- **Incorrect error code** (IC): to indicate cases when Criterion appropriately coded a structure as incorrect but gave it a wrong code, or when Criterion offered a confusing error message that failed to pinpoint the nature of the problem

  *Example student text 1:* Someday, he find another career make him happier is journalist.

  *Error message:* The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb. (Subject-verb agreement)

  *Example student text 2:* First, staying in the same career, it is easy for some people to working because it dose not waste much time to start with new jobs.

  *Error message:* The subject and the verb in this sentence may not agree. Reread the sentence and look closely at the subject and the verb. (Subject-verb agreement)

- **False positive** (FP): to indicate that the system provided a false alarm by flagging a correct structure in essays as an error

  *Example student text 1:* The prevalence of changing jobs has been a growing concern in the past few years

  *Error message:* You may be using the wrong preposition. (Preposition errors)

  *Example student text 2:* Actually, it leads to many bad things that affect to our environment and after that are our health.
Error message: You have used to in this sentence. You may need to use too instead. (Confused words)

B. Students’ textual operations

Students’ revision operations in response to Criterion ACF are coded using five categories adapted from Lee and Hegelheimer (2012, as cited in Chapelle et al., 2015, p. 389).

1. **No change**: Instances when students did not make any change to their text. In other words, the original text remained unchanged as the student avoided addressing the identified error or decided to keep the original correct form after consulting the automated feedback which is a false positive

*Example student text:*

*First draft:* It can't not be denied that there are advantages of changing job. (Negation error)

*Revised:* It can't be denied that there are advantages of changing job.

2. **Removal**: Instances when students deleted the part of the essay (phrase, clause, or whole sentence) that included the highlighted error in the revision process

*Example student text:*

*First draft:* Moreover, another benefits of staying in the same jobs is that it can open door for employers to learn and advance their skill or their jobs in their career. (Determiner-noun agreement)

*Revised:* Moreover, changing career helps them gain expertise in a new area and make them have more opportunities in the future.

3. **Addition**: Instances when students revised the essay by adding word(s), phrase(s), or sentence(s)

*Example student text:*

*First draft:* This news has been receiving many different comments and strongly influencing to people's thoughts, especially the youth about money and happiness. (Preposition error)

*Revised:* This news has been receiving many different comments and strongly influencing to people's thoughts, especially the youth's thought about money and happiness. (Adding a word)

4. **Deletion**: Instances when students revised the essay by deleting word(s), phrase(s), or sentence(s)

*Example student text:*

*First draft:* The question if money is a main indicator of people's being a controversial one. (Possessive error)

*Revised:* The question if money is a main indicator of people being a controversial one. (Deletion of the apostrophe ‘ and the s)

5. **Substitution**: Instances when students revised the essay by changing word(s), phrase(s), or sentence(s) with alternatives

*Example student text:*

*First draft:* They are not keen on there job anymore. (Confused words)
Revised: They are not keen on their job anymore.

6. Reordering: Instances when students revised the essay by reordering certain parts of a clause or swapping the positions of phrases in a sentence

*Example student text 1:*

*First draft:* This not only affect to human’s health when they use water from rivers or lakes sources, but also threaten many kinds of fishes, shrimps in the seas. (Proofread this!)

*Revised:* This not only affect to people's health when they use water sources from rivers or lakes, but also threaten many kinds of fishes, shrimps in the seas.

*Example student text 2:*

*First draft:* I think successful person is person who have a good characteristics and be looked up to by everyone. (Garbled sentences)

*Revised draft:* I think those who has good characteristics and be looked up to by everyone is a successful person.

**C. Revision outcomes**

The changes are further coded based on the outcome using the binary coding of (1) correct versus (2) incorrect revision. Those previously coded as *No change* and *Removal* will be merged under the code of students’ avoidance of addressing ACF, except for *No change* cases which are responses to false positives and learners decided to retain the originally correct forms of their texts. As a result, this second tier of coding will have three codes:

- **Correct revision** (CR): instances when students were successful at addressing a tagged error by producing a correct revised form

  *Example 1*

  *First draft:* For example, not all of people achieve the goal which they put in their life, they have to do works they don’t have passion for, after a long time, when they have enough abilities for the major they dream about, they can change the job to find a new motivation and energy. (Run-on sentences)

  *Revised:* For example, not all of people achieve the goal which they put in their life, after graduating, they have to do works they don't have the passion in. For a long time, when they have enough abilities for the major they dream about, they can change the job to find a new motivation and energy.

  *Example 2*

  *First draft:* In conclusion, finding a new job or profession during careers is a best idea for everyone. (Wrong article)
Revised draft: In conclusion, finding a new job or profession during careers is the best idea for everyone.

• Incorrect revision (IR): Instances when students failed to address the tagged error, resulting in an incorrect revised form

Example

First draft: It is a good chance for their to gain more experiences than they had. (Confused words)

Revised: It is a good chance for they to gain more experiences than they had

• Avoidance (AVOID): Instances when students avoided addressing the tagged error by either making no change to an incorrect structure or removing the part of text that included the tagged error

Example

First draft: The enviroment need to defend by human, Humans should have been educated more enviroment. (Extra comma)

Revised text: The enviroment need to defend by human, Humans should have been educated more enviroment.

• Retention of correct form (RE): Instances when students decided not to make any change to a tagged error as it was a false positive generated by Criterion. In other words, the originally correct form is retained despite the error code.

Example

First draft: What is more, you will make the most of your youth. (Missing question mark)

Revised text: What is more, you will make the most of your youth.
Appendix F
Guidelines for Coding Think-aloud Protocols

Adapting two broad levels of engagement with CF, “substantive” and “perfunctory” noticing (Qi & Lapkin, 2001) or “extensive” versus “limited/no engagement” (Storch & Wigglesworth, 2010), the following list of codes for think-aloud protocols was generated and comprised of nine categories addressing different strategies when students came into contact with the feedback. The use of this list of codes aims to shed light on more fine-grained representations of what took place as students engaged with the corrective feedback generated by Criterion. These nine categories are divided under perfunctory and substantive noticing episodes as below:

1. Perfunctory noticing
   • Adopting ACF
   • Guessing the correct form
   • Quickly revising the form using ACF as a red flag
   • Noticing without attending to the error

2. Substantive noticing
   • Drawing on stored metalinguistic knowledge
   • Getting cues from ACF message to revise
   • Looking up online resources
   • Seeking peer/teacher feedback
   • Translating into L1

Further explanations of the categories and illustrative examples are below:

1. Adopting ACF: Instances when students read Criterion’s error explanation and its suggested corrections, then replacing their original text with one of the suggested corrections.

Example: And kindness, you may need to use a comma after this word. OK I see, kindness comma and sometimes I. OK. Kindness….creativity, pride, comma, kindness, comma, and sometimes as simple things…OK.

Extra comma? I think the comma is correct here. You may need to remove this comma? Delete? Another example is that the salary of…Delete, Need to remove this comma, need to remove this comma. Delete the comma? Let’s try deleting the comma here.
2. **Guessing the correct form**: Instances when students did not make any attempt to use any resources. They simply came up with an alternative form to quickly replace the tagged error. During these instances, students explicitly verbalized their guessing using expressions such as “I guess…, I think…, Let’s try…, maybe…”

*Example 1*: Opportunity, maybe, maybe add the article the. The opportunity, yeah, I think so.

*Example 2*: Choose the choose different ways, maybe I will. Choose the different ways. I will, I will choose the different ways. Choose different ways or choose the different ways? Choose, maybe I will delete the.

*Example 3*: The second answer, second answer, one answer, second answer, missing, second answer. Why one answer is … second answer, what is second answer, the second answer? I think so. Maybe, the second answer.

3. **Quickly revising the form using ACF as a red flag**: Instances when students very promptly recognized their errors thanks to Criterion’s highlighting of the errors. Simply seeing the highlighting is sufficient to bring them to a correct form. No further reading of error explanations or seeking other resources was needed.

*Example*: My god, salry. Ah. Miss the letter A. Determiner. These, these trend. I will add the s after trend, trends.

4. **Noticing without attending to the error**: Instances when students briefly looked at the error tags or Criterion’s error explanations, then quickly decided not to respond to it before moving on to the next error.

*Example 1*: The affect of this problem OK. To reduce the affect of this problem, OK. Don’t know how to check, so I will make it… will keep it.

*Example 2*: Ah, a lot of money, last but not least, but why, I often use it, I think it's right. So, I don't correct it.
5. **Drawing on stored metalinguistic knowledge:** Instances when students engaged with the feedback by searching for some clues in their previously acquired knowledge about the English language for the relevant information to produce a correct form for the tagged error.

*Example 1:* *Family.* You may need to use an article before this word. Use an article before this word? No, friends or family is, is general. I don't think use an article before this word.

*Example 2:* Ill-formed verbs. Let me see. *Which makes them to …* used wrong form of this verb. Ah OK. *Make somebody.* OK. *Make somebody* plus verb infinitive, *not to.* OK. I’ll delete *to.*

6. **Getting cues from ACF message to revise:** Instances when students examined *Criterion* error explanation in details to look for some clues to correct the tagged error.

*Example:* *Important.* I strongly agree…I believe this is correct. *Successful.* Let’s see what suggestions does it offer here. *You may need to use an article before this word.* Consider using *article a.* Adding *a* here? *Is a important factor.*

7. **Looking up online resources:** Instances when students, after noticing the highlighted error or reading *Criterion* error explanation, referred to an online website to seek a correct form for the tagged error. They can either use their smartphone or the computer they are working on.

*Example:* *Strengths.* *Strengths.* Where is this? OK, S T R E N G T H T, now let’s see S T E A N G H T. Now I want to I want the help of the dictionary [use google translate] *Strength.* *Strength?* S T … S … S T R E N G T H. OK. G T H, not G H T. Yeah that’s right.

8. **Seeking peer/teacher feedback:** Instances when students engaged with the feedback explanation from *Criterion* and consulted their peer or teacher to seek help in dealing with the error being tagged.

*Example:* And one missing comma. Where is it? Show me show me … Yeah, here. Where are you, missing comma? *Last but not least. Last but not least,* I thought it is like a collocation. … OK I will ask Ms G here. I think I’m right, not wrong. Ms G, can I ask you something? *Is this right or wrong?* Do I need a comma before, after *last?*
9. **Translating into L1**: Instances when students translated their own writing or *Criterion* error explanation into L1 in search of some clues for a correct answer to the tagged error.

*Example*: What’s this about? What's up. ok, Grammar, *the, in the sing my song*, Cao Bá Hưng đã tham gia [Vietnamese translation]. ok, *they* instead of *the, they*. This has to do with *they*. I think this has nothing to do with *they*. Just delete *the*, delete *the*.

In order to arrive at the final decision regarding whether a *Criterion* error tag was substantively or perfunctorily noticed in a revision episode, the following guidelines apply:

1. In cases where students only used **one** of the nine strategies above to engage with the feedback and resolve the highlighted error, the noticing category (either *perfunctory* or *substantive*) to which the strategy belongs determines the level of noticing (or level of cognitive engagement with the feedback).

2. In cases where students used more than one strategy listed above, the decision is made using the following rules:
   a. Two or more strategies under perfunctory noticing are considered **one** perfunctory noticing episode.
   b. Any strategy under perfunctory noticing in conjunction with another strategy/strategies under substantive noticing results in a substantive noticing episode.
Appendix G

Ethics Approval Letter

16 Nov 2016

Dr Neomy Storch
Languages and Linguistics
The University of Melbourne

Dear Dr Storch

I am pleased to advise that the Faculty of Arts Human Ethics Advisory Group (HEAG) has approved the following Minimal Risk Project.

Project title: Exploring Automated Corrective Feedback in EFL Writing Classroom Context: A Case Study of Criterion

Researchers: Dr N Storch, A/Prof C Roever, T Hoang

Ethics ID: 1648139

The Project has been approved for the period: 16-Nov-2016 to 31-Dec-2017.

It is your responsibility to ensure that all people associated with the Project are made aware of what has actually been approved.

Research projects are normally approved to 31 December of the year of approval. Projects may be renewed yearly for up to a total of five years upon receipt of a satisfactory annual report. If a project is to continue beyond five years a new application will normally need to be submitted.

Please note that the following conditions apply to your approval. Failure to abide by these conditions may result in suspension or discontinuation of approval and/or disciplinary action.

(a) Limit of Approval: Approval is limited strictly to the research as submitted in your Project application.

(b) Amendments to Project: Any subsequent variations or modifications you might wish to make to the Project must be notified formally to the Human Ethics Advisory Group for further consideration and approval before the revised Project can commence. If the Human Ethics Advisory Group considers that the proposed amendments are significant, you may be required to submit a new application for approval of the revised Project.

(c) Incidents or adverse effects: Researchers must report immediately to the Advisory Group and the relevant Sub-Committee anything which might affect the ethical acceptance of the protocol including adverse effects on participants or unforeseen events that might affect continued ethical acceptability of the Project. Failure to do so may result in suspension or cancellation of approval.

(d) Monitoring: All projects are subject to monitoring at any time by the Human Research Ethics Committee.

(e) Annual Report: Please be aware that the Human Research Ethics Committee requires that researchers submit an annual report on each of their projects at the end of the year, or at the conclusion of a project if it continues for less than this time. Failure to submit an annual report will mean that ethics approval will lapse.

(f) Auditing: All projects may be subject to audit by members of the Sub-Committee.

Please quote the ethics ID and the name of the Project in any future correspondence.

On behalf of the Ethics Committee I wish you well in your research.

Yours sincerely

A/Prof Ingrid Volkmer
Chair Faculty of Arts HEAG Human Ethics Advisory Group

Faculty of Arts
The University of Melbourne, Parkville Victoria 3010 Australia
T: +61 3 8344 7991 E: arts-research@unimelb.edu.au
W: www.arts.unimelb.edu.au