A Study of Student and Teacher Perceptions of Criterion,
an Online Writing Program

by

Ronald Shutler

A thesis submitted to the Faculty of Graduate and Postdoctoral
Affairs
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Applied Linguistics and Discourse Studies

Carleton University
Ottawa, Ontario

© 2012

Ronald Shutler

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-91598-1

Our file Notre référence

ISBN: 978-0-494-91598-1

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

This thesis reports on a study of student and teacher perceptions of the Criterion computer program — an online system that allows students to enter short essays and then immediately receive automated evaluation and feedback. The research explored six English as a Second Language (ESL) classes at a local college and used a mixed method approach composed of a quantitative questionnaire administered to all 73 students as well as qualitative oral interviews with six volunteer students and the six teachers of the classes. Drawing perspectives from sociocultural theory, genre theory, ESL writing pedagogy, and other Computer Assisted Language Learning (CALL) research, the study only found weak statistical correlations between background student data and their satisfaction levels with various components of the program. Through analysis of the interview data, five key themes emerged: feedback, satisfaction with the program, lack of understanding of the program, time constraints, and technical issues. These issues and their implications are discussed.

Acknowledgements

I would first like to thank my advisor, Jaffer Sheyholislami, for guidance and encouragement, and for sharing his vast store of knowledge about linguistics. Many other Carleton University faculty members provided help and assistance at various times as well. I especially wish to thank Joan Grant for always knowing the answers to administrative questions. Thanks are also due to Mike Barker for his encouragement, help and valuable suggestions for various computer-related issues. I am grateful to Maggie Addison for her help in diligently reviewing my Literature Review section. Any remaining errors are, of course, my fault.

I would also like to thank the many people from the college where I conducted my research. I am indebted to Sylvia Garcia who not only allowed access to the students and teachers but facilitated and helped co-ordinate things for me. Thanks are also due to the computer lab personnel including Iryna Pimashenko and Beth Thompson who allowed me to use lab space and let me use their license to access Criterion. I would especially like to thank the unnamed teachers and students who filled in questionnaires and participated in one-on-one interviews. This research would not be possible without them.

Finally, I would like to thank my family and friends and fellow students for providing support and encouragement throughout my thesis work and for tolerating my cranky moods that I had no right to inflict on them. Most of all, I wish to express my gratitude to my wife, Kathy Neufeld, who did proofreading work and took the brunt of the cranky moods but still provided unwavering support and encouragement.

Table of Contents

Abstract	ii
Acknowledgements	iii
List of Figures	vii
List of Tables.....	viii
List of Appendices	ix
List of Abbreviations.....	x
Chapter 1: Introduction	1
1.1 Criterion.....	3
1.2 Research Question	4
1.3 Method.....	5
1.4 Framework.....	5
Chapter 2: Literature Review	8
2.1 Background Theory for Teaching of ESL Writing.....	8
2.1.1 Sociocultural Theory	8
2.1.2 Genre Theory.....	10
2.1.3 Attitude and Motivation	14
2.1.4 ESL Teaching Methods.....	16
2.1.5 Teaching of ESL Writing	16
2.1.6 ESL Writing Feedback	18
2.1.7 ESL Writing Evaluation	22
2.2 Computer Assisted Language Learning - CALL.....	26

2.2.1 History of CALL	26
2.2.2 Natural Language Processing by Computers	28
2.2.3 CALL and Writing	29
2.3 Criterion and related programs	31
2.3.1 Automated Writing Evaluation (AWE).....	31
2.3.2 General description of Criterion.....	33
2.3.3 Criticisms of Criterion and similar programs.....	34
2.3.4 Validity.....	36
2.3.5 Do these programs improve writing?	37
2.3.6 Perceptions	40
2.4 Summary.....	42
Chapter 3: Methodology.....	44
3.1 Research Question	44
3.2 Ethical Considerations.....	45
3.3 Participants	46
3.4 Data Collection.....	47
3.4 Data Analysis Method	48
3.4.1 Quantitative Analysis	49
3.4.2 Qualitative Analysis	50
Chapter 4: Results and Analysis.....	54
4.1 Quantitative Results.....	54
4.1.1 Description of Data	54
4.1.2 Statistical Tests.....	64

4.2 Qualitative Results.....	68
4.2.2 Emergent Themes.....	73
4.3 Feedback Usefulness	74
4.4 Satisfaction with the program.....	82
4.5 Understanding of the program.....	85
4.6 Time Constraints	90
4.7 Technical Issues.....	93
4.8 Student and Teacher Recommendations.....	98
4.8.1 More Training	98
4.8.2 Continue to Use the Program	99
Chapter 5: Discussion and Conclusion	100
5.1 Summary of Findings and Discussion.....	100
5.2 This Study's Contributions	108
5.3 Limitations of this Study	110
5.4 Implications for Further Research Directions	111
5.5 Summary.....	112
References	113
Appendices	125

List of Figures

Figure 1 Age distribution of students.....	56
Figure 2 Gender distribution	56
Figure 3 Distribution of ESL level.....	57
Figure 4 Distribution of time spent in Canada	58
Figure 5 Distribution of hours spent on a computer per day.....	59
Figure 6 Distribution of number of essays written.....	60
Figure 7 Distribution of student approval of various Criterion components	61
Figure 8 Average numerical agreement with questions	62
Figure 9 The decomposition of e-rater into features (Enright & Quinlan, 2010)	126
Figure 10 Immediate feedback in Criterion (http://www.ets.org/criterion/ell/demo/)....	128
Figure 11 Grammar feedback in Criterion (http://www.ets.org/criterion/ell/demo/).....	128
Figure 12 Organization and Development feedback in Criterion	129

List of Tables

Table 1 Spearman's Rho Correlation Coefficients	66
Table 2 Axial Coding categories	69
Table 3 Sample excerpt of transcription data.....	130

List of Appendices

Appendix A - Detailed Description of Criterion.....	125
Appendix B - Sample of Transcription Data.....	130
Appendix C - Student Written Questionnaire	132
Appendix D - Student Interview Questions	134
Appendix E - Teacher Interview Questions	136

List of Abbreviations

AI	Artificial Intelligence
AWE	Automatic Writing Evaluations
BASIC	Beginner's All-purpose Symbolic Instruction Code
CALL	Computer Assisted Language Learning
CCCC	Conference on College Composition and Communication
CMC	Computer Mediated Communication
ESL	English as a Second Language
ESP	English for Specific Purposes
ETS	Education Testing Services
FCAT	Florida Comprehensive Assessment Test
GMAT	Graduate Management Admission Test
IADE	Intelligent Academic Discourse Evaluator
L1	First Language
L2	Second Language
NLP	Natural Language Processing
PEG	Project Essay Grade
PLATO	Programmed Logic for Automated Teaching Operations
RGS	Rhetorical Genre Studies
SPSS	Statistical Package for the Social Sciences
TOEFL	Test Of English as a Foreign Language

Chapter 1: Introduction

Imagine that someone you know — an adult — wants to learn a second (or third) language. Imagine a world where she can get unlimited practice in the language by carrying out two-way interactive spoken and written conversations about engaging and interesting topics with her computer and have that computer provide feedback, corrections and instructions until she becomes fluent. Imagine a world where that computer is a desktop, a laptop, a smartphone, a robot, or a surgically implanted chip. That world does not exist yet, but the field of Computer Assisted Language Learning (CALL) holds great potential for providing invaluable tools like that for language learners. Today's personal computer technology does not have human constraints like location, schedules, or amount of available time; computer technology can work at a time and place that is convenient for the user.

From the very early days of the development of computers, using them as a teaching tool has been promoted. According to Ahmad, Corbett, Rogers and Sussex (1985), using computers for education was discussed as far back as the 1940s. In the mid 1960s, Stanford University developed one of the first CALL systems to teach Russian. In this system, students used a teletype machine to enter responses to questions delivered in Russian via a tape recorder (Ahmad et al., 1985). Today, the field of CALL is so large that there are several journals such as *CALICO* and *ReCALL* which are dedicated to the subject. An entire issue of *The Modern Language Journal* (Special Issue 1 of 2009) was also dedicated to the subject. The teaching of language by computers is closely related to the research field of Natural Language Processing (NLP) by computers. NLP research

looks into the ability of computers to "understand" and use language in a natural way.

Arguably, it could be said that the ability to properly communicate in a language requires knowledge of not just grammar, syntax, and vocabulary, but also of extra-linguistic information about context and culture — not just linguistic competence but socio-linguistic competence. Unfortunately, there is no computer system today that can truly "understand" and use language in a natural way like humans (Abramson, 2011). This is a serious limitation on the ability of CALL to effectively teach a language today.

My interest in the field of CALL started with an interest in computers — an industry where I worked for over twenty-five years. My role in the computer industry included a great deal of teaching about computers. Recently, I changed careers to become a different kind of teacher, an English as a Second Language (ESL) teacher, partly because teaching a language seemed more important (and more rewarding) than teaching about computers. However, I soon discovered that the two fields were not so independent from each other. I realized that not only did computer technology have the potential to help language learners in many ways, but that it was already doing just that. My background with computers has allowed me to be aware of the potential for computers, but also to be aware of the limitations of current technology. This is why when I first heard about a computer program called Criterion (described below), I was very skeptical about its claims because I knew what computer technology was or was not capable of. I must admit that I was initially impressed with the program, although I could see that it is by no means perfect. This has led me to make Criterion the subject of this thesis.

1.1 Criterion

First, an explanation of the Criterion program is in order. Criterion is an online computer program which provides guidance for and evaluation of essays written by ESL students (Education Testing Service, 2011a). The program is expressly designed to help students learn writing skills. The Criterion program was developed by the company ETS (Education Testing Services). The Criterion program evaluates student essays and gives feedback in five categories: Grammar, Usage, Mechanics, Style, and Organization & Development (Education Testing Service, 2008). The ETS company has published some research about Criterion (Enright and Quinlan, 2010) which makes claims about the validity of the evaluation part of the program. It has also published a study showing that Criterion has a small beneficial effect on the writing skills for a native English ninth-grade class (Rock, 2007). Some authors connected with ETS (Burstein, Chodorow & Leacock, 2004) have reported on a longitudinal study conducted over a 20 week period that compared students writing essays using Criterion with students writing in traditional classroom settings. Although that study showed no significant differences between Criterion users and the control group in the results from the FCAT (Florida Comprehensive Assessment Test), the authors did claim student improvements in writing production and a reduction in the number of writing errors tracked Criterion. Despite the positive views of Criterion, this research by a company about its own product cannot be considered completely objective. I have been unable to find research on Criterion not connected to ETS. It would seem that an impartial independent study is needed to really understand how well the program works.

1.2 Research Question

Ideally, an independent longitudinal study similar to the Burstein et al. study mentioned above comparing the learning outcomes of a large group of students using Criterion for an extended period of time with another large control group of students not using Criterion should be conducted. However, this would require more extensive resources and a longer period of time for data collection which are not available to me. Instead, an in-depth exploration of the perceptions of people using the program (both students and teachers) has the potential to be very informative and is also within the scope of this Master's level thesis. Peters, Weinberg and Sarma (2009) argue that the computer activities that students like the most are also judged by the students to be the most effective. According to Leki (2000), "Within the L2 writing classroom, arguably the most potent aspect of the context for learning is the writing teacher's response to the students' writing" (p.104). If the writing teacher's response is to be replaced by a computer's response, it therefore becomes even more important to better understand how the students perceive this.

Perceptions are important because McDowall and Jackling (2006) showed that positive perceptions of the usefulness of computer assisted learning (in the accounting field) improved performance. Perceptions are coloured by attitudes and there is a correlation between attitudes that students have and their achievements in language learning (Masgoret & Gardner, 2003). In addition, an exploration of Criterion's teaching of writing must logically consider its adherence to pedagogical theories about teaching writing. Therefore, this study sets out to answer the following research questions:

What are the perceptions of both students and teachers about the Criterion computer program?

How well does Criterion adhere to current pedagogical theories and practices?

1.3 Method

In order to answer these questions, I have conducted a mixed method study which included quantitative questionnaire data and qualitative interview data. By using both quantitative and qualitative data, triangulations between the two can often be done to strengthen the findings. I studied six classes (at ESL Level 7 and Level 8) at a college in Ottawa where the Criterion program is being used for ESL writing classes. The study was conducted during the October 24 to December 16, 2011 session of these classes. All students were asked to fill in a written questionnaire — a total of 73 students. In addition, I conducted one-on-one recorded interviews with all six teachers and six volunteer students. For the student interviews, I also recorded the computer screen as we were discussing various components of the program. In order to analyze the qualitative interview data, I have used a variation of the grounded theory approach of Glaser and Strauss (1967) which could simply be called a thematic analysis. My analysis involved three stages — open coding, axial coding and selective coding and is based on techniques documented by Flick (2009) and Gibbs (2007). The key themes that I will analyse in the Results and Analysis chapter emerged through this coding process.

1.4 Framework

Throughout this thesis, I have used sociocultural theory and genre theory as a framework in which to view the teaching of writing in a second language. Sociocultural

theory posits that language learning can only be understood by considering all the social, cultural, and historical factors in which that learning is embedded (Donato & McCormick, 1994). Part of the theory as described by Wertsch (1991) includes the concept of mediated action. Wertsch (1991) says that all human action employs mediational means and that the mediational means shapes the action. Students whose action is learning to write in English use Criterion as a mediational means. By using sociocultural ideas in my study, this leads to an appreciation for the great number of environmental factors that have a bearing on Criterion's ability to aid students with their language learning. As well, genre theory points to the idea that writing itself must be considered in terms of the rhetorical purposes of the piece of writing and that an understanding of genres is critical to successful writing (Freedman & Medway, 1994). Genre theory is related to sociocultural theory since it views writing as a social process (Coe, 1994). Within genre theory there are different viewpoints. One school of thought (Berkenkotter and Huckin, 1993) have argued that the knowledge of what genres are and how and when to use them is not teachable while Martin (2009) suggests that genres should be taught. Therefore, Criterion should be examined in light of these beliefs.

This thesis is organized into five chapters. After this introduction chapter, chapter two contains a literature review composed of three main sections. The first section of the literature review explores the teaching of writing in the context of sociocultural theory as well as genre theory. The second section will delve a bit more into the field of CALL, and the third section will look into the literature about Criterion and related programs. Chapter three will describe the methodology including the questionnaire, the interviews, and the data coding approach. Chapter four contains a description of both the quantitative

and qualitative results, and some interpretation of those results. The coding of the qualitative results leads to the emergence of the main themes which are then analysed in this same chapter. Finally, in chapter five, the five themes and their relevance to other research is discussed. The chapter also discusses some implications of the results and some potential future directions for research.

Chapter 2: Literature Review

In this chapter, I will attempt to construct a theoretical framework within which the usefulness of the Criterion program can be analysed. To that end, the following theories will be drawn upon: sociocultural theory and genre theory, ESL writing pedagogy theory, as well as Computer Assisted Language Learning (CALL) research. This chapter is organized into three sections. The first section will discuss the background theory for the teaching of ESL writing. Included in this first section will be a discussion of sociocultural theory and genre theory and their relevance to the teaching of ESL writing. Also in this section, I will examine the factors of student attitudes, student motivation, feedback and evaluation. The second section contains a review of Computer Assisted Language Learning (CALL) including Natural Language Processing (NLP) by computers and a review of CALL in the specific domain of writing. Finally, the third section reviews the literature about Criterion and related programs.

2.1 Background Theory for Teaching of ESL Writing

2.1.1 Sociocultural Theory

Sociocultural theory states that all mental processes (such as the process of writing) must be viewed not only in terms of the individual, but in terms of the entire social and cultural contexts in which the mental processes occur. The theory is largely based on the work of a Russian psychologist Lev Vygotsky (Donato & McCormick, 1994). Sociocultural theory uses the term "genetic analysis" to highlight that many aspects of mental processes can be better understood if one understands the origin (genesis) and historical transitions that the mental processes have undergone (Wertsch,

1991, p. 19). In other words, sociocultural theory considers not only the present sociocultural context but also the historical factors that have led to the current context. Sociocultural theory views activity as being composed of four parts: subject (typically the person taking the action), an object (the goal or purpose of the action), actions (what the subject undertakes to achieve the goal such as reading a textbook), and operations (the context of those actions) (Donato & McCormick, 1994). For this study, ESL students are the subjects, the object is a desire to function in an English environment, the action is mainly attending an ESL class in which Criterion is used as a tool, and the operations are all the surrounding environmental factors of the college, the city, and the country.

Sociocultural theory suggests that the mental processes involved in learning how to write in a second language can be viewed in terms of a mediated activity — whether that mediator is a teacher, a textbook, a computer, or language itself (Donato & McCormick, 1994). Wertsch (1991) defines mediated action by saying that all human action employs mediational means and that the mediational means shapes the action. In other words, human actions must be viewed as actions within the larger sociocultural context and with respect to the mediator. Sociocultural theory also points out that in order to truly understand the learning strategies that students use, the mediator must be considered in its context. The context also includes numerous levels of sociocultural influence such as the classroom, the educational institution, the country, and the student's own history. Donato and McCormick (1994) go on to say that mediators lead the learners from natural impulses to higher cognitive processes. They summarize this by stating that "Mediation is, thus, the instrument of cognitive change" (p.456).

In a paper discussing the use of technology for language learning, Kern (2006) comments about the sociocultural connectedness of learning when he states that "[t]here is consensus in CALL research that it is not technology per se that affects the learning of language and culture but the particular uses of technology" (p. 200). Kern goes on to say that language teachers need to have a heads-up approach to teaching in that they need to be aware of the relationship between the technology, culture, ideology and the outside world. By examining the Criterion program as a mediator in the context of sociocultural theory as well as how "heads-up" the teachers are, it allows us to be more aware of the many additional social and cultural factors involved that have a bearing on the program's success. Sociocultural theory may be useful for examining the more general concepts of learning, but a related theory which is much more specific to learning how to write is Genre theory.

2.1.2 Genre Theory

I will now provide a brief overview of genre theory and its relationship with the teaching of writing. In recent literature, the concept of genre-based writing pedagogy has become popular. Earlier definitions of genre focused on the textual form and substance of the written work; however, today's definition of genre in the field of linguistics also focuses on the rhetorical purposes of writing (Freedman & Medway, 1994). A piece of writing fits into a given genre (e.g., a business letter, or a lab report) based on its purpose but also based on the idea that it occurs repeatedly and therefore acquires recognizable features (Freedman & Medway, 1994). As Freedman and Medway put it, genres are "seen as typical ways of engaging rhetorically with recurring situations" (p. 2).

Of the various genre approaches to language use, three key ones I will describe here are New Rhetoric studies which are also called Rhetorical Genre Studies (RGS), the Sydney School, and English for Specific Purposes (ESP) (Hyon, 1996). The ESP approach focuses on both the function and form of genres with an emphasis on the patterns of text that occur in a given genre (Swales, 1990). The RGS approach tends to look at genres in terms of sociocultural actions and in a holistic way. The Sydney School, which was started from the field of Systemic Functional Linguistics, also considers genre as a social action but initially focused on what primary and secondary school students needed to understand about genres so it has tended to focus more on the detailed linguistic features. One characteristic of the Sydney School's view of genre is that genres are connected to the values of the ruling elites (Freedman & Medway, 1994). Those who are in positions of power are better able to utilize the appropriate genres to further their aims while those outside that elite group are disadvantaged because the existing preferred genres did not stem from their own values and beliefs. Freedman and Medway explain how this is a feedback loop because "[t]he perception of the genre contributes to the perpetuation of those forms of power and exclusion" (p.14).

One of the key differences between RGS and the Sydney School is in their view of how students learn to write and whether genre can be taught. Some scholars from the RGS group express doubts about whether it is even possible to teach genres in the classroom. For example, Berkenkotter and Huckin (1993) have argued that the knowledge of what genres are and how and when to use them is not teachable. They say that this knowledge, "[r]ather than being explicitly taught, is transmitted through enculturation as apprentices become socialized to the ways of speaking in particular

disciplinary communities" (p.482). On the other hand, Martin (2009), from the Sydney School, argues that genres should be taught. He says that genre theory describes "the ways in which we mobilize language - how out of all the things we might do with language, each culture chooses just a few, and enacts them over and over again - slowly adding to the repertoire as needs arise, and slowly dropping things that are not much use" (p. 13). A familiarity with the genres of a culture allows people to know how to interact with others appropriately and to be active, contributing members within that culture. Therefore, it becomes critically important to teach these genres to people who are learning a second language (and its culture). Because of the Sydney School's focus on the pedagogy side of genre studies, it has created extensive teaching material based on a genre approach (Hyon, 1996). The approach used is based on three phases: Deconstruction (where the teacher breaks down the given genre and explains its context and linguistic features), Joint Construction (where the teacher works with students to collaboratively write something in the given genre) and Independent Construction (where the student writes a text in the given genre by herself) (Rose, 2008).

Ken Hyland tends to support both (New Rhetoric Studies and Sydney school) sides with his focus on both the sociocultural view and the teaching aspect. His view of genre reflects the sociocultural theory when he states that one of the key principles of genre is that "[w]riting is a social activity" (2007, p. 152). Still further emphasis on the social aspect of genre comes from Coe (1994) who says that "genre epitomizes the significance of approaching reading and writing as social processes in which individuals participate without necessarily being entirely conscious of how social the processes are" (p.159). Writing as a social activity or process can mean that writers collaborate or

interact during the process of writing but it also includes the sole writer working alone. This is because the lone writer still draws on previous social interactions and his or her social awareness when creating a piece of writing within a genre. Even in isolation, the writer is still responding to what someone else has said (Freedman & Medway, 1994). The social context of writing also includes the reader since a good writer anticipates how the reader will respond (K. Hyland, 2004). As Freedman and Medway put it, "[w]riting is social in its essence" (1994, p. 5).

The discussion of genre in the context of teaching writing is important because as Ken Hyland (2007) says "genre approaches provide an effective writing pedagogy" (p. 149). He summarizes seven key advantages saying that genre-based writing pedagogy is explicit (it makes clear what is to be learnt), systematic (provides a framework for language and context), needs-based (can ensure that it is based on student needs), supportive (lets teachers provide the scaffolding for students), empowering (provides students with possibilities for expression), critical (provides students with the ability to challenge discourses), and consciousness-raising (increases teachers awareness of genres allowing them to help students more). In an earlier book on the subject, Ken Hyland (2004) states that "genre pedagogies have emerged in L2 writing classes as a response to process pedagogies, as an outcome of communicative methods, and in consequence of our growing understanding of literacy" (p. 7). Swales (1990) adds his voice to the idea of genre teaching by saying that an approach using genre in the classroom is a "workable way of making sense of the myriad communicative events that occur" (p. 1). One extra note to add about genre teaching is that often the knowledge of a genre may exist in the learner's first language and so it is easy for them to transfer this knowledge to their

writing in a second language (Gentil, 2011). Of course, if the knowledge of the genre does not exist in the learner's first language, then this suggests that there is a need to learn the genre and raises the question of whether explicit instruction could help in this regard. Examination of the Criterion program in the context of genre theory allows us to look at the genre(s) that the program is implicitly teaching as well as the question of whether it is explicitly using any kind of genre pedagogy.

2.1.3 Attitude and Motivation

The genre approach to pedagogy may improve learning outcomes, but there are many other factors to consider when viewing language learning from a sociocultural perspective. It is important to understand some of these factors — such as attitude and motivation — that are related to outcomes in language learning and which can be discovered by asking students what their perceptions are about a program such as Criterion. Gardner and Lambert's (1972) sociopsychological theory of second language learning states that there are two types of motivation — integrative and instrumental. He defines integrative as being the desire to fit into the target language culture and community and instrumental as the desire for the practical advantages of knowing the target language. If an ESL student's object (goal) is to get a job, then they are instrumentally motivated, while if their object is to become part of the wider Canadian culture, they are integratively motivated. Gardner claims that both forms of motivation are important but the integrative one leads to greater success in language learning. Expressed in the terms of sociocultural theory, integrative and instrumental motivations are related to the objects (goals) of mediated action. Although Dornyei (2003) has some different views on motivation, he does point out that Gardner's integrative component of

motivation has been consistently shown in empirical studies to account for a significant portion of overall motivation in learning. Gardner also points out that affective factors like attitude and motivation are just as important for predicting achievement in a second language as language aptitude (Gardner, 1985). In a study of a group of language learning students using some basic CALL tools, Ayres (2002) found positive correlations among three factors: motivation, a positive attitude to CALL, and the belief that CALL is relevant to the students. This does not directly show that these factors contribute to success in language learning; however, it does suggest that the factors are related. If one accepts that higher motivation in general contributes to greater language learning success, and combines this idea with the correlation found by Ayres (2002), then this suggests that a positive attitude about CALL tools is correlated with greater success in language learning.

McDowall and Jackling (2006) showed that positive perceptions of the usefulness of computer assisted learning (in the accounting field) improved performance. In addition, Davis, Bagozzi, Warshaw (1989) showed that effective use of computers is influenced by whether one has previous positive or negative experiences with computers. Finally, Masgoret and Gardner (2003) conducted a meta-analysis of 75 studies looking into the relationships between attitude/motivation and achievement in learning a second language. They concluded that there is a positive correlation between achievement and both motivation and attitudes toward the learning environment. In this study, a key component of that learning environment is Criterion. Therefore, there is value in examining the attitudes of students (as revealed by their perceptions) about using a program such as Criterion so that something about its effectiveness can be inferred.

2.1.4 ESL Teaching Methods

Effectiveness of teaching is a central theme for all teaching methods. To be able to study how well Criterion teaches writing to ESL students, one first needs to look at how research and practices in ESL instruction and then, more specifically, ESL writing instruction. Over the years, many methods have been promoted for teaching ESL. These include such methods as the Grammar Translation Method, the Direct Method, the Audiolingual Method, Total Physical Response and the Natural Approach (Brown, 2007). Brown discusses how all these methods have their merits; however, he points out that they lead to a sense of confusion about which is the preferred one. However, more recently, Kumaravadivelu (1994) has promoted the idea of a post-method world where the methods are no longer the focus. Instead, he lists ten macrostrategies for teaching English. The macrostrategies include ideas like "facilitating negotiated interaction" (providing opportunities for two-way communication), "contextualizing linguistic input" (ensuring that language examples have context), "promoting learner autonomy" (giving control of learning to the learner) and "ensuring social relevance" (awareness of the sociocultural influences). These macrostrategies are meant to apply to teaching of ESL in general, but they are also applicable to how the specific skill of writing is taught. When examining how well the computer program Criterion works, one needs to consider its ability to "facilitate negotiated interaction", "contextualize linguistic input", "promote learner autonomy", and "ensure social relevance".

2.1.5 Teaching of ESL Writing

Most language teaching methods and approaches are usually discussed in the context of teaching all aspects of language (such as reading, writing, speaking and

listening), but since Criterion is focused only on writing, it is necessary to understand in more depth the teaching of the specific skill of writing in an ESL context. Unfortunately, as Silva (1990) has pointed out, "There simply are no comprehensive theories of L2 writing" (p.20). Early pedagogical approaches focused on the product of writing (the spelling, grammar, structure, meaning, etc. of a completed paragraph or essay) (Silva, 1990), but more recent approaches focus more on the process of writing (the steps of writing such as researching, outlining, revising, etc.) (Harmer, 2004). As various authors have said, there is a much greater depth to the idea of writing than just product and process. Writing includes a reflection and reinforcement of power, identity, and privilege. According to Ricento (2002), identity is closely linked with language. Giles and Johnson (1987) support this idea saying that language is a prominent marker of identity. Effective teaching must therefore take into account the much deeper and complex aspects of what learning to write in a new language means. Genre pedagogy as described by various authors (K. Hyland, 2004; 2007; Swales, 1990) might be a approach for ensuring that writing as a social and mediated action is connected to the real sociocultural world.

One important suggestion put forward for effective teaching of writing to ESL students is to integrate it with other skills. Kumaravadivelu (1994) has emphasized the importance of integrating the different language skills. In addition, one of Kumaravadivelu's (1994) proposed macrostrategies for teaching a second language is "Contextualized Linguistic Input" (p.32) which he defines as providing input for the student that is an integration of syntactic, semantic, pragmatic, and discourse phenomena, not as isolated parts. Not only should all the language skills be integrated, but the entire social context should be integrated, as made clear by sociocultural theory ideas. One

illustration of the integration of reading and writing skills is described in a study by Weigle (2004) that compared traditional writing tests in which students responded to a given topic to writing tests that were integrated with reading passages to which students responded with a short essay. The study showed that integrating reading and writing in a competency test improved the results markedly. Having passages to read before the students were expected to write allowed them to activate their knowledge and provided something that they could respond to. As well, this is a more realistic way of writing since one does not often write something (an essay, a letter, etc.) in isolation from any other context. This is also supported by the Sydney School's three-phase (deconstruction, joint construction and independent construction) approach to genre pedagogy (Rose, 2008) since that also involves reading in the process of writing. Similarly, the sociocultural theory reminds us that the mediation that creates learning strategies is embedded in a real world context (Donato & McCormick, 1994). The act of writing in the real world is normally embedded in a context to which other language skills are closely related. For Criterion, one needs to consider how well its teaching of the writing skill is integrated into the overall teaching of all language skills but also how it integrates into the overall sociocultural context. Two other important factors in the teaching of ESL writing are providing feedback to students and evaluation of student writing. The next two sections are devoted to these topics since offering feedback and providing evaluation are precisely what the Criterion program does.

2.1.6 ESL Writing Feedback

Feedback is a key technique used in the teaching of ESL writing. According to Leki (2000), "[w]ithin the L2 writing classroom, arguably the most potent aspect of the

context for learning is the writing teacher's response to the students' writing" (p.104).

Feedback can come in many different forms; it can come from teachers or peers; it can be given orally, written and can even come from automated systems (K. Hyland & F. Hyland, 2006). Hyland and Hyland point out that teachers make numerous decisions about the kind of feedback that they provide by considering the students' preferences for feedback, their level of proficiency, the content of their writing, and whether the student has done a first, second, or third draft. A great deal of the literature about how to teach writing is all about how to respond appropriately to students' work (Harmer, 2004; Brown, 2007; K. Hyland, 2003; White, 1985). In addition, the immediacy of feedback is important because, as Scheeler and Lee argue "If feedback is delayed, it allows learners to practice errors" (2002, p. 232). Therefore, judgement of Criterion needs to be based on whether it can provide a response that is customized to the individual students and whether it can respond immediately.

There has been a great deal of research about how students respond to feedback (by making appropriate corrections, etc.), but a more difficult question is whether they learn anything from the feedback (and therefore whether there is any value in giving feedback). The question is difficult to definitively answer because a true longitudinal study about whether feedback improves students' actual learning over the long term is extremely difficult to undertake (although some authors have recommended such studies; see e.g., K. Hyland and F. Hyland, 2006; Kern, 2006). Intuitively, it would seem that providing constructive criticism should help students to learn. However, in a seminal article, Truscott (1996) argues that grammar correction in writing classes should be abandoned altogether. This article generated a great deal of response (Ferris, 1999;

Truscott, 1999). It should be noted that Truscott did not advocate the elimination of feedback — just feedback about grammar.

In a later article, Truscott and Hsu (2008) showed that students were able to immediately improve their writing on an assignment based on feedback they received (compared to students who received no feedback); however, one week later, there was no difference in error rates between students who had received feedback and those who had not. Part of Truscott's argument is that students usually learn grammar in a specific sequence. If, through feedback, a grammar point is being given that is out of sequence, then the student is unlikely to remember it. This idea echoes concepts from Vygotsky's Zone of Proximal Development which states that students should be given help (scaffolding) just at the level they need it (Wertsch, 1991). In the same vein, Aljaafreh and Lantolf (1994) looked into what kind of scaffolding is most effective and when it is most effective. In general they found that the learning must occur only at the student's level and that the learner has to eventually appropriate control for self-correction. This requires a knowledge of what level of expertise and knowledge each individual learner has already attained and a negotiation between that learner and the expert who is providing the scaffolding. Once again, consideration of whether this is possible with a program such as Criterion is needed.

One concern related to providing feedback to students is the ownership of the writing task itself. In a report about a study of two students and how they managed the feedback that their teachers provided them, Fiona Hyland (2000) warns of the dangers of the students relinquishing control over their own writing. She suggests that teachers should encourage their students to maintain control over their own work and that teachers

should be careful about appropriating the student's writing. Brannon and Knoblauch (1982) also warn about the problem of appropriation of the students' writing by the teachers. Reid (1994) refutes this position by saying that teachers should "use their roles as writing experts and cultural informants to empower students in their writing" (p. 273). She echoes sociocultural theory by saying that the appropriation role of the teachers is desirable because writing is a social construct and the teachers are the cultural informants. A logical extension is then that teachers are genre informants as well. For this study, the possibility of Criterion's appropriation of students' work needs to be considered.

Feedback has long been considered by psychologists as a key way to intervene in and reinforce the learning process (Kulhavy & Stock, 1989). Therefore, it is worthwhile to consider the psychological aspects of feedback. One of the key things to note about providing feedback to students is that they are all individuals. A study by Fiona Hyland (2003) of six students and how they responded to feedback illustrated the individuality of students. When receiving feedback, one student was willing to risk trying more complicated forms while another always used the safe and tried forms that he was familiar with. In this case, the teacher needs to be aware of the fact that negative feedback can push students away from trying more complex forms in their writing. Another study by Fiona Hyland (1998) showed that different students will react differently to the feedback they receive and different students will have different expectations about the feedback that they should receive. In addition, Fiona Hyland's (2003) study discovered that all the students believed that by continuously receiving form-focused feedback, they would be able to correct these errors and improve their writing. Even if this belief is

wrong, it cannot be ignored since the belief itself can have an effect on the writing development.

2.1.7 ESL Writing Evaluation

The Criterion program on which this thesis is focused provides feedback but also an evaluation so it is worthwhile examining what evaluation of writing entails. The question is then about which aspects of a piece of writing are relevant to evaluation. Which aspects to evaluate depends on the purpose and genre of the writing. A piece of writing could be evaluated based on its meaning or its grammar or its spelling but the level of importance to assign to each of these is difficult to agree on. Although there have been a very large number of characteristics that have been proposed as indicative of the level of language proficiency, the only characteristic which might predict the level of proficiency is accurate article use (Cumming & Mellow, 1996). One successful way of looking at ESL writers' texts is the rating scheme developed and validated by Hamp-Lyons (1991). This rating scheme defines multiple traits of second language writing: Communicative Quality, Organization, Argumentation, Language Accuracy, and Language Appropriacy. Similarly, Shermis, Koch, Page, Keith, and Harrington (2002) list five traits that come out of the rating of essays in general (not just ESL essays): content, organization, style, mechanics, and creativity. Obviously, no rating system is perfect, but for the purposes of this study, consideration of which aspects of a piece of writing the Criterion program considers when it performs its evaluation is needed.

The computer has become a useful research tool for measuring certain features of writing proficiency. Grant and Ginther (2000) describe using computer programs to tag various features of writing texts and trying to correlate these tagged features with the

quality of the writing. Interestingly, one of the features that statistically is very good at predicting the quality of the writing is simply the length of the essay. Another good predictor is the average word length. It should be noted that these may be good predictors of quality in a given essay; however, it is easy to write a very long essay with lots of long words without writing a good essay. There are many other problems associated with the practice of tagging various features in student essays (Ferris, 1993) — both manually and by computer. For one, these systems assume that the material is grammatically correct in the first place. Many of the systems created were designed for L1 writers — not L2 writers (Bley-Vroman, 1983). Also, the measurement of how often a tagged feature is used may be useful; but the appropriateness of a feature is not being considered. An essay written with lots of complex features used in a completely inappropriate and meaningless way is not a good essay. Finally, a measure of the number of tagged features does not measure what was not included in an essay (Grant & Ginther, 2000). Often, lack of content is the key issue in determining that an essay is poor.

I have illustrated some of the problems related to computer tagging of features, but, even independent of the use of computers, there are also problems related to evaluation in general. Of course, evaluations are constrained by the limits of time and financial resources, so there is always compromise in how one evaluates. Most writing evaluation tends to be done on short essays about fixed topics. Many have questioned how success at this very specific task can be representative of true writing proficiency (Weigle, 2004). Among the problems with this is the time constraint imposed on students to write, the topic about which they must write and even the imposed genre that they must follow. In particular, the topic imposed on the student may cause serious problems

if that student does not have the same cultural background as the person who defined the topic. Real-life writing tasks are rarely like this. Weigle (2004) provides an example of one poor topic prompt used: "Do you like shopping? Why or why not?" (p.32). This tended to result in poor essays since students often have nothing to say about the topic.

Yet another problem is the overemphasis on evaluation itself. According to Cumming (2002), almost all ESL writing tests are for placement/admission purposes which are high-stakes and high-pressure. Very few tests are intended to measure achievement. Related to this is an article by Cope, Kalantzis, McCarthy, Vojak, & Kline (2011) discussing the issue of formative assessments (assessments that are intended to direct the student toward improvement in the final version) versus summative assessments (assessments that evaluate how well the student has done) in the context of computerized assessments. Cope et al. suggest that there needs to be more focus on formative assessments. The authors observe that there is an unhealthy tendency to put a great deal of energy into ensuring that students pass tests instead of actually learn. The authors recommend several principles of meaningful assessment including that it should "be situated in a knowledge-making practice" (p. 81). They also go on to suggest areas of technology that can help with this such as Natural Language Analytics (another term for Natural Language Processing by Computers). Whether Criterion is providing evaluation in a way that directs the student toward improvement and whether it is situated in a knowledge-making practice needs to be considered.

In the spirit of sociocultural theory, consideration of the evaluation of a piece of writing should also consider the context of the evaluation. That context includes the evaluator. An interesting paper by Attali (2011) studied the effects of sequence on human

raters (the people who mark essays). Attali showed that raters who see a very bad paper are more likely to give the next paper a higher mark. This is an important finding since it implies that the rating in high-stakes tests is not fair. Aside from the sequence bias, Atkinson (2003) describes cultural bias in raters' evaluations. He says that writing evaluation and research about writing tends to be done by a very exclusive cultural group — that of white, male, academics from English-speaking countries. Their definition of a good essay is inherently biased. For example, Atkinson points out that academics say that good essays are linear, direct and efficient in their expression of ideas. This is based in the culture of capitalism which values directness and efficiency as illustrated by the common expression "time is money". As well, our culture's emphasis on originality and its denigration of plagiarized content is probably just a reflection of our culture's legal system's focus on private ownership. So, although creativity is considered to be a valuable asset in writing, one should be aware that it is a culturally biased view — some cultures consider the emulation and replication of other respected writers' words as having greater value.

As mentioned earlier, computers have become a useful tool in the tagging of writing texts. The logical extension of this capability is then to use computers to evaluate essays based on that tagging (which is one of the things that Criterion is actually doing). Of course, computers are becoming more and more prevalent in almost all aspects of daily life. Language education is no exception. A brief description of Computer Assisted Language Learning (CALL) and some of the background about the field of CALL is necessary before I discuss Criterion in more detail to help set the context in which Criterion has been created.

2.2 Computer Assisted Language Learning - CALL

The definition of CALL (Computer Assisted Language Learning) is controversial (Kern, 2006). Kern points out that in 1997, CALL was defined as "the search for and study of applications of the computer in language teaching and learning" (Levy, 1997, p. 1 as cited in Kern, 2006). A more modern definition from 2005 is that "CALL means learners learning language in any context with, through, and around computer technologies" (Egbert, 2005, p. 4 as cited in Kern, 2006). In this thesis, I use CALL in the more recent sense as defined by Egbert.

The use of computers in the teaching of English as a second language has become increasingly prevalent. As this study is examining one particular CALL program — Criterion — it is worthwhile to first quickly review the field of CALL itself. In this section, I will provide a brief history of CALL and then a brief discussion of Natural Language Processing (NLP) by computers. I will then focus in on teaching ESL writing through various CALL technologies. .

2.2.1 History of CALL

In 1912, Edward Thorndike described his view of how education might make use of some mechanical automation: "If, by a miracle of mechanical ingenuity, a book could be arranged so that only to him who had done what was directed on page one would page two become visible, and so on, much that now requires personal instruction could be accomplished in print" (as cited in Beatty, 2010, p.95). Even before computers existed, Thorndike was prophetically describing what they might do.

The idea of Computer Assisted Language Learning has been around since the advent of computers. According to Ahmad et al. (1985), the educational potential was

being discussed as far back as the 1940s. Some of the first language learning programs were created during the Cold War to teach Russian to Americans (Beatty, 2010, p.20). One of the first CALL systems was developed at Stanford University in the mid 1960s to teach Russian where students used a teletype machine to enter responses to questions delivered in Russian via a tape recorder (Ahmad et al., 1985). Dartmouth College, in New Hampshire, developed the computer programming language known as BASIC in the late 1960s. The BASIC language was used by language teachers in 1970 to create drill-type programs to teach Danish, French, German, Latin and Spanish (Ahmad et al., 1985). One of the most famous early examples of CALL was called PLATO — which stood for Programmed Logic for Automated Teaching Operations. This system, developed at the University of Illinois, was a general learning system that provided a drill and practice. It was first used for language learning in 1972 to teach students how to translate written Russian to English (Ahmad et al., 1985, p.30).

Since then, the technology has advanced greatly. A very good summary of technological capabilities was written 20 years ago by Nina Garrett (Garrett, 1991). Interestingly, Garrett wrote a follow-up article which stated "An update to my 1991 overview now takes up this [*The Modern Language Journal*] entire volume, with contributions from 18 experts in language technology pedagogy, theory, and research. Their contributions explore topics in computer-assisted language learning (CALL), some of which had not yet been imagined in 1991, and the rest of which have changed out of all recognition" (Garrett, 2009, p. 719). For more detail on the history of CALL, see the special issue (S1, 2009) of *The Modern Language Journal*. More recently, an entire

special issue of the *CALICO journal* was dedicated to Canadian research in the CALL field (Caws, Hamel & Schulze, 2011).

2.2.2 Natural Language Processing by Computers

The teaching of language by computers is closely related to the research field of Natural Language Processing (NLP) by computers. The Criterion program is a key example showing some fruition in the research field of NLP. NLP research looks into the ability of computers to "understand" and use language in a natural way. NLP is really part of the Artificial Intelligence (AI) field of research because the ability to properly use language is considered a hallmark of "intelligence". In 1950, Alan Turing created a definition of computer intelligence. He said that if a computer could interact in a natural language conversation with a human and that human could not tell that it was a computer, the computer would be said to have intelligence (Turing, 1950). In essence, he equated the NLP problem with the AI problem. Arguably, it could be said that the ability to properly communicate in a language requires knowledge of not just grammar, syntax, and vocabulary; but also of extra-linguistic information about context and culture — not just linguistic competence but socio-linguistic competence. Research in NLP is ongoing (Chapelle & Chung, 2010; Chowdhury, 2003); however, with some exceptions, the accomplishments to date have not lived up to the expectations. Despite what Criterion has achieved, it is important to understand the limitations of the field that Criterion comes up against.

Unfortunately, there is no computer system today that can truly pass the "Turing Test" (Abramson, 2011). It follows then that this seriously limits the ability of CALL to effectively teach a language. Kristensson (2009) discusses some of the current NLP needs

and limitations for text entry systems (things such as standard computer keyboard, cell phones, automatic handwriting recognition systems, speech recognition systems, etc.). He describes five challenges that must be met to improve the current state of text entry — localization (functionality in different languages), error correction, editor support (the ability of users to correct errors), feedback, and context of use (typing at a keyboard versus typing on a cell phone on a moving bus). Woven through all of these challenges is the limited ability of computers to manage natural language. As an example, many cell phone text message utilities automatically correct or even auto complete words. Once again, without a proper ability to naturally process language, this can lead to many problems. There are entire web sites devoted to the incorrect text messages that get sent because of this (e.g. <http://www.autocorrectgonebad.com/>). Sociocultural theory provides a glimpse into the difficulty of language processing by computers. Language use is embedded in an extremely complex social and cultural environment which means that its use is still too difficult for computer programs.

2.2.3 CALL and Writing

Given the broader definition of CALL (any learning contexts involving computer technologies) CALL has many uses in the field of teaching L2 writing. Common word processing programs (like Microsoft Word) that facilitate the act of writing can be considered a type of CALL. In today's academic and business environments, the ability to write using a computer's word processing capabilities is probably much more important than the ability to write by hand and so it would seem appropriate to teach students to write using computers. Although some might be concerned about the technical barriers to using computers for writing in a learning environment, there are some studies that seem

to indicate that often this is not an issue. Yoon and Hirvela (2004) report on students using online corpora (large databases of real language usage examples) to help with their writing and they say that, in general, students liked the activity. In a study comparing handwritten essays to essays written using a computer, Lee (2004) reported that habitual computer users preferred doing their work on the computers. One might assume that the increasingly wide usage of computers today is continuing to reduce the technical barriers for language learning students.

In addition to basic word processing programs, a related aspect of CALL for writing is that of Computer Mediated Communication (CMC). This typically involves communication via computer in real time (such as chat rooms) or asynchronously (email and newsgroups). Belcher (1999) describes how there are advantages and disadvantages to using CMC in the classroom. She states that non-native speakers benefit from CMC since it empowers them and democratizes the environment. It does this by removing many of the barriers like shyness and also allows them more time to construct their sentences. At the same time, Belcher points out that since some native speakers might be at a disadvantage with CMC due to their lack of technical abilities, this could level the playing field even more between native and non-native speakers. With an awareness of sociocultural theory and genre theory, CMC can be thought of as a genre that empowers its users by allowing them to express their thoughts and ideas to a greater audience making it a truly "social activity" (K. Hyland, 2007, p. 152). Sheyholislami (2012) goes further and suggests that CMC may enable "marginalized and oppressed minorities to revitalize their languages, reinforce their cultural symbols, and reify their regional and transnational identities" (p. 236).

An extension of CMC for which computers are well-suited is that of collaborative writing. Kost (2011) describes a study of German students who used a Wiki (an online system for collaboratively sharing information) to collaboratively write an essay. She concluded that the environment helped students with collaborative writing and that it also aided their revision behavior. Given how both sociocultural theory and genre theory emphasize the social aspect of writing, it would seem that computer-based collaborative writing is a good technique for teaching writing.

I have now given a brief overview of how CALL has grown and become very prevalent in the area of teaching writing. In the next section, I will look more specifically at computer programs like Criterion that actually perform the writing feedback and evaluation functions involved in teaching writing.

2.3 Criterion and related programs

I will begin this section by describing Automated Writing Evaluation (AWE) programs. This discussion will then lead to a detailed description of the Criterion program. Next, I will discuss some of the literature published about the program. This includes criticisms as well as studies into the validity of the evaluations. After that, I will review some of the literature that asks whether programs like Criterion lead to learning. Finally, I will discuss perceptions of the program and perceptions' importance as the rationale for this study itself.

2.3.1 Automated Writing Evaluation (AWE)

The idea of Automated Writing Evaluation (AWE) has been around for a surprisingly long time. In a groundbreaking paper, Page (1966) described the imminence of computer programs being able to grade essays. He was able to create a simple system

that measured a small number of essay characteristics (such as essay length) to provide a mark which was just as reliable as human raters. Page explained how this could be done without the computer actually understanding the essay. The computer is simply able to measure some characteristics that closely correlate with good essays. Eventually, the College Board — an organization of universities in the United States — funded a program called Project Essay Grade (PEG) which was one of the first commercial systems (Warschauer & Ware, 2006). Another early program was Writer's Workbench (Macdonald, Frase, Gingrich, & Keenan, 1982) which provided detailed textual analysis of writing. One effect of Writer's Workbench is that it pioneered the idea of a computer giving feedback — not just evaluation. Today there are numerous computer programs that do both evaluation and feedback for writing. Three popular ones are MyAccess, Holt Online essay Scoring, and Criterion (Warschauer & Ware, 2006). MyAccess uses an evaluation program called Intellimetric; Holt Online Essay Scoring uses an evaluation program called Intelligent Essay Assessor; and Criterion (the focus of this study) uses an evaluation program called e-rater.

In a study of the ratings for 1,193 essays using the Project Essay Grade (PEG) program, 807 essays were used to train the system (teach the program how to weight each of its measured characteristics) and 386 were used to determine how well PEG then rated the essays (Shermis et al., 2002). The study looked at trait evaluations in addition to overall holistic scores. The traits were defined as content, organization, style, mechanics, and creativity. The study showed that the reliability of both the holistic scores and the trait scores were close to humans. The authors point out that human teachers do not tend to rate these traits in a uniform way. Therefore, not only can AWEs such as PEG provide

overall holistic scores for essays, but they also have the potential to help students with the all-important formative feedback on the traits of the essay.

Another AWE program is called Intellimetric (the rating program used by MyAccess). A study was done of how well Intellimetric does in placing students in the appropriate composition courses (James, 2006). Its ratings were compared with human teacher ratings. The study concluded that the most accurate placements resulted when Intellimetric's ratings were combined with the teacher ratings. Although this study seems to show the validity of the evaluations provided by Intellimetric, one concern is that the validity might not be true for different prompts (i.e. when students are writing essays on different topics). As a result of this concern, another study was done to see if Intellimetric's rating validity was still applicable across different user prompts (James, 2008). It showed that the validity still held. The sample size was not too large (77 essays based on three different prompts) which makes its conclusion somewhat weak. The development of AWE programs like Intellimetric, Intelligent Essay Assessor, and e-rater led to the creation of programs like Criterion which I will describe in the next section.

2.3.2 General description of Criterion

Criterion is an online computer program which provides guidance for and evaluation of essays written by both native English speakers and ESL (English as a Second Language) students (Education Testing Service, 2011a). The program is expressly designed to help students learn writing skills. The Criterion program was developed by the company ETS (Education Testing Services) and is actually composed of two parts — *e-rater* to provide an evaluation of student essays and *Critique* which provides feedback to students (Burstein, Chodorow & Leacock, 2004). The Critique

portion of Criterion provides feedback to the students in five different areas: grammar, usage, style, mechanics, and organization and development. A careful examination of Enright and Quinlan's (2010) description of the features that e-rater looks at for evaluating essays reveals that a very small percentage of e-rater's work is devoted to evaluating organization and development of the essay (probably because of the difficulty involved). This is an important issue to consider when looking at the perceptions of Criterion. ETS makes various claims about the validity of the rating portion of Criterion; however, the company does make it clear that Criterion, by itself, is not sufficient to teach writing. They point out that Criterion is "intended to be used under the instruction of a classroom teacher" (Burstein et al., 2004, p. 27). See Appendix A - Detailed Description of Criterion for details on how Criterion works.

2.3.3 Criticisms of Criterion and similar programs

There are various criticisms of Criterion and other similar computer-based evaluations of written work. One of the key issues is the focus of form over content. Criterion's e-rater and other similar programs do not understand the meaning of any of the essays it evaluates — they are simply performing a statistical analysis to help predict how a human is likely to mark the essay. As a result, it is not truly looking at the content, it is looking at the forms. In a study of students using the MyAccess program, Fang (2010) reported that although there are many potential benefits to the system, it caused students to focus more on form than content. Given that MyAccess works in a very similar way to Criterion, it is also possible that Criterion leads the students to a greater focus on form instead of content. Shermis, Shneyderman and Attali (2008) did a study of 1668 essays by students in Grade 6 and Grade 8. Their statistical analysis is probably

open to some interpretation (for example, they define "content" as similarity of vocabulary to proper essays); however, they conclude that the content portion of an essay only accounts for less than 6% of the variance in scores when the essays are evaluated by e-rater. This would strongly suggest that the Criterion program is not largely concerned with content and that this would lead students to focus more on the items it does care about (such as spelling and mechanics).

There are many scholars in the field that also warn about the dangers of using machines for writing evaluation. In an evaluation of the MyAccess program based on case studies of 12 ESL students, Dikli (2010) writes that human teacher feedback tended to be short and more focused while MyAccess feedback tended to be more generalized and sometimes redundant. She recommends that the companies developing software of this nature need to improve the feedback in these areas. In an article analyzing writing assessment software in general, Vojak, Kline, Cope, McCarthey, and Kalantzis (2011) point out that the 2004 Conference on College Composition and Communication (CCCC) concluded with a statement opposing the use of machine scoring of writing. Similar to my discussion earlier of the relevance of sociocultural theory to understanding how programs like Criterion work, the authors draw on Vygotskian and Semiotic theory to describe their key understandings of writing. Their description includes the statement that "[w]riting is a socially situated activity" (p. 98). Vojak et al. suggest that a student working alone with a computer is not a socially situated activity so they find that automatic writing assessment software comes up short. This contradicts Freedman and Medway's earlier claim that writing is "social in its essence" (1994, p. 5). Similar to

Vojak et al., Xi (2010) urges caution when using such systems — especially in high-stakes evaluations.

2.3.4 Validity

A great deal of research has gone into the question of the validity of the e-rater portion of the Criterion software. The program evaluation is considered valid if it is close enough to the evaluations of trained human raters. A comprehensive analysis by Attali and Burstein (2006) of 25,000 essays looked at human rating versus e-rater rating. It used essays submitted through Criterion (about 7,000 essays) for grades 6 through grade 12, GMAT (Graduate Management Admission Test) essay tests (about 12,000) and TOEFL tests (about 6,000). It concludes that they have proven e-rater's reliability in terms of "single-essay agreement results with human scores, correlations between scores on different prompts, and descriptions of the scoring process and how it contributes to the validity of the system" (p. 23). Attali and Powers (2009) also did a study of 12,000 essays looking into the comparability of scores from different essay prompts. They were trying to determine if e-rater could properly manage essay scores based on different grade levels and different prompts. They concluded that e-rater could do this and once again showed that its validity is as good or better than human raters. Finally, Weigle (2010) conducted a study to address the issue of validity of the e-rater scores as compared to other external indicators of a students' writing proficiency. This was done by comparing e-rater's scores with the students' own self-assessment, their teachers' assessment and other writing samples. The study looked at 386 ESL students over a 15 month period, and it concluded that there were moderate correlations between e-rater's scores and the other external indicators. This conclusion provides further support for e-rater's validity.

One concern about the validity of marks provided by e-rater might be that since e-rater does not actually understand the meaning of essays, it might be possible to trick the program and thereby receive an incorrect mark. Powers, Burstein, Chodorow, Fowles, and Kukich (2002) report on a challenge issued to various writing experts to try to trick e-rater into giving them an incorrect (too high or too low) mark. The challenge produced some interesting ways to trick the program (e.g. repeating the same paragraph 37 times). The report also provides suggestions for how to fix e-rater (and other automatic essay scoring systems) for the future. It is probably safe to assume that the program was fixed to catch this kind of thing since four years later, Higgins, Burstein and Attali (2006) report on the ways that Criterion is able to catch off-topic or bad faith essays. It appears that there is some convincing evidence for the validity of the evaluations provided by programs like e-rater.

2.3.5 Do these programs improve writing?

Another question that many reported studies try to address is whether programs such as Criterion actually help students to learn. It turns out that it is a very difficult question to answer. Most studies about these programs look at the validity of their marks or whether students responded correctly to the feedback (in the short term). It is much more difficult to determine if there has been a long-term effect on the students' learning. In an article by Xi (2010) introducing the July special issue of Language Testing, he points out that there is a need to know if the feedback from programs like Criterion leads to learning but also if the automated writing evaluation systems (like e-rater) lead to learning. As Warschauer and Ware (2006) point out, "[r]esearch to date on AWE has largely been funded by the companies that have produced commercial AWE products and

carried out by the staff of these companies" (p. 163) and "[r]esearch [on these products] conducted to date should be considered with a highly critical eye" (p.163).

Burstein, Chodorow and Leacock (2004) report on a study of 1072 students — half of whom used Criterion over a period of 20 weeks (most did about 4-5 essays). The control group wrote essays in traditional classroom settings. At the end, all the students wrote the FCAT (Florida Comprehensive Assessment Test). No significant differences were found. The authors put a positive spin on it by saying "Though there were no significant differences in FCAT writing assessment scores between the treatment and control groups, the study indicated that students who used Criterion had significant increases in writing production, scored significantly better on later prompts, and significantly reduced a number of writing errors that are tracked by the Criterion software" (p.34). There is no indication as to whether the control group made the same improvements. In a longitudinal study of 401 students using Criterion, it was found that students achieved the "expected gains" in writing skills over a one year period (Attali & Powers, 2009, p. 978). However, this is not a comparison to students not using the program, but is more a test of Criterion's ability to provide consistent evaluations between different prompts used at different levels of study. Another study reported on students' attempts to learn proper use of articles and prepositions (a common difficulty for ESL students) (Chodorow, Gamon, & Tetreault, 2010). In the study, there were 463 students who used Criterion. The article claimed that students improved (they made fewer errors with articles); however, this study did not compare to how well they would have done in a traditional classroom setting. In yet another study of 1464 students, those students who used Criterion several times a week over a 4-week period received higher

scores at the end of the period on their essays than students who did not (Rock, 2007). The effect was considered small but statistically significant (at $p < 0.05$). It is hard to generalize this result too much since the period was relatively short. In addition, it seems that students were writing a lot of essays with Criterion. It is not clear if the control group was writing the same number of essays without Criterion. The effect could simply be attributed to getting more practice writing — not the Criterion program itself.

Finally, a study was done on an automatic feedback system called IADE — Intelligent Academic Discourse Evaluator (Cotos, 2011). In a very cautious conclusion, the author states that the program's "feedback possesses *potential* for facilitating language learning, a claim supported by evidence of focus on discourse form, noticing of negative evidence, improved rhetorical quality of writing, and increased learning gains" [my italics] (p.1).

For all of these studies discussed above which look into whether Criterion leads to actual learning, it should be noted that at least one author from each study is employed by ETS — the company that developed Criterion. As a result, as Warschauer and Ware (2006) have pointed out, they should be considered with a critical eye. Overall, it seems that the evidence supporting the idea that Criterion helps students learn is therefore somewhat weak and may be biased which points to a potential research gap. Creating an impartial longitudinal study that determines whether ESL students learn to write from Criterion would require a great deal of time and resources outside the range and scope of this study. However, it is somewhat easier to study perceptions of the program.

2.3.6 Perceptions

The literature about Criterion and related programs is almost exclusively focused on validity and whether the programs evaluate as well as humans. One of the few studies that looked at perceptions of users was done by Fang (2010). Fang looked at student perceptions of the MyAccess program (a program very similar to Criterion) with a small sample size of 45 students. Students generally recommended the program for its feedback although they did not like it as much for its marking (Fang, 2010). There seem to be some weaknesses in some of the questions used in the study. For example, one of the questions to which students had to respond was "I will read the computer feedback and revise my essays after using MyTutor of MyAccess" (p. 251). It is possible that students answered affirmatively simply because they would be expected to. Such an answer would not necessarily be indicative of their attitudes toward the program.

This study will explore issues about the program by studying the perceptions that both students and teachers have about it. As discussed earlier there is a correlation between attitudes that students have and their achievements in language learning (Masgoret & Gardner, 2003). An exploration of perceptions can reveal the underlying attitudes and motivations. As Gardner (1985) says, "[a]ttitudes influence the success with which another language is acquired" (p. 4). Therefore, it may be possible to make some inferences about language learning based on the perceptions of the students about the program. As sociocultural theory explains, perceptions are coloured by many aspects of experience and environment. These include some of the factors that this study has gathered from participating students such as age, gender, level of acculturation and computer anxiety. In addition, one aspect of any computer application that affects

perception is that of practicality. Jamieson, Chapelle, and Preiss (2004) report that CALL applications need to be evaluated based on a number of factors including practicality. If the program is to be useful, it has to be useful within the constraints (such as money, time and people) of the environment in which it is used.

One aspect of experience and environment is that of age itself. The idea of age being a factor for language learning has often been raised. Bialystok (1997) discusses this idea but concludes that it is not the physical age that makes a difference, rather it is the societal and experiential factors that tend to favour younger learners. This also makes sense from a sociocultural theory perspective. Abrahamsson and Hyltenstam (2009) examined the idea that achieving native-like proficiency in a second language might be constrained by the physical maturity of a person's brain. They concluded that adult learners never truly do and even children rarely do achieve native-like proficiency. Finally, Arecco and Ransdell (2002) speculate that earlier exposure to language learning correlates with improved L2 writing ability. Another aspect of experience and environment is the learner's level of acculturation. Students in this study have been in Canada for varying numbers of years and this may colour their perceptions of Criterion. Fox, Cheng, Berman, Song, and Myles (2006) have stated that acculturation into both the society and the academic environment is a key factor in language learning success. The factor of gender may also be relevant to perceptions. Chu (2010) suggests that women's perceptions of the usefulness of the internet are less positive than men's, while Todman (2000) states that women's levels of computer anxiety were higher than men and the difference between men and women was increasing. Computer anxiety is relevant to the perceptions of a computer program like Criterion. Conti-Ramsden, Durkin and Walker

(2010) also link the levels of computer anxiety with gender so it is worth gathering information about both of these to see how they affect the perceptions. As Conti-Ramsden et al. state, "[i]ndividuals who are anxious about computers are less likely to make use of them" (p. 136). Age, gender, level of acculturation and computer anxiety can therefore all have a bearing on students' and teachers' perceptions about a program such as Criterion.

2.4 Summary

In this chapter, I have described sociocultural theory, genre theory, attitudes and motivation, ESL writing pedagogy, and CALL leading to Criterion and perceptions of Criterion. I described sociocultural theory as a way of looking at language learning through consideration of its social, cultural and historical contexts (Donato & McCormick, 1994). Therefore, in studying students who are attempting to learn English, one must consider all the environmental factors such as the classroom, the college, and the culture of the country in which the learning is taking place. Within sociocultural theory, Criterion can be viewed as a mediator that affects students' efforts to learn English writing. Genre theory describes writing as "typical ways of engaging rhetorically with recurring situations" (Freedman & Medway, 1994, p. 2). Genre theorists have conflicting views about whether genre can or even should be taught. I have emphasized that attitudes and motivation are very important to consider when studying the use of Criterion because Gardner points out that affective factors like attitude and motivation are just as important for predicting achievement in a second language as language aptitude (Gardner, 1985). I have mentioned that although there is no agreed upon theory to describe the process of ESL writing (Silva, 1990), and there have been numerous

methods suggested for teaching English, Criterion's ability to “facilitate negotiated interaction”, “contextualize linguistic input”, “promote learner autonomy”, and “ensure social relevance” as promoted by Kumaravadivelu (1994) should be considered. I have stated how feedback is an important aspect of teaching writing and that Hyland and Hyland (2006) suggest the teachers customize their feedback to students based on many factors. One consideration of judging Criterion needs to be whether it can provide a response that is customized to the individual students. I have also pointed out that evaluation of writing must take into consideration many factors such as knowledge of what aspect of writing is most important (meaning, grammar, spelling, etc.), the appropriacy of evaluations based on single short essays on specific topics (Weigle, 2004), the overemphasis on evaluation itself (Cumming, 2002), and the potential biases of those who mark essays. Our examination of Criterion must be done while asking whether it can overcome these difficulties. I have provided a brief background on the field of CALL and NLP as an introduction to a description of Criterion itself. Finally, I have concluded with a discussion of the importance of perceptions in learning about Criterion's usefulness as well as a discussion of some of the factors (age, gender, level of acculturation, computer anxiety and practicality) that can affect these perceptions.

Chapter 3: Methodology

The aim of this chapter is to provide a description and rationale for the research methodology used in this study. I will first explain my research question. Next I will discuss ethical considerations of the study followed by a description of the participants, the data collection, and the data analysis. This study uses a mixed method approach for data collection and data analysis — there is both quantitative and qualitative data.

3.1 Research Question

As explained in the literature review, the field of CALL is growing and the capabilities of computers are becoming more and more sophisticated. Although there is no definitive correct way to teach writing to ESL students, my study has suggested that sociocultural theory and genre theory help to frame research into the topic. This is because writing is commonly understood as a social process (K. Hyland, 2007). In the context of this study of the Criterion computer program, sociocultural theory helps us to consider the actions of students as being mediated by Criterion and connected to all of the social and cultural influences around the learning process (Donato & McCormick, 1994). Genre theory extends the idea of sociocultural theory by saying that writing is a cultured response to recurring situations and that writing has a rhetorical purpose (Freedman & Medway, 1994).

As discussed in the literature review, a longitudinal study trying to determine the effectiveness of the program is beyond the scope and range of this study. However, an examination of the perceptions of both students and teachers who have used the Criterion program is valuable since a great deal of useful information can be gathered in this way.

The perceptions of students and teachers are a window into all the social and cultural aspects of the environment which are necessary to understand in evaluating a program such as Criterion.

The research question is therefore this: What are the perceptions of both students and teachers about the Criterion computer program? By probing into the perceptions of students and teachers, the study aims to better understand how Criterion is actually being used in an authentic environment. This will help in also trying to answer the question: How well does Criterion adhere to current pedagogical theories and practices?

3.2 Ethical Considerations

The students questioned in this study were all ESL students. Due to students' disadvantage with both the English language and Canadian culture, there is a very small risk that students may not be fully aware of the implications of their participation in the study. To mitigate these risks, a number of steps were taken. An in-class explanation of the study was done for all students. All students who participated were given a consent form to sign which was carefully explained to them. The written questionnaires were all filled in anonymously. All students names for the one-on-one interviews are kept confidential — this study will only refer to pseudonyms. To avoid identifying the interviewees' gender which might also compromise their anonymity, and to avoid the cumbersome use of he/she and his/her, all teachers are arbitrarily referred to as male and all students are arbitrarily referred to as female. I conducted all the interviews myself and I am a certified ESL teacher and am therefore aware of many of the cultural sensitivities and vulnerabilities of students. Since I myself conducted the interviews, I was able to be as careful as possible to protect the students. It should also be noted that I have done

some teaching at the college where the study took place so there was a possibility that some of the students in the study could have been my former students; however, this did not seem to be the case. This study was approved by both the Carleton University Research Ethics Board and the Applied Research Ethics Approval Board of the college where the study took place.

3.3 Participants

The Criterion program is being used for ESL writing classes at a local college in Ottawa. This is a very large college with enrolment of about 16,000 offering numerous post-secondary programs ranging from nursing to computer science to construction. In addition, it offers extensive ESL programs with approximately 500 students enrolled in the various levels. The study was conducted with all six ESL writing classes that were using the Criterion program at the college during the October 24 to December 16, 2011 session. These classes were at Levels 7 and 8 which correspond to the Canadian Language Benchmarks 7 and 8 (as defined on the Centre for Canadian Language Benchmarks website at <http://www.language.ca>). Students at this level have a relatively good grasp of English - many of them intend to go on to post-secondary studies after completing Level 8. In general, all students in these writing classes had to write at least two essays using Criterion although the timing, and use of Criterion's evaluation varied from class to class. All students were asked to fill in a written questionnaire — a total of 73 students. A request was made for volunteer students to submit to one-on-one interviews — six students participated. In addition, interviews were conducted with all six teachers teaching the courses.

3.4 Data Collection

During late November and December of 2011 (towards the end of the term), short information sessions about this research were provided to five out of the six classes. Students then signed the consent form, and then filled in the anonymous written questionnaire. At this point, a request was made for extra students to volunteer to be interviewed one-on-one. Six student volunteered and were later interviewed. For the sixth class, the teacher indicated that there was not sufficient time for the class to be interrupted, so she explained the research to the students herself and had them fill in the consent form and questionnaires on their own time. During this same period, audio-recorded interviews were conducted with all six teachers. The one-on-one interviews with the students were conducted in the college's computer lab and were audio recorded. In addition, a screen-capture program (<http://screencast-o-matic.com/>) was used which recorded the actions on the computer screen as the students discussed the program.

In summary, there were three different sets of data captured:

1. Written Questionnaires (See Appendix C - Student Written Questionnaire) -

The written questionnaire was used to determine some basic background information about the students (age, course level, etc.) as well as Likert-scale information about their perceptions of the various aspects of the Criterion program. The main purpose of this questionnaire was to get an idea of the students' general perceptions of various aspects of the program. A secondary purpose was to see if there were any relationships between the various background factors and the Likert-scale ratings of how well they liked the program.

2. Student Interviews (See Appendix D - Student Interview Questions) - Semi-structured oral interviews were conducted with six students to determine more details about their perceptions of the program. The student interviews were audio-recorded and the actions on the computer screen were also recorded. Conducting the interviews while the student was using the Criterion program facilitated easier explanation and discussion of the program features. During the interview with the sixth student (student F), it was revealed that the student had not yet completed an essay using Criterion although he/she had seen the output for other students. As a result, a second interview was conducted the next week with this same student after he/she had completed an essay using Criterion.

3. Teacher interviews (See Appendix E - Teacher Interview Questions) - Semi-structured oral interviews were conducted with the six teachers involved in the program. Audio recordings were made of the interviews. The purpose of these interviews was to determine the teachers' perspective on the program and to extract as many insights about the utility of Criterion as a tool for teaching writing.

3.4 Data Analysis Method

This study used a mixed-method approach — with quantitative questionnaire data and qualitative interview data. Therefore, the majority of the focus will be on the qualitative data. Wherever possible, attempts will be made to triangulate the results from the two different types of data. I will first describe the quantitative analysis and some justification for why this analysis was done. Next, I will describe the theory and steps followed for the qualitative analysis.

3.4.1 Quantitative Analysis

A quantitative analysis of the written student questionnaire data was conducted. First, the data was encoded into a Microsoft Excel spreadsheet. This created 73 rows (one for each student) of data. One of the questions in the questionnaire asked how many essays the student had submitted to Criterion. Five of the students answered zero and therefore these students' responses were treated as invalid reducing the number to 68. A summary of the written questionnaire data is shown in the results section.

The data was transferred to the SPSS (Statistical Package for the Social Sciences) software package in order to conduct some statistical analyses on the results. First, a Cronbach Alpha test was conducted to ensure the internal consistency of the data. The test was run on the results from Questions 7 through 20 which were all 5-point Likert-scale responses ranging from Strongly Agree to Strongly Disagree indicating how positive students' perceptions of various aspects of the Criterion program were. One would expect that if students have overall positive or overall negative feelings about the program, then their responses would consistently reflect this and therefore show a high Alpha value in this test.

There are many potential statistical tests that could be run to determine if any of the background data responses are related to how students perceive the Criterion program. To check for a simple bivariate correlation (a correlation between two items), a Pearson Correlation test could be run, but this test assumes that the data is interval data (and it is debatable whether the Likert scale is interval or ordinal). Therefore, a more appropriate test to run is the Spearman's Rho test. This was done (using SPSS version 19) comparing the fourteen Likert scale responses with the variables of Age, Gender, ESL

Level, Years in Canada, Number of Hours per day on a computer, and Number of Essays written for Criterion.

3.4.2 Qualitative Analysis

For the qualitative data, all the recorded interviews (both student and teacher) were transcribed into tables in Microsoft Word. Columns were created in the tables to indicate interview number, timestamp, speaker, words spoken, and computer screen observations (the computer screen observations were for the student interviews only). The computer program "Express Scribe" (<http://www.nch.com.au/scribe/index.html>) was used to aid in the transcription of the interviews. An example showing a portion of one of these tables appears in Appendix B - Sample of Transcription Data.

Glaser and Strauss (1967) first proposed their grounded theory approach to qualitative data analysis and suggested what they called the "Constant Comparative Method" (p. 101) for coding and interpreting qualitative data. They pointed out that much previous research with qualitative data was based on the assumption that the researcher was looking for proof of a given hypothesis in the data. They were suggesting an opposite approach in that they felt they could look for theory from the data without having any a priori hypothesis to prove. Since the qualitative data gathered for this research was done without a hypothesis to test, some of their ideas have been used in the approach to data analysis for this thesis. It seems highly unlikely that one could look at data without having at least some preconceived ideas and hypotheses in mind (unless one knows nothing about the subject under study); however, it would seem that an open mind might lead to new discoveries.

Others have built on Glaser and Strauss' idea by defining a more specific and rigorous approach to data coding. To a large extent, this thesis has followed Flick (2009) and Gibbs (2007) who provide a step-by-step approach to the coding of data. A number of studies have used this approach effectively. For example, Murday, Ushida & Chenoweth (2008) studied student and instructor satisfaction with some online language courses at Carnegie Mellon University by interviewing the subjects. Stracke (2007) conducted interviews with three students who had quit a language course that involved a mix of face-to-face learning and some CALL to determine their reasons for leaving. Murray (1999) studied French language students who were using an interactive video CALL program. Finally, Bidlake (2010) studied students using self-instructed CALL programs. The approach of this thesis most closely emulates that of Bidlake which uses three levels (or stages) of coding: Open, Axial and Selective.

To perform the Open coding, a new column in the tables of transcriptions was created (see Appendix B - Sample of Transcription Data). In this column, a basic description of what was happening was entered based on the spoken words as well as the video screen information. The idea of Open coding is to reflectively tease out the conceptual categories. As Stracke (2007) suggests, the coding at this stage should involve the researcher asking what the data is referring to and what the action/interaction is all about. Another way of looking at this data is described by Aronson (1994) who says, in explaining a similar technique, that coding simply involves breaking down patterns of experience.

A list of the conceptual categories (Axial coding) that emerged from this coding was then created. These categories were sorted in a hierarchical manner with up to two

levels of branching. This required several iterations since, as Glaser and Strauss (1967) suggest "while coding an incident for a category, compare it with the previous incidents in the same and different groups coded in the same category" (p.106). The iterations continued until it reached a point where the categories covered all the concepts that emerged from the data and was of a manageable size. Each of these categories was assigned a five-letter code and number. Next, another new column was created in the tables of transcriptions for Axial coding. One or more of the Axial codes were then assigned to each piece of dialogue in the transcriptions (See Appendix B - Sample of Transcription Data).

In order to do the Selective coding, the Microsoft Word tables were imported into Microsoft Excel. There are numerous software tools available to analyse qualitative data, but the important roles that the software needs to perform are database functions, data retrieval functions, searching and sorting (Seror, 2005). Microsoft Excel meets all these requirements. Using Microsoft Excel, all the data was sorted to group all the Axial coding categories together. This helped to determine the categories that occurred most often and also helped to isolate various other salient topics. In this part of the analysis, this study differed from Bidlake (2010) slightly since a count of the most often occurring themes was not the only way used to determine the key themes. This is partly because of a recognition of the biases introduced by the interview questions themselves. For example, there were questions about all five aspects of Criterion's feedback, so they all appear as themes, whether students felt they were relevant or not. To help overcome this potential bias, salient themes were also looked for that may not have occurred often but seemed important based on my understanding of the entire context. As Glaser and Strauss (1967)

point out, "[t]he root sources of all significant theorizing is the sensitive insights of the observer himself (sic)" (p. 251). The selective coding also took place with an appreciation for the context as made clear by sociocultural theory. As Huckin (2004) says in his description of how to interpret findings in qualitative research, "attention to context often yields insight" (p.170). Similarly, Roberts (1989) says "[t]he coder must be familiar with both the context in which a statement is made and the cultural universe within which it was intended to have meaning" (p. 164). I tried to rely on my knowledge and understanding of the sociocultural context of ESL writing as well as the context provided by the literature review chapter of this thesis to help me in this way.

This chapter has summarized the details of the research methodology including the data gathering and the data analysis. In the next chapter, I will describe the results of both the quantitative and qualitative analysis of the data.

Chapter 4: Results and Analysis

The objective of this chapter is to provide a full description of all the results as well as provide a detailed analysis of these results. To do this, I will first present a summary of the quantitative questionnaire data and describe certain key parts of the questionnaire data in more detail. Next, I will describe the statistical tests performed on the questionnaire data and interpret the meaning of these tests. After that, I will present a summary of the qualitative interview data including a description of the coding that was performed on the data. Next, I will list the five key themes as well as student and teacher recommendations which resulted from the coding analysis of that data. Finally, I will provide a more detailed analysis of those five themes and the recommendations by drawing on examples from the interview transcripts themselves.

4.1 Quantitative Results

4.1.1 Description of Data

The questionnaire data is composed of 68 students' responses to a two-page written questionnaire (See Appendix C - Student Written Questionnaire for the questionnaire). The first six questions in the questionnaire were for background information about these students — age, gender, ESL level, number of years in Canada, number of hours spent on a computer per day, and number of essays written with Criterion. The rest of the questions were Likert scale assessments of how positive the students felt about various aspects of the Criterion program.

The first question on the questionnaire asked for the student's age. The question of age is of interest for a number of reasons. As mentioned in the literature review chapter,

the idea of a critical period or younger age being ideal for language learning has often been discussed (Bialystok, 1997; Abrahamsson & Hyltenstam, 2009; Arecco & Ransdell, 2002). Also, it is possible that younger students could be more comfortable with computers and therefore have more positive perceptions of the Criterion program. It is also possible that younger students are more able to learn a language and this could lead to a more positive attitude to their ESL class in general and also a more positive attitude toward the Criterion program. The distribution of ages is shown in Figure 1 and indicates that the students had a wide variety of ages, with about half of them being 25 years old or less. The average age of the students was 28.5 years old and the median age was 24 years old. One could speculate that as people get older, they are less likely to make an intense effort to learn a language and this is why the median age is young. It could be that younger people are more likely to be in a position where they are striving to start their careers in an English environment and therefore recognize the need for language skills. It should be noted that the classes involved in this study (Level 7 and Level 8) are relatively high-performing students — many of whom are likely to go on to post-secondary education programs. In general, students that graduate from Level 8 are ready to attend a regular post-secondary program. Note that six students did not provide their ages.

As shown in Figure 2, 18 students were male and 50 were female. The students were asked to report their gender due to the often discussed idea that men and women have different attitudes and perceptions toward computers (Chu, 2010; Todman, 2000). There are significantly more females than males and the reasons are not obvious. There may be a relationship between the skew in the gender numbers and the specific recruiting

techniques of the college. There simply may be more women than men that enroll in higher level ESL classes. It may simply be just chance.

Figure 1 Age distribution of students

Figure 2 Gender distribution

The majority of the students (47) were in Level 7 and the rest (21) were in Level 8 as shown in Figure 3. It is possible that students who have been enrolled at the college for a longer period of time and who have been exposed to the Criterion program more may be more comfortable using Criterion and so it was worthwhile obtaining this information. Of course, not every student in Level 8 had also attended the Level 7 ESL course at the

college and so may not have been exposed to Criterion before. Based on informal discussions with teachers at the college, there are usually more students in Level 7 than Level 8 for various reasons. The reasons include the fact that some have a sufficient level of English on completing Level 7 so they do not continue. Limited funding is another factor since some students only have funding for a certain time period and so cannot afford to continue at Level 8.

The average time spent in Canada by the students was 3.1 years. The distribution of years spent in Canada is shown in Figure 4. As mentioned in the literature review chapter, Fox et al. (2006) have stated that acculturation into both the society and the academic environment is a key factor in language learning success. The time spent in Canada could be an indirect measurement of the amount of acculturation that each student has undergone. Therefore, it was worth asking this question since students' success in language learning could have a bearing on their perception of the Criterion program.

Figure 3 Distribution of ESL level

The fourth question in the questionnaire asked how many hours per day students spend on a computer on average. Less time spent on a computer is possibly a reflection of

anxiety and whether the student has a positive or negative attitude towards computers. As mentioned in the literature review chapter, Conti-Ramsden et al. (2010) link computer anxiety and likelihood of using computers. They also link computer anxiety to gender so it is worth gathering the gender and time spent on a computer data from the students.

Figure 5 illustrates the distribution of hours spent on the computer and shows a somewhat normal distribution with the amount tailing off at the high end. The average time (3.0 hours per day) seems high but this may simply be a reflection of today's average students. Although I do not have statistics for comparison, it would be reasonable to guess that students today probably spend a lot more time on the computer per day than even five years ago. As a result, this points to the possibility that anxiety with computers may not be an issue at all for the students.

Figure 4 Distribution of time spent in Canada

Students were also asked how many essays they had written using Criterion and Figure 6 shows the distribution of the responses. Note that the questionnaires of students

who answered zero to this question were excluded from the entire study. The average number of essays was 1.8. The number of essays written for almost all students was either 1 or 2. This is somewhat disappointing since it indicates that most students are probably not extremely familiar with the program. It should be noted that the study was constrained by the eight-week session length for the classes. Conducting the questionnaires near the beginning of the session would not have worked because students would not have used the program at all. Conducting the questionnaire in the last two weeks of the session was also inadvisable because there are many final tests being run during that period and the teachers did not want disruptions to their classes.

Figure 5 Distribution of hours spent on a computer per day

The rest of the questionnaire was composed of Likert scale questions to indicate the level to which students agreed or disagreed with the usefulness of various components of the Criterion program. Figure 7 shows a summary of all of this data. The axis along the bottom indicates which question, the axis along the right indicates which answer (from Strongly Disagree to Strongly Agree) and the height of the bars indicates

how many students provided that particular answer. The figure shows that for all but one of the questions, the most common answer was "Agree". The only exception was for the question about whether the college should continue to use the program and, in this case, the "Strongly Agree" answer was the most common. There were no cases of students answering "Strongly Disagree".

Figure 6 Distribution of number of essays written

By assigning a numerical value to each of the Likert scale responses (1 for Strongly Disagree up to 5 for Strongly Agree), the average agreement for each of the Likert scale questions can be determined. These values are shown graphically in Figure 8. All fourteen of the Likert Scale questions had average responses that were favourable of the program. Numerically, the average response to the questions was 3.87 (where Strongly Disagree was 1 and Strongly Agree was 5). The variation in responses between each question was relatively small. Given that the student attitudes are so positive, this would imply that students are more likely to apply themselves to working with the

program as McDowall and Jackling (2006) suggest. I will now analyze the questions which resulted in the highest and lowest responses.

Figure 7 Distribution of student approval of various Criterion components

The question with the lowest response (though still a positive response) was whether students were satisfied with the overall holistic scores that Criterion provides. The numerical value of the response to this question was 3.63 (where Strongly Disagree was 1 and Strongly Agree was 5). This is not too surprising since students normally would prefer higher marks and this would be reflected in a question of this nature. In other words, the responses to the question may not show the level of satisfaction with Criterion's fairness in marking — it may simply mean that students would prefer that they had higher marks.

Figure 8 Average numerical agreement with questions

Three questions (Writer's Handbook, Style Satisfaction, and Organization & Development Satisfaction) tied for the next lowest response at a value of 3.76. The first of these three was regarding the usefulness of the Writer's Handbook utility of the program. This is a utility that students can click on to get more detailed information about various pieces of feedback. For instance, they could find out what subject-verb agreement means or they could find out what the characteristics of a good thesis statement are. It is possible that this response is a reflection of whether the students had ever used the utility. Given that this utility was never mentioned in any of the student interviews, it seems that this is probably the case. The second of those three questions was regarding Criterion's Style category scores. The Style category provides information to students such as repetition of words and run-on sentences. Many students had comments about the "repetition of words" error during interviews. Criterion identifies words that have

occurred many times and suggests that students should use synonyms to add variety. This is often a good suggestion; however sometimes, it is not appropriate as the word in question has no synonyms. An example of this is a word like "telephone". There are other descriptions of telephones but using phrases instead of the actual word can sometimes make an essay worse. As well, an extensive knowledge of synonyms is likely lacking in most ESL students. So this slightly lower response to the question may be a reflection of frustration by some students. The third of those three lower response questions was regarding Criterion's Organization and Development category scores. This part of Criterion analyzes the students' essays and tries to identify the thesis statement, the introductory material, main ideas, supporting ideas and the conclusion. This is difficult to do for humans, so the Criterion program, that does not actually understand the meaning of anything that the students have written, often misidentifies these parts. Unlike the other feedback categories for which Criterion usually provides specific instructions on how to fix them, the feedback provided in this category is, by necessity, very general. As a result, students often do not know how to fix problems in this category with the hope of improving their mark. Their frustration is probably reflected in this slightly lower response for this question.

The question with the highest level of agreement in the responses was regarding the Grammar Feedback which had a numerical value of 4.15. In this category, Criterion provides feedback on Sentence Fragments, Garbled Sentences, Subject-Verb Agreement, Ill-formed Verbs, Pronoun Errors, Possessive Errors, Wrong or Missing Words, and a general category of "Proofread This". With the possible exception of "Proofread This", these are all errors that students can easily fix. Students also tend to focus a great deal on

the grammar when writing (whether the theory says they should or not). The fact that Criterion provides this area of feedback seems to meet their needs quite well and so is reflected in the high level of agreement with the question. The question with the second highest level of agreement was the one asking if the college should continue to use the Criterion program. The numerical value was 4.13. This seems to reflect a very strong preference for doing that. As mentioned earlier, if students like the program, it is likely to lead to improved learning outcomes. It is worth noting here that the qualitative data — to be discussed later in this chapter, also strongly supports the continued use of the program. In addition to this summary of all the quantitative data, I also performed some statistical analysis of the results which I will describe next.

4.1.2 Statistical Tests

As described in the Methodology chapter (see 3.4.1 Quantitative Analysis), statistical tests were performed on the questionnaire data to look for consistency and potential correlations. The two tests used were the Cronbach Alpha test and the Spearman's Rho correlation test. The Cronbach Alpha test was run (using SPSS version 19) to test whether the Likert scale responses were internally consistent. All of the Likert scale questions were asking the students about their level of positive attitudes towards various components of the program. If a student had a certain attitude (positive or negative) toward the program in general, it would be expected that all the responses to questions about the individual components would reflect that in a consistent way and the Cronbach Alpha test would provide an Alpha value close to 1.00. On the other hand, if a student gave responses that were inconsistently positive or negative, then the Cronbach Alpha test would provide a low value (lower than 0.7 is the threshold often used) which

would indicate inconsistent data and a potentially poorly designed questionnaire. The test gave an Alpha value of 0.916 which indicates strong consistency in the data.

To check for simple bivariate correlations (correlations between two items), the Spearman's Rho test was run. This was done (using SPSS version 19) comparing the fourteen Likert scale responses with the variables of age, gender, ESL level, years in Canada, number of hours per day on a computer, and number of essays written for Criterion. The results are shown in Table 1. Each entry in the table corresponds to the correlation coefficient from a Spearman's Rho test between the item indicated in the column heading and the item indicated in the row heading. For example, to find the correlation coefficient between the Likert scale response for "Make a Plan" and the students' ESL Level, find the entry for the second row of numbers (Make a Plan) and the third column (ESL Level) which shows the numerical value of 0.036. The closer that a coefficient is to 1, the greater the level of correlation between the two factors. An asterisk (*) indicates a potential correlation (at $p < 0.05$) so this value shows that, for this case, no correlation has been found. The potential correlations are shaded in the table. All of these correlations are somewhat weak and it should be noted that they are at a level of $p < 0.05$ which means that they could just be the result of chance in 1 out of 20 cases. Since there are 84 correlation tests (14 Likert scale responses times 6 personal data variables), it is not surprising that 5 cases are shown to be significant.

To summarize, there are five potential correlations between:

1. Writer's Handbook and Age (Positive Correlation)
2. Mechanics Feedback and Gender (Positive Correlation, where male is 0 and female is 1)

3. Style Satisfaction and ESL Level (Negative Correlation)
4. Style Feedback and ESL Level (Negative Correlation)
5. Organization/Development and Number of Essays (Negative Correlation)

Table 1 Spearman's Rho Correlation Coefficients

	AGE	GENDER	ESL LEVEL	YEARS in CANADA	COMPUTER HOURS PER DAY	NUMBER of ESSAYS
Overall scores	0.139	-0.012	0.084	0.209	-0.017	-0.124
Make a Plan	0.067	0.116	0.036	0.123	-0.066	-0.125
Writer's Handbook		0.107	-0.166	0.242	-0.032	-0.238
Grammar Satisfaction	-0.101	0.079	-0.077	-0.027	-0.021	0.032
Grammar feedback	-0.066	0.154	0.048	0.016	-0.022	0.009
Usage Satisfaction	-0.02	0.053	-0.021	0.024	-0.033	-0.145
Usage feedback	0.065	0.224	-0.021	0.017	0.032	-0.084
Mechanics Satisfaction	-0.089	0.058	-0.064	-0.132	0.002	0.013
Mechanics feedback	0.05		-0.037	0.096	0.018	-0.066
Style Satisfaction	0.025	-0.006		0.189	0.094	-0.195
Style feedback	-0.01	0.066		0.207	-0.09	-0.212
Org&Dev Satisfaction	0.053	0.05	-0.148	0.176	-0.036	
Org&Dev Feedback	0.075	0.115	-0.064	0.188	-0.001	-0.232
Continue to use	0.006	-0.03	-0.131	0.189	-0.023	-0.205

* - indicates significance at $p < 0.05$

The writer's handbook and age correlation is suggesting that approval of the writer's handbook increases with the age of the participant. This might be because older students appreciate the idea of an old style reference text book (the Writer's Handbook is simply an online grammar text book linked in to the program). It is also possible that older students are more patient with their writing work and therefore are willing to take

the time to study the details of the grammar theory, whereas younger students might be more impatient and simply want to try a quick fix to see if that eliminates a particular error.

The Mechanics feedback and Gender correlation suggests that females are more likely to approve of the Mechanics feedback (females were coded as "1" and males as "0" in SPSS). The mechanics feedback includes items like spelling and punctuation so perhaps females are more concerned with spelling and punctuation than males.

The two negative correlations between ESL Level and both Style Satisfaction and Style Feedback are suggesting that at the higher ESL Level, students are not as happy with the Style category of feedback. There is a particular error provided in the Style feedback category ("repetition of words") which is somewhat problematic (and will be reviewed later in this chapter). One possibility is that the higher level students have been more exposed to this problem and therefore are more likely to have a less favourable attitude to the Style category.

The last correlation to mention is the negative correlation between Organization and Development satisfaction and the Number of Essays written. The value suggests that students who have written more essays feel less satisfaction with the Organization and Development category of feedback. This may be similar to the last two correlations because, as I will describe later in this chapter, there are several important problems related to the Organization and Development category. Initially, students may be impressed with what feedback is provided in this category, but as they get more experience and start to understand in more depth what Criterion is telling them about

their organization and development, they may become a little disappointed and they may feel less positive about it.

4.2 Qualitative Results

The qualitative data was gathered from a total of 13 recorded interviews — six teacher interviews and seven student interviews with six different students. One of the students was interviewed twice since she indicated during the first interview that she had not yet completed an essay with Criterion. For this reason, the student was also interviewed a second time after she had completed an essay with Criterion.

As described in the Methodology chapter, a three stage coding method was used. After the first stage of Open coding where each piece of dialogue was coded to simply describe what was happening, I then developed a list of key themes that had occurred in the various open coding entries. I tried to limit the number of main themes to about fifteen and also specified branches (subthemes) and branches of branches (sub-subthemes). I also tried to make sure that themes did not overlap too much. I then tried to see if this set of themes properly represented all the Open coding entries. After a few iterations, I came up with the final hierarchical structure of Axial coding categories as shown in Table 2. For ease of sorting and tracking purposes, as the table shows, I assigned both five-letter codes and a hierarchical number representation to each theme and its branches. For the second phase of coding, each piece of dialogue was then assigned an Axial code as shown in Table 2.

Table 2 Axial Coding categories

Main Category	First Branch	Second Branch
1.0 Time (TIME)	1.1 Teacher Time	1.1.1 Teacher time configuring Criterion
		1.1.2 Teacher time marking Criterion essays
	1.2 Student time	1.2.1 Student time writing Criterion essays
		1.2.2 Student time writing classroom essays
2.0 Workload (WRKLD)		
3.0 Computer literacy (COMPL)	3.1 Teacher computer literacy	
	3.2 Student computer literacy	
4.0 Ease of use (EASE)	4.1 Teacher ease of use	
	4.2 Student ease of use	4.2.1 Ease of use with outlines
		4.2.2 Ease of use with feedback
5.0 Understanding of program (UNDER)	5.1 Teacher understanding of program	
	5.2 Student understanding of program	5.2.1 Awareness of feedback options
		5.2.2 Awareness of Criterion limitations
6.0 Training (TRAIN)	6.1 Teacher training	
	6.2 Student training	
7.0 Satisfaction (SAT)	7.1 Teacher satisfaction	
	7.2 Student satisfaction	
8.0 Technical (TECH)	8.1 Program errors	
	8.2 Program misleading	
	8.3 Losing essays	
9.0 Validity of marks (MARKS)	9.1 Teacher perceptions of validity of marks	
	9.2 Student perceptions of validity of marks	

Main Category	First Branch	Second Branch
10.0 Validity of feedback (FDBCK)	10.1 Teacher perception of validity of feedback	10.1.1 Grammar
		10.1.2 Usage
		10.1.3 Mechanics
		10.1.4 Style
		10.1.5 Organization and Development
	10.2 Student perception of validity of feedback	10.2.1 Grammar
		10.2.2 Usage
		10.2.3 Mechanics
		10.2.4 Style
		10.2.5 Organization and Development
11.0 Essay topics (TOPIC)	11.1 Teacher assigning topics	
	11.2 Student use of topics	
12.0 Researching for essays (RSRCH)		
13.0 Outlining (OUTLN)	13.1 Criterion's Make a Plan utility	13.1.1 Teacher perceptions of Make a Plan
		13.1.2 Student perceptions of Make a Plan
		13.1.3 Problems caused by outlining
	13.2 Other Outlining tools	
14.0 Five-paragraph essay (FIVEP)		
15.0 Comparisons to Microsoft Word (MSWRD)	15.1 Microsoft's grammar/mechanics corrections	
	15.2 Use of Microsoft for initial draft	
16.0 Recommendations for the future (RECMD)	16.1 Teacher recommendations	
	16.2 Student recommendations	

The meanings for each of the axial coding main themes shown in this table are as follows:

- 1.0 Time (TIME)** - This theme describes the issue of time constraints for both teachers and students. In some cases it was specific to the time writing Criterion essays (students) or configuring Criterion to assign essays (teachers)
- 2.0 Workload (WRKLD)** - This theme describes the occurrence of comments about workload for both students and teachers.
- 3.0 Computer literacy (COMPL)** - This theme describes the levels of computer literacy and how comfortable both teachers and students are with computers.
- 4.0 Ease of use (EASE)** - This theme is used to describe comments about the ease of use or lack of ease of use of the Criterion program.
- 5.0 Understanding of program (UNDER)** - This theme describes the level of understanding about the Criterion program by both students and teachers. Its subthemes include understanding of some specific aspects of the program.
- 6.0 Training (TRAIN)** - This theme is used to describe the topic of training on how to use Criterion for both students and teachers.
- 7.0 Satisfaction (SAT)** - This theme describes comments about the level of general satisfaction with Criterion.
- 8.0 Technical (TECH)** - This theme is used to describe occurrences of technical issues and technical problems related to Criterion.

- 9.0 Validity of marks (MARKS)** - This theme describes comments about the marks that Criterion provides for students and the perceptions of those marks.
- 10.0 Validity of feedback (FDBCK)** - This theme describes comments about the feedback that Criterion provides and the perceptions of that feedback
- 11.0 Essay topics (TOPIC)** - This theme is used to describe comments about the topics used for the Criterion essays.
- 12.0 Researching for essays (RSRCH)** - This theme describes comments about students performing research for their essays before writing them in Criterion.
- 13.0 Outlining (OUTLN)** - This theme describes comments about the Make a Plan utility in Criterion that allows students to create an outline before writing their essay.
- 14.0 Five-paragraph essay (FIVEP)** - This theme describes comments about the five-paragraph essay format that Criterion uses.
- 15.0 Comparisons to Microsoft Word (MSWRD)** - This theme describes comparisons made between the capabilities of Criterion and Microsoft Word.
- 16.0 Recommendations for the future (RECMD)** - This theme describes the various recommendations that both students and teachers made with respect to the Criterion program.

4.2.2 Emergent Themes

For the third and final phase of coding, selective coding was done by counting the number of occurrences of each axial coding theme in the interview data. The key themes that emerged were based on the highest counts for that theme including its subthemes. The four most commonly occurring themes were Usefulness of feedback — especially Organization and Development (204 occurrences for Feedback, 40 for Organization and Development Feedback), Satisfaction with the program (119 occurrences), Understanding of the program (102 occurrences), and Time constraints (82 occurrences). In an attempt to compensate for the bias towards particular themes introduced by the nature of the questions in the interviews, I also tried to extract themes based on what the literature review pointed to. When considering sociocultural theory, and therefore viewing Criterion as a mediational means of language learning, it becomes important to consider technical issues with the program as an important contextual factor that has a bearing on the success of language learning efforts. Therefore, the theme of Technical Issues seems relevant as well. To summarize, the themes that have emerged are:

1. Usefulness of feedback (especially Organization and Development)
2. Satisfaction with the program
3. Understanding of the program
4. Time constraints
5. Technical Issues

These five themes will be discussed in detail in the following five sections.

Although there is a separation between themes, there is also an extensive overlap. These themes emerged due to their prevalence in the interview data; therefore in the elaboration

of those themes, I will draw on the extensive interview records. Since the interview records have been sorted by Microsoft Excel according to the themes (based on the axial coding), it is easy to extract examples relevant to each theme. As mentioned earlier, I will be referring to examples from interviews with six teachers and six students. To protect their anonymity, I will simply refer to them as Teachers 1 through 6 and Students A through F. As mentioned earlier, to avoid identifying their gender which might also compromise their anonymity, and to avoid the cumbersome use of he/she and his/her, I am arbitrarily referring to all teachers as male and all students as female.

4.3 Feedback Usefulness

The theme of feedback came up very often — possibly because it was a topic that many of the interview questions addressed during the interviews. As mentioned earlier, a great deal of the literature about how to teach writing is all about providing feedback (Harmer, 2004; Brown, 2007; K. Hyland, 2003; White, 1985). The Criterion program itself is fundamentally constructed with two functions — providing feedback and evaluation. Therefore, it is quite important to examine how Criterion's feedback is perceived. In general, it can be concluded that the feedback was viewed in a mostly positive light. The quantitative questionnaire data strongly suggest that students like all the different feedback components. All the fourteen Likert scale questions about feedback had positive responses. As well, there were numerous comments from the interview data that corroborate this support. A typical example comes from the interview with Student E, in which the following exchange took place:

[REDACTED]

Student E was happy with the feedback and seemed especially happy about the grammar feedback. The quantitative questionnaire data provided some triangulation to this result since the agreement with the "Grammar feedback" question was the highest of all at an average response of 4.15 (where strongly disagree is 1 and strongly agree is 5). The positive view of the feedback — especially the grammar part — suggests that Criterion is providing what the students want.

In addition, a couple of teachers felt that the grammar feedback was good. Teacher 2 said "I think it does a very good job at evaluating the students' grammar uhh for the most part" (Teacher 2, Interview). Teacher 6 gave somewhat more qualified support saying "If they are strong writers... and they're writing coherent, cohesive essays, then they're working on umm... very specific things such as mechanics and grammar... it [Criterion] is catching some of their errors" (Teacher 6, Interview). In other words, if the essays were coherent and cohesive to start with, then Criterion worked well. The implication is that Criterion does not do a good job for incoherent or non-cohesive essays. Note that, as mentioned in the literature review chapter, the ETS company (the makers of Criterion) have dedicated some effort to ensuring that completely incoherent essays are rejected and not even evaluated by Criterion (Higgins, Burstein & Attali, 2006). It's still possible that mildly incoherent essays would get through the filtering and

this is the kind of situation to which Teacher 6 is referring. So although students seem to have a very positive view of the feedback, the teacher's perceptions are a little more nuanced.

Another key measurement of good feedback is how customized it is. One key aspect of teaching that Hyland and Hyland (2006) refer to is how important it is for teachers to customize the feedback that they provide for students based on that student's preference for feedback, what stage of writing the student is at (first, second or third draft) and what the student's level of proficiency is. Of course, this is one pedagogical practice that Criterion is not capable of. Criterion gives the same level of feedback to all students — regardless of any of the above factors. This fact is illustrated when Teacher 6 compares himself/herself to Criterion by saying "[s]o my comments are much more focused on the individual writer and focused on... umm... the major problems" (Teacher 6, Interview). It should be noted that the teachers can configure Criterion to provide feedback in certain areas. For instance, it can be configured to ignore spelling errors. This never came up in the interviews possibly because the teachers were unaware of this capability or possibly because the teachers did not feel that they had the authority to change those settings for their class. In any case, these settings are done on a per class basis — not for individual students. It would seem that this is one key area where Criterion falls short of the suggestion from pedagogical theory.

In addition, one problem with Criterion's feedback is the fact that it is one-way. That is, the feedback is provided to the student, but there is no two-way dialogue between Criterion and the student about the feedback. This means that there is no checking for

understanding that a teacher would be able to do. This is expressed clearly by Teacher 6 in the following dialogue excerpt:

Another shortcoming related to the ability to individualize and contextualize feedback has to do with the Style category of feedback. One of the errors that Criterion highlights in students essays is the "repetition of words" error. This is probably a good style suggestion — good writers generally try to vary their word choice in an essay. However, it depends on the context. In certain cases, repetition of words is unavoidable because the topic of the essay is a specific word like "music" for which there are no useful synonyms. This feedback causes some confusion and concern for the students as illustrated in the following dialogue with Student F:

Student F is concerned that she has an error in her essay (according to Criterion) but she does not have a choice since there are not always available synonyms. Teacher 2 also commented about the repetition error in the following exchange:

Once again, Criterion is failing to contextualize the feedback as pedagogical theory says it should. Note that in the quantitative questionnaire data, there was a potential correlation negative correlation between the ESL level and the satisfaction with the Style section and its feedback. It is possible that this is caused by students who have

more experience with the program getting annoyed with the "repetition of errors" problem.

One more area of feedback which has a bearing on the previously mentioned pedagogical theories, and which came up very frequently as a subtheme in the interviews, is the area of Organization and Development. The Organization and Development feedback category highlights the student's introductory material, thesis statement, main ideas, supporting ideas, and conclusion. This is the one feedback category for which a computer seems least suited since one would expect that a thorough understanding of the essay is needed to decide on issues like the validity of a thesis statement or whether the arguments to support the thesis are well structured. To be clear, Criterion does not evaluate the thesis statement — it simply takes a guess (mostly correct, but not always) about which sentence is the thesis statement. If the student mouses over the highlighted thesis statement, the program pops up a display that asks the student if it is the thesis statement, and makes some general comments about what a thesis statement is. It is possible that this causes some confusion since the other feedback sections (Grammar, Usage, Mechanics, and Style) highlight only the errors. As a result, students might think that the highlighted thesis statement is in error. It would seem that a human can provide much more individualized and contextualized feedback for this area. Teacher 3 illustrates this when he says "[t]he trouble I see is that, when the student goes for the feedback, it just highlights a sentence... it asks them the question, you know, umm... have you done this well? And it... it doesn't key in on... and then maybe something mechanical can't, but it doesn't key on... key in on what's really wrong with that sentence" (Teacher 3, Interview). Teacher 6 also comments on the shortcomings of this feedback (he

mistakenly says Style feedback when he is actually referring to the Organization and Development part of the feedback): "Criterion doesn't really give feedback on style... it just identifies... in the introduction, you can use this, this, or this, like a hook, you know, look at the writer's handbook for ideas, but it doesn't tell the writer what they... do they need a hook?... do they [need] a thesis?" (Teacher 6, Interview).

An even greater concern about the Organization and Development feedback is that Criterion sometimes gets it wrong. Teacher 6 says "... it [Criterion] doesn't like really identify them... like it will say oh this is your thesis but it doesn't always pick out the right one" (Teacher 6, Interview). Student F expressed this same concern while pointing to her thesis statement and saying "I wrote this as my thesis and I don't know why it tells I like the other" (Student F, Interview). So Criterion's Organization and Development feedback is important to help focus more on function over form; however, it does not provide enough in this area to truly help the students to focus on function. Interestingly, the weak negative correlation that was mentioned in the Results section for the quantitative questionnaire data was between the number of essays written and satisfaction with the Organization and Development. This suggests that the more the students had seen the Organization and Development feedback, the less they liked it.

One issue that applies to all the categories of feedback — not just Organization and Development is that of appropriation. As mentioned in the literature review chapter, some scholars are concerned about the issue of appropriation (F. Hyland, 2000; Brannon & Knoblauch, 1982). The concern is that by providing too much directed feedback, the teacher takes control of the essay. Although Reid (1994) argued that teachers have a feedback role to play in spite of this danger, it is still worth examining whether

appropriation is more or less likely to happen. One idea is that having multiple sources of feedback reduces the danger that one source appropriates the student's work. The idea that multiple sources of feedback is good is expressed by Teacher 3 who says "I think just having a different... OK you've got your teacher who's also giving you feedback. Having another means of feedback is excellent" (Teacher 3, Interview). The danger of a human teacher appropriating the students work is also reduced when the student puts more faith in the computer's feedback. The following dialogue with Student D shows how she sometimes felt discouraged by the teacher's negative comments and how she felt better about getting comments from Criterion:

Student E also points to the advantage of the computer over a human when she says "[w]hen I using Criterion, there is not much pressure from the teacher and from other people so you can umm... relax" (Student E, Interview). Comments by these two students indicate that they are putting more faith in the computer program. That reduces the risk of the teacher's appropriation of their work but then raises the risk of the computer's appropriation of their work. It would seem that the latter risk is lower because

Criterion does not provide enough feedback (especially in the Organization and Development category) to be able to appropriate the student's work.

4.4 Satisfaction with the program

In general, there was a very high level of satisfaction with the program. The quantitative questionnaire results strongly showed satisfaction with the program. Numerically, the average response to the questions about whether students approved of various program components was 3.87 (where Strongly Disagree was 1 and Strongly Agree was 5). In particular, the question that asked whether students agreed that the college should continue to use the program had the second highest numerical response (after Grammar Feedback satisfaction) at 4.13. Arguably, this question could be considered as most representative of overall satisfaction. Therefore this positive rating implies a positive attitude and should lead to higher L2 learning achievement (Masgoret & Gardner, 2003). That level of satisfaction was also reflected in the interviews. When discussing her impressions of Criterion, Student C said that "it's amazing" (Student C, Interview). Student E went further and said "I want to emphasize that I'm really impressed by that... the Artificial Intellect...of this program" (Student E, Interview). Teacher 5 was very impressed with the program as shown with the following dialogue:

Teacher 5's comments suggest that there is a connection between that satisfaction and how "engaged" the students are with the program.

Some teachers qualified their level of satisfaction such as Teacher 3 who says that "they [the students] see this as another...writing task, rather than a way of improving themselves. The students who, you know, are strong learners... they tend to see the good side of that" (Teacher 3, Interview). So Teacher 3 is suggesting that the program works better for the stronger students. Teacher 4 commented: "I also got feedback from the students so it [using Criterion] is helping them to a certain extent. But it's uhh... I think uhh it has its limitations" (Teacher 4, Interview).

Admittedly, the satisfaction was not unanimous. Teacher 6 was not happy with how students were left on their own when he said "... maybe it [Criterion] identifies the error correctly, but there's also no information of really how to change it. So... so that's missing... so then the students are kind of left on their own or to come ask me or someone else how to do it" (Teacher 6, Interview). Teacher 2 politely pointed out that he was not satisfied with the actual essay text entry screen as well as the outlining tool (Make a Plan) when he said "I think its [Criterion's] word-processing screen has a lot of... little room for a lot of growth and I feel that its umm outlining screen has lot of room for a lot of growth... specifically, not allowing students to revise their outline" (Teacher 2, Interview). Teacher 2 is referring to the input screen in which students type their essay. Unlike Microsoft Word or other common word processing program, Criterion's input

screen does not have nearly as many formatting capabilities. The outlining tool has some shortcomings mainly because it does not allow students to revise their outlines.

Regardless of whether participants are satisfied or not, satisfaction is related to expectations and therefore to the idea, as documented by Davis et al. (1989), that one's previous experiences with computers influences one's effective use of computers. This seems to be a reflection of the problem of computer anxiety in general. As the quantitative data about average hours spent on a computer per day show (average 3.0 hours per day), there seems to be much less computer anxiety today than there was in the past. One can speculate that this lack of anxiety helps to reduce the number of previous bad experiences and therefore would help students to make more effective use of computers. In all the recorded interviews (six teachers and six students), there were no comments from anyone complaining about using computers instead of working with humans (although there were some complaints about certain aspects of the program). This suggests that the use of computers is so universal and omnipresent, that computer anxiety is no longer a major factor.

Satisfaction is also related to whether students and teachers believe that the program is useful. As McDowall and Jackling (2006) pointed out, positive perceptions of the usefulness of a tool help learning. A couple of students had very positive comments that showed their belief in the usefulness of Criterion. Student D pointed to her belief that using the program helps her to not make the same mistakes the next time she writes: "... I know I have problems with my grammar if there's run-on sentences, punctuation... it [Criterion] give you... tell you where you have the error... and... to not have... to not have the same error next time" (Student D, Interview). Student F shows confidence in the

program when she says "I think it is the perfect uhh... software to work with in... to improve your academic skills" (Student F, Interview). Satisfaction, a perception of usefulness, engagement, and positive learning outcomes are all related. Teacher 4 commented on this when she said "I've noticed that they take it [writing with Criterion] pretty seriously... and they're very involved in their writing" (Teacher 4, Interview). This implies that the perception of usefulness is causing the students to be more engaged with their writing work.

Although the satisfaction with Criterion was by no means unanimous, overall, it was quite positive. As mentioned in the literature review chapter, Lee (2004) reported that habitual computer users preferred doing their work on computers. Given the high level of computer use by the study participants (average 3.0 hours per day usage), it is fair to say that part of their satisfaction was simply related to their preference of writing essays on a computer instead of hand-written essays. Gardner (1985) has documented the relationship between attitude and L2 learning achievement. Therefore this generally positive satisfaction level with the program suggests that it leads to better learning outcomes. As well, it is worth considering that the students and especially the teachers already have a very good understanding of how effective language learning takes place. Given that they would expect Criterion to aid in effective learning, it can be inferred that their satisfaction is based on their own internal evaluation of Criterion's effectiveness.

4.5 Understanding of the program

A lack of understanding of the program came up as an issue for both teachers and students. This would seem to have a strong bearing of whether the program will be effective. As sociocultural theory tells us, an awareness of all the environmental factors is

necessary to be able to understand how learning takes place (Donato & McCormick, 1994). This includes an understanding of the mediator — Criterion in this case — that assists the students in their learning goals. Teachers, in particular, need to have a solid understanding of Criterion in order to be able to extract the most benefit that they can from the program. It could be argued that the students' understanding of the Criterion mediator is also beneficial if those students are to take advantage of Criterion to its fullest.

In addition to sociocultural theory, the theory of genre pedagogy suggests that both students and teachers need to be aware of the constraints and purposes of the genre in which the writing is taking place and also how to successfully write within that genre. A quick run-through of the program reveals that Criterion teaches the "five-paragraph essay" genre — both implicitly and explicitly. A discussion of the advantages and disadvantages of the five paragraph genre is not the focus of this thesis, but suffice it to say that this is the genre that students must know in order to pass important university admission tests like TOEFL. Although some genre theorists (Berkenkotter & Huckin, 1993) say that genre should not be taught, Ken Hyland (2004) and Swales (1990) both say that teaching genre is a good idea. At least one student — Student E — seems to be aware of the focus on the form aspect of the genre when she implies that following the structure is more important than the content. She says "In my opinion, you can trick this system. I think I would use... write some sentences... that doesn't maybe make any sense together and computer wouldn't really find... It's still a machine, right?" (Student E, Interview). Student A explained how Criterion provides sample essays which she referred to as the models. She explains that following the sample essay as a model is the easiest

way to get a good mark — "... so I can know how to get the high mark... not using all the writing skills... just follow the model and umm... to avoid some mistakes" (Student A, Interview). While Student A may not be able to express her ideas in the terms used by scholars of genre theory, she is showing her awareness of the importance of the form aspect of the five-paragraph essay genre. One teacher, when asked about Criterion's Organization and Development feedback, showed his awareness of the constraints of the genre when he said "[i]n I mean regards to stifling creativity, I think it does a very good job of that. Umm... It teaches the students how to do very classic five paragraph essays" (Teacher 2, Interview).

Overall, it was somewhat surprising that teachers did not have as extensive an understanding of the program as one might expect. It is possible that one of the key reasons for this lack of understanding is the time constraint issue — which I will also discuss later. A couple of Criterion's utilities are good illustrations of the teachers' level of awareness. One of the utilities provides a summary of all the class marks for the teachers and can be used to find out how well the class is doing, but also can indicate which areas (grammar, style, mechanics, etc.) are the most problematic. This can be useful for the teacher to decide which areas require more focus in class. Another utility allows the teachers to add their own personal comments to the automated ones that Criterion has provided. This can overcome some of the shortcomings of Criterion's non-individualized feedback. Four out of the six teachers had not used or had no knowledge of the summary utility (Teacher Interviews). Four out of the six teachers had not used or had no knowledge of the additional comment utility (Teacher Interviews). Even more surprising was that two of the teachers had not seen the Organization and Development

feedback that Criterion provides (Teacher Interviews). This seems critical since, as discussed earlier, the Organization and Development category of feedback is particularly problematic. Criterion highlights guesses at the structural elements (thesis statement, etc.) from the student's essay but does not provide very focused help for the student to improve her essay. As Kern (2006) has pointed out, teachers need to be aware of the environment in which technology tools are used, but an understanding of that environment would be greatly enhanced by an understanding of exactly what Criterion does.

In the course of the student interviews, it became clear that students also had gaps in their understanding of the program. For instance, I interviewed Student B about an essay which she had already submitted and for which she had already viewed the feedback. When viewing the Organization and Development category of the feedback, the following dialogue took place:

Student A showed how she was impressed with Criterion but that she also realized its limitations when she said "... the people who make it [Criterion] is talented, but machine's always a machine" (Student A, Interview). These two students are not under any illusions about the infallibility of computers.

4.6 Time Constraints

The theme of time constraints came up quite often. It should be noted that one of the reasons that the theme occurred was simply because it was one of the interview questions for both students and teachers (See Appendix C - Student Written Questionnaire for the student interview questions). In any case, one can expect that time would be a common theme in any formal teaching environment where there is an emphasis on class schedules, semester deadlines, and high-stakes testing.

Based on the interview data, in general, teachers felt that time constraints limited their ability to fully understand and manage the program. Among the comments in this vein was the following from Teacher 4 who was concerned in the interview that he was not able to provide me with much information. He wanted to become more familiar with Criterion but felt that lack of time made it impossible.

In a similar vein, Teacher 5 claimed that he could not use the utility for providing personal comments to the students due to the time factor:

It is clear that ESL Teachers are under great time pressures for numerous reasons. Those reasons are not the concern of this thesis; however, it does need to be noted that Criterion's effectiveness is adversely affected by those time pressures.

This problem with time seems paradoxical since one of the purposes of Automatic Writing Evaluation systems like Criterion is to save time (Enright & Quinlan, 2010). In support of this, some teachers and students (mostly students) do mention that Criterion can save time. Student B said "[i]t [using Criterion] saves time... not just the students... also saves the teachers' time" (Student B, Interview). Student E said "[i]t's [Criterion] a wonderful opportunity to test people without spending valuable time of teachers" (Student E, Interview). Student F said "[i]t [Criterion] is useful 'cause umm... umm... it

can help you to... to write um... as many as writings as you like 'cause your... your teacher cannot correct for you everything" (Student F, Interview). Student F's idea that more writing is better is also expressed by Teacher 2 who says "[t]he number one greatest thing about the program, in my opinion, is that, it provides an opportunity for students to write the necessary amount of essays it takes to write to become good at writing essays" (Teacher 2, Interview). One way to put this then is that although an initial time investment is needed by teachers to make the most effective use of Criterion, it would seem that, in the long run, there are time-saving benefits to the program.

An important benefit of Criterion with respect to time is its immediacy. Feedback in class from teachers can take several days whereas Criterion feedback comes as soon as the student clicks on the submit button. Student F explained this, saying "When you're writing Criterion, it gives you right away your... your writing... your feedback" (Student F, Interview). This is a key advantage for computers over humans. Immediate feedback means that students can make a direct association between their current mistake and the correction to allow them to avoid making that mistake in the future (Attali & Powers, 2010). Delaying the feedback means that students may not remember everything that was going on when they made the mistake. "If feedback is delayed, it allows learners to practice errors" (Scheeler & Lee, 2002, p. 232). This is a clear case where Criterion is following the edicts of language learning pedagogy in a way that is not practical for human teachers in a classroom setting.

So it would seem that time constraints limit teacher's full understanding of the program, but the program may paradoxically be able to save time in the long run. The perception of Criterion is that of a large tool which requires some initial time investment

to fully understand, but it is perceived as a time-saver in the long run. Time constraints are an important factor to be considered in an evaluation of a program such as Criterion (Jamieson et al., 2004). Criterion adheres well to pedagogical theory regarding the importance of immediate feedback (Scheeler & Lee, 2002) and seems to even exceed human teachers for whom practical constraints of large classroom settings limit the immediacy of the feedback that they provide.

4.7 Technical Issues

The first type of problem to be analysed is that of technical glitches — those problems not related at all to language learning — but simply related to computer bugs. One such occurrence happened when I was interviewing Student D. During the interview, the Grammar feedback section displayed just a blank page. Later on we tried it again and it worked, but there was definitely a problem with the display. This one small technical glitch in all my interactions with the program seems pretty minor. However, the other technical glitch worth mentioning is probably more closely related to ETS's management of their program, because, as Teacher 6 revealed, the Criterion web site was down for 48 hours at one point during the term (Teacher 6, Interview). This means that all of the worldwide users lost their access to the system. Presumably, ETS was doing some kind of maintenance and/or upgrades on their web server, but this length of time for such a widely used system seems quite excessive.

Another type of problem which was brought up quite often was that of "lost essays". Both Teacher 1 and Teacher 6 complained of this problem with Teacher 1 specifically mentioning how this happens when students write their essays at home — "[t]his semester I had two students complaining about it... the disappearing essay... you

write it at home... you think you've saved it and think everything is hunky-dory. You come back the next day in the lab, you try to retrieve it... poof... gone... thin air" (Teacher 1, Interview). This is a difficult problem to analyze because it can occur for several reasons and it is hard to isolate the reasons that caused each incident. First of all, a student's claim that the system lost her essay could simply be a convenient excuse for not doing assigned work. Second, the lost essay could be the result of a student unknowingly disconnecting their computer from the web site before saving the essay. Third, a lost essay could be caused by a network problem of any sort that disconnects the student computer from Criterion's web site before saving. Fourth, it is possible that something was truly wrong with Criterion and a Criterion software error caused an essay to be deleted. A fifth possible reason, which is more subtle, is that the user interface for Criterion might not be clear enough for students to know how to save their essays. Laying blame in the case of a poor user interface is not so clear-cut. The problem is serious enough that Teacher 6 has taken to sometimes having students email her the essay to get around the risk of Criterion losing it (Teacher 6, Interview). The user interface in general is then an important technical issue to discuss.

Both students and teachers commented that the user interface in general was good and easy to use. However, one complaint about the user interface, which seemed quite basic, was that there is no "undo" button for students (Teacher 6, Interview). Most student actions on the web site probably do not need an "undo" button because they are simply displaying information — if they accidentally display the wrong thing, it is easy to just go back to another display page. The item that might cause trouble then is the actual submission of the essay. This is something that ETS does not want to be undone

because it is the basis of how the college is charged for the program. Each essay submission invokes a license charge. Another more serious problem has to do with the user interface for the Make a Plan utility. This is the utility that allows students to make an outline. As mentioned earlier, there is already a problem with the outline in that students cannot change the outline once they have submitted it. Due to this, and possibly some other confusion, I discovered that Student D had cut and pasted an outline into the top of her essay (Student D, Interview). The outline was a simple point form outline. As a result, her submitted essay had numerous errors flagged because Criterion was treating the outline as part of the essay itself.

Teachers had other complaints about some of the more subtle aspects of the user interface. For instance, in order to see Criterion's feedback on their essays, students must click on the button "View Trait Feedback Analysis". As Teacher 2 mentioned, "students often ask me what that means" (Teacher 2, Interview) because the word trait in this context is not obvious — especially to ESL students. When discussing the Style feedback section, Teacher 3 said "I'm not sure how easy it is for the student to interpret that" (Teacher 3, Interview). He was concerned with the feedback displayed being hard for students to interpret, but that this difficulty with the user interface was made worse due to the one-way nature of the feedback. Students are unable to ask for clarification or explanation about what the program tells them. The fact that Criterion's feedback is one-way contrasts with the "social activity" (K. Hyland, 2007, p. 152) definition of writing that suggests interaction with the teacher regarding feedback is important.

There was another class of technical error that I consider to just be due to the inherent complexity of trying to interpret English writing. As mentioned earlier, the

Organization and Development category feedback is very difficult for a computer program to provide. In the section on Feedback earlier in this chapter, I demonstrated that Criterion only tries to identify the thesis statement and other structural elements without determining if they are logical and make sense. This seems to strike to the core of the matter which is that Criterion cannot really help the student with their thesis statement any more than a text book that describes how to write a good thesis statement. Another technical issue related to the complexity of English is the "repetition of words" error mentioned earlier that several teachers and students complained about. As Teacher 6 explained, "... it will identify words that they used repetitively, umm... however, it doesn't take into any con-... umm... into consideration the context. So for example, they were writing on electronic devices. Well, they're going to have those words 'electronic devices' repetitively throughout their... umm... essay... but in the context, there is no synonym for that. But Criterion identifies it and says you need to have a synonym" (Teacher 6, Interview). The ability to manage some of these aspects of language complexity is delving into the field of Natural Language Processing. Unfortunately, as stated in the literature review chapter, Abramson (2011) says that there is no computer system today that can truly pass the "Turing Test" (a test that indicates true language understanding).

A couple of other errors related to the complexity of the language can be attributed to incorrect analysis by Criterion of English structure. An interesting example is from Student B who had written the phrase "When in Rome do as the Romans" in her essay (Student B, Interview). Criterion highlighted the word "do" and suggested that there was a subject-verb agreement error. In fact, what was needed was a comma after "Rome" — something that a human would very easily be able to determine. Another

example is from Student E who wrote in her essay "... many people want to be beautiful and willing to pay an ultimate price for it" (Student E, Interview). Criterion highlighted the article "an" and indicated that this was the wrong article — it should be "the" instead. What is really needed is the verb "are" before willing. This helps to illustrate that Criterion is not perfect and some of the mistakes it makes would be easy for a human to correct.

As I stated at the beginning of this section, true technical glitches by Criterion are very minimal. Since technical issues are probably the source of much computer anxiety and that computer anxiety can inhibit students from using computer programs such as Criterion (Conti-Ramsden, et al., 2010), this relative stability of the program is a very positive thing. It may also explain a certain amount of the high satisfaction levels shown with the program. No complex computer program is perfect, so it should be mentioned that Criterion has some technical limitations due to some minor design and user interface issues. Most importantly, it is technically limited by its inability to truly understand language. Sociocultural theory hints at this problem because it states that an understanding of learning requires an understanding of the virtually limitless environmental and contextual factors involved. If this is required to understand how learning takes place, it is certainly required to understand the content of written essays on numerous topics. Computer technology today is not very good at this and therefore some of Criterion's areas of feedback — particularly the Organization and Development category — are not as good as human feedback.

4.8 Student and Teacher Recommendations

In addition to the emergent themes, there were many recommendations from students and teachers that were raised during the interviews. The two key recommendations that came up most often were:

1. More training on Criterion is needed for both students and teachers, but more critically for teachers.
2. The college should continue to use the program.

4.8.1 More Training

The idea that more training is needed for students and teachers is related to the emergent themes of understanding of the program and time constraints. A few of the teachers commented on how they felt that were not very familiar with the program. It was felt that more training for the teachers themselves was needed on how the program works and how the teachers should make use of it. There were many features of the program that teachers only became aware of because I asked about them during the interviews. The issue of time constraints may be an explanation for the lack of training for teachers. One teacher commented that teachers should be aware that the program is not standalone — it requires oversight (Teacher 6, Interview).

Some students and teachers were very explicit about the need for more training and also talked about some of the barriers to getting that training. Teacher 5 said "[w]ell, personally, I'd like a little more time. It would be nice if the college offered a little small workshop" (Teacher 5, Interview). Teacher 2 said "I think the school really needs to be trained on it because it's not quite intuitive enough" (Teacher 2, Interview). Some provided helpful suggestions for how to get more training. Teacher 5 suggested that

teachers can become better familiarized with the program by reviewing Criterion's sample essays: "I would suggest umm... umm... looking at some of the sample essays, which I did not do, and the evaluations of those essays" (Teacher 5, Interview). Similarly, Teacher 6 suggesting that teachers should "[t]ake an essay that you've marked, cut and paste it, and compare your... your feedback to Criterion feedback" (Teacher 6, Interview).

4.8.2 Continue to Use the Program

The suggestion to continue to use the program was almost unanimous amongst the interviewees. Almost all of the students and teachers felt that the college should continue to use the program despite the various issues that were analysed above. All six students recommended that the college should continue to use the program (Student Interviews). Most of the teachers recommended that the college should continue to use the program. Teacher 2 recommended that the college continue to use the program but said that he would like to see some changes such as a better input screen and a better outlining utility. Teacher 5 recommended that the college continue to use the program but was concerned with the students' workload — the number of essays students were being asked to write. Finally, Teacher 6 was the only one who did not recommend the college continue to use the program. Teacher 6 had several criticisms of the program (many of which were discussed earlier in this chapter) and therefore did not feel that the program was a good idea. The almost unanimous support for the idea of continuing to use the program is also reflected in the result from the quantitative questionnaire data that indicated very strong support for the suggestion with a numerical value of 4.13 (where Strongly Disagree was 1 and Strongly Agree was 5).

Chapter 5: Discussion and Conclusion

In this discussion and conclusion chapter, I will revisit the five themes that emerged in the findings. I will do this with a view to addressing the research questions and I will attempt to frame the discussion in the context of sociocultural theory, genre theory, views about attitude and motivation, ESL writing pedagogy approaches, CALL, and research on Criterion and related programs. In this chapter, I will first present a summary of the findings and a discussion of how they relate to the literature. I will then describe the study's contributions to the field, followed by a discussion of the limitations of the study. Next, I will describe some implications for further research possibilities and I will close with a summary.

5.1 Summary of Findings and Discussion

This thesis has attempted to explore the use of the Criterion computer program which is used as a tool for teaching ESL students writing skills. The study attempted to answer the following two questions:

What are the perceptions of both students and teachers about the Criterion computer program?

How well does Criterion adhere to current pedagogical theories and practices?

The fourteen Likert scale questions from the written questionnaire were shown to be internally consistent through the Cronbach Alpha test. Spearman's Rho correlation tests showed some potential weak statistical correlations between the following five pairs of question responses:

1. A positive correlation between approval of the Writer's Handbook utility and Age (older people liked the Writer's Handbook more)
2. A correlation between Mechanics Feedback and Gender (females are more likely to favour the mechanics feedback)
3. A negative correlation between Style Satisfaction and ESL Level (higher level classes liked the style section less)
4. A negative correlation between Style Feedback and ESL Level (higher level classes liked the style feedback less)
5. A negative correlation between Organization/Development Satisfaction and Number of Essays written (people who had written more essays with Criterion like the Organization and Development section less)

Pure chance is one possible reason for these correlations being flagged as significant (since a $p < 0.05$ value means that the tests will find significance by chance in 1 out of 20 tests and there were 84 correlation tests run). The three negative correlations could indicate that more experience leads to less satisfaction with two noted trouble areas of the program (Style and Organization and Development). The potential positive correlation between gender and mechanics feedback suggests that females are more focused on spelling and punctuation correction. Finally, the potential correlation between the age and approval of the Writer's Handbook might be because older students appreciate the idea of an old style reference text book or because older students are more patient with their writing work and therefore are willing to take the time to study the details of the grammar theory.

No correlations were found to support the suggestion by Conti-Ramsden et al. (2010) that people who have high levels of computer anxiety are less likely to use computers. Support for Conti-Ramsden et al. (2010) could have come from a correlation between the number of hours spent on a computer per day and satisfaction levels with the program or certain parts of it. This potential correlation would occur if one assumes that the number of hours spent on a computer per day relates to lower computer anxiety, and that higher use of the program leads to more positive perceptions. No correlation was found, but the more likely explanation is that greater use of the program doesn't lead to more positive perceptions.

The coding of the qualitative interview data yielded five key themes that emerged as the most important ones. Here is a summary of those themes and what can be concluded about each of them:

1. Usefulness of feedback (especially Organization and Development)

In general, Criterion is perceived very favourably for its feedback. This supports the finding by Fang (2010) who found that students liked the feedback provided by the similar program, MyAccess. Among Criterion's shortcomings is the fact that its feedback is not customized. This goes against the pedagogical practice encouraged by Hyland and Hyland (2006) who say that good feedback is individualized for each student. It also agrees with Dikli (2010) who said that teacher feedback was more focused than the similar MyAccess program feedback. Another shortcoming is Criterion's lack of context whereas Kumaravadivelu (1994) listed "contextualize linguistic input" as one of the key macrostrategies for language teaching. As well, Criterion's feedback is one-way (e.g. students cannot ask Criterion for clarification) whereas Kumaravadivelu (1994) listed

"facilitate negotiated interaction" as another key macrostrategy for language teaching. Finally, Criterion does not understand meaning for the organization and development part of an essay. This is part of the reason for the above listed shortcomings, but it also makes it difficult for Criterion to follow Kumaravadivelu's (1994) macrostrategy of "ensuring social relevance". This fact also agrees with the criticism of Vojak et al. (2011) who say that working alone with a computer is not a socially situated activity. This lack of sociocultural focus probably leads to the students focusing more on form instead of content which supports the claim of Fang (2010) for the MyAccess program. The fact that Criterion cannot do these things supports Abramson's (2011) claim that there are no computer systems today that can use language as naturally as humans and pass the "Turing Test". As sociocultural theory says, the context is critical for an understanding of how learning takes place (Donato & McCormick, 1994) and since Criterion is not aware of all the contextual elements, it cannot adapt to the specific needs of the situation. In terms of the genre pedagogy described by Rose (2008), Criterion has the feedback in the wrong order. Rose (2008) describes the first phase of genre teaching as being deconstruction. Through its feedback, Criterion is providing that deconstruction but only after the student has written an essay (the independent construction phase).

2. Satisfaction with the program

Sociocultural theory would suggest that the level of satisfaction with the tool used in the mediated action of learning (in this case, Criterion) is one of the many contextual elements to be considered. Although the satisfaction with Criterion was by no means unanimous, overall, it was quite positive. Gardner (1985), Masgoret and Gardner (2003), and McDowall and Jackling (2006) have documented the relationship between attitude

and L2 learning achievement. Therefore this generally positive satisfaction level with the program suggests that it would lead to better learning outcomes with Criterion. As well, it is worth considering that it is likely that the students and especially the teachers already have a very good understanding of how effective language learning takes place. Given that they would expect Criterion to aid in effective learning, it can be inferred that their satisfaction is based on their own individual internal evaluation of Criterion's effectiveness. The high level of satisfaction therefore provides some weak support for the literature that suggests benefits from Criterion (Burstein et al., 2004; Chodorow et al., 2010; Rock, 2007). The notable fact that students did not express any concerns with computer anxiety suggests that the earlier focus on computer anxiety (Todman, 2000; Conti-Ramsden et al., 2010) is not as great a concern as it once was. The general satisfaction with the evaluation component of Criterion and the fact that there were very few complaints about evaluation provide some support for the research establishing the validity of AWEs (Shermis et al., 2002; James, 2006; James, 2008; Attali & Burstein, 2006; Attali & Powers, 2009; Weigle, 2010). It should be pointed out that Kumaravadivelu's (1994) post-method description of macrostrategies for teaching ESL does not take into account the issue of satisfaction so this might be a limitation of his teaching theory. In addition, genre pedagogy as described by Martin (2009) does not directly address the issue of how satisfaction affects the learning situation so this is also a possible limitation of his genre theory.

3. Understanding of the program

It would seem that there is some awareness by students and teachers of the constraints of the genre used with Criterion (its explicit teaching of the five paragraph

essay genre). By only teaching one specific genre, Criterion is not satisfying the suggestion by Martin (2009) that the many genres of a culture should be taught. An awareness of this limitation is therefore important so that teachers can supplement Criterion's single genre with others. The fact that Criterion is explicitly teaching a genre generally supports the position of the Sydney School (Martin, 2009); however, it does not, by itself, support the Sydney School's idea of the three phases of teaching genres (deconstruction, joint construction, and independent construction) (Rose, 2008) because its deconstruction is somewhat embedded in the feedback, there is no joint construction, and the sole focus is on the independent construction. Among Ken Hyland's (2007) list of key advantages to genre-based writing pedagogy are the fact that it can be explicit, needs-based, supportive, and critical. Criterion does not really provide these advantages since it only explicitly describes the five-paragraph genre after the student has entered an essay, its essay topics are not necessarily based on students needs, its support only comes after the student has entered an essay, and it does not allow any opportunity for students to be critical of the discourse of the five-paragraph essay. Teachers need to be aware of these shortcomings so that they can attempt to compensate. For example, teachers could provide time in the classroom where students are encouraged to question and criticize the five-paragraph genre being taught.

The notable lack of understanding of some of the details of the program also limits Criterion's effectiveness because sociocultural theory states that an awareness of all the environmental factors is necessary to be able to understand how learning takes place (Donato & McCormick, 1994). This includes an understanding of the mediator — Criterion in this case — that assists the students in their learning goals. Teachers, in

particular, need to have a solid understanding of Criterion in order to be able to extract the most benefit that they can from the program. According to Kumaravadivelu (1994), one of his macrostrategies for teaching ESL is to "minimize perceptual mismatches", so this validates Kumaravadivelu's belief in the importance of understanding in the learning context. The research cited in this thesis about Criterion and similar programs does not seem to directly examine the importance of students' and teachers' understanding of the program so this suggests an area where there should be more focus. Finally, the lack of understanding and awareness of the program has an indirect bearing on the value of the perceptions of students and teachers about the program. If they do not understand some of the important details of the program, it weakens the value of their perceptions.

4. Time constraints

Although sociocultural theory helps researchers to be more aware of the numerous issues involved in the sociocultural context and how important they are in order to understand learning (Donato & McCormick, 1994), the theory's applicability to actual ESL teaching is limited by certain practical constraints. Time is one of those constraints as ESL teachers cannot possibly deal with all of the contextual issues involved in the day to day practice of teaching. It would seem that time constraints limit teacher's full understanding of the program, but the program may paradoxically be able to save time in the long run. The time constraint issue is an important aspect of practicality to be considered in an evaluation of a program such as Criterion (Jamieson et al., 2004), but the resolution of the time constraint problem may have more to do with the institutional and related environmental factors than with Criterion itself. Since Kumaravadivelu's (1994) list of macrostrategies does not directly consider time

constraints, his list could potentially be supplemented with a consideration of this idea. As well, genre pedagogy as described by Martin (2009) does not take into consideration the constraints imposed by time, so genre pedagogy could also potentially be strengthened by taking this into account.

Criterion adheres well to pedagogical theory regarding the importance of immediate feedback (Scheeler & Lee, 2002) and seems to even exceed human teachers for whom practical constraints of large classroom settings limit the immediacy of the feedback that they can provide. It would seem that this immediate feedback advantage should therefore be exploited as much as possible.

5. Technical Issues

Many of the student and teacher perceptions of Criterion centred on or were coloured by technical issues with the program. An awareness of these technical issues is necessary because it is an important part of the context in which the learning takes place and sociocultural theory argues that the context is important. Overall, it seems that the program is quite technically sound and reliable, but there are a few technical problems, pedagogical problems and problems that may appear to be technical but are really related to a lack of understanding. As described in the literature review chapter, CALL capabilities have grown substantially over the years (Garrett, 2009) to the point that the use of computer tools for language learning is quite commonplace. It means that the theme of technical issues applies to many aspects of language learning — not just Criterion. Most importantly, Criterion is technically limited by its inability to truly understand language. Computer technology today cannot do this (Abramson, 2011) and therefore some of Criterion's areas of feedback — particularly the Organization and

Development category — are not ideal and should probably be supplemented by human feedback. Because of its inability to truly understand language, Criterion is therefore failing to follow Kumaravadivelu's (1994) macrostrategies of "contextualizing linguistic input", "facilitating negotiated interaction", and "ensuring social relevance".

Supplementing Criterion with human feedback is therefore in agreement with Burstein et al. (2004) who say that Criterion should be used "under the instruction of a classroom teacher" (p. 27). Although Rose (2008) provides a rich description of how genre pedagogy should be done, the descriptions do not take use of computer technology into consideration. Given that CALL is becoming more and more prevalent in the field of language learning, the important of technical issues should be incorporated into genre pedagogy theory.

5.2 This Study's Contributions

These results provide a small incremental addition to an understanding of how students and teachers view the Criterion program, but also to the field of CALL in general. To the best of my knowledge, all the published research on the Criterion program has at least one of its authors connected to ETS so this study represents a first independent investigation of the program. The results provide guidance for both CALL application developers and users of CALL applications. First, the results can be used as guidance for researchers who continue to develop the Criterion program as well as programs that are similar to it. By taking into account the input from teachers who indicated that they rarely used certain utilities and the input from students who sometimes demonstrated a lack of understanding about some aspects of the program (as well as some positive feedback), developers of Criterion can create an even better product. The

developers can do this in a number of ways such as by providing better, more concise help and online visually-oriented tutorials and by improving the user interface to ensure that unused utilities are more visible. Not only is this useful for Criterion developers, but it is also useful for developers of other similar computer programs used to teach writing (such as MyAccess). Some of the feedback gathered in this study (such as problems with the user interface and how users' ability to spend time learning how the program works is constrained by the amount of time available) is applicable to all developers of CALL applications. Anyone who is developing a CALL application needs to ensure that the user interface is as simple and intuitive as possible and that learning how to use the application does not require a large time investment.

Secondly, the findings are relevant to those who wish to use any kind of computer technology as a tool for helping students learn a language. The findings have shown that a lack of understanding about the computer program or application in use can cause problems; therefore, it might be advantageous to put more energy into ensuring that students understand the computer program that they are working with. More specifically, this study can be used to provide guidance to teachers who use the Criterion program for their ESL writing classes. If teachers are aware of some of the weaknesses of Criterion (such as its feedback on organization and development), they can compensate by providing extra feedback themselves or educating their students about what to expect from the program. The recommendations described in the section 4.8 Student and Teacher Recommendations also provide guidance since one of the two key suggestions was that more training about the program should be provided for both students and teachers. If teachers are better trained on the use of the program, they will be more aware

of its weaknesses. The other recommendation was that the college should continue to use the program, so this is helpful since it implies that the program is useful.

5.3 Limitations of this Study

This study had several limitations which should be mentioned here. First, due to constraints of resources and time, only one session (8-week term) of classes could be studied. A more extensive study could look at several sessions of the level 7 and level 8 classes. As well, the study only looked at one college where the program is being used. In order to get a wider view of how the program is used in other contexts, it would be useful to study similar classes using Criterion at other institutions. The qualitative interview data for students was necessarily drawn from a subset of the students using the program (six of a total of 73 students). This means that it is quite possible that many different views were not represented. As well, the interviewed students were volunteers, and although I made it clear to students when I presented to their classes that any volunteers were welcome, the fact that they were volunteers introduces a bias to the data. It was not possible to randomly select students since I was not in a position to force anyone to submit to an interview. Another important limitation to mention is that most students did not have a great deal of experience with Criterion. The average number of essays written on Criterion by the students in the study was only 1.8. It would be more valuable to work with students who had much more experience with the program and therefore might have more insight into its strengths and weaknesses. I should acknowledge my own bias in this area. I am an ESL teacher and someone with a computer background and although I wish to see programs like Criterion succeed, I made every attempt to analyse the data impartially. I should acknowledge as well that although all reasonable attempts were

made to avoid it, subjectivity in the interpretation of the data, especially the qualitative interview data, is inevitable.

5.4 Implications for Further Research Directions

There are many areas for furthering this research. First, as mentioned earlier, a large independent longitudinal study of the effectiveness of Criterion would be very valuable. Do students who use the program learn better writing skills than students who use more traditional teacher-led training? Based on the idea suggested above about teachers providing additional feedback to students, it would be useful to study the effectiveness of a hybrid of traditional teacher-led and Criterion-led writing instruction. Due to constraints of time and space, I chose to highlight only five themes from the extensive interview data which I gathered. However, other researchers could analyse the same data and perhaps find other important themes or report on the data more extensively especially if they use a different method of data analysis such as discourse analysis. Many of the specific aspects of Criterion's functionality could become the focus of research. For example, a study of the effectiveness of just the Style feedback could be conducted. Does this Style feedback help at all? Although many claimed that the "repetition of words" error was annoying, in the long run, does it improve students writing? Similar studies could be done for the Mechanics feedback, the Grammar feedback, the Usage feedback, and the Organization and Development feedback.

Independent of Criterion, there are areas in the field of computer technology and Natural Language Processing (NLP) by computers which need more research. Considering what sociocultural theory tells us about the need for an understanding of everything about social context and culture and history, is it even possible for computers

to learn to understand and use language naturally like humans? Without this ability, programs like Criterion are limited especially in the area of Organization of Development of written essays since they cannot tell the students very much about what they have written.

5.5 Summary

This thesis has reported on a study of the perceptions of ESL students and teachers using the Criterion computer program to aid in the acquisition of English writing skills. The study used a mixed method approach involving written questionnaires and oral interviews. This quantitative data was analysed statistically and the qualitative data was analysed using a thematic analysis which was based on a Grounded Theory approach. The results have shown that Criterion is a useful program but both teachers and students should be aware of its limitations (such as its focus on one genre and its weakness when providing feedback for organization and development) so that they may better plan their learning activities. Ten years ago, programs as sophisticated as Criterion were not possible. Even the worldwide access through any computer on the internet to a system like Criterion was not as universal as it is now. Given how unpredictable advances in the field of computers can be, it is very difficult to know what technology will provide for us ten years in the future. The field of CALL will undoubtedly provide many surprising products to help the language learners of the future.

References

- Abramson, D. (2011). Philosophy of mind is (in part) philosophy of computer science. *Minds and Machines*, 21(2), 203-219.
- Abrahamsson, N., & Hyldenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, 59(2), 249-306.
- Ahmad, K., Corbett, G., Rogers, M., & Sussex, R. (1985). *Computers, language learning, and language teaching*. Cambridge: Cambridge University Press.
- Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal*, 78(4), 465-483.
- Arecco, M. R., & Ransdell, S. (2002). Early exposure to an L2 predicts good L1 as well as good L2 writing. In S. Ransdell & M. Barbier (Eds.), *New directions for research in L2 writing* (pp. 123-131). London: Kluwer Academic Publishers.
- Aronson, J. (1994). A pragmatic view of thematic analysis. *The Qualitative Report*, 2(1). Retrieved from <http://www.nova.edu/ssss/QR/BackIssues/QR2-1/aronson.html>
- Atkinson, D. (2003). Writing and culture in the post-process era. *Journal of Second Language Writing*, 12(1), 49-63.
- Attali, Y. (2011). Sequential effects in essay ratings. *Educational and Psychological Measurement*, 71(1), 68-79.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning and Assessment*, 4(3), 3-30.

- Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement*, 69(6), 978-993.
- Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement*, 70(1), 22-35.
- Ayres, R. (2002). Learner attitudes towards the use of CALL. *Computer Assisted Language Learning*, 15(3), 241-249.
- Beatty, K. (2010). *Teaching and researching computer-assisted language learning*. London: Pearson Education Limited.
- Belcher, D. D. (1999). Authentic interaction in a virtual classroom: Leveling the playing field in a graduate seminar. *Computers and Composition*, 16(2), 253-267.
- Berkenkotter, C., & Huckin, T. N. (1993). Rethinking genre from a sociocultural perspective. *Written Communication*, 10(4), 475-509.
- Bialystok, E. (1997). The structure of age: in search of barriers to second language acquisition. *Second Language Research*, 13(2), 116-137.
- Bidlake, E (2010). *Going solo: Case-studies of learners grappling with self-instructed CALL*. University of Newcastle upon Tyne.
- Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1), 1-17.
- Brannon, L., & Knoblauch, C. H. (1982). On students' rights to their own texts: A model of teacher response. *College Composition and Communication*, 33(2), 157-166.

- Brown, H. D. (2007). *Teaching by principles - An interactive approach to language pedagogy*. White Plains, NY: Pearson Education.
- Burstein, J., Chodorow, M., & Leacock, C. (2004) Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25(3), 27-36.
- Caws, C. G., Hamel, M., & Schulze, M. (2011). CALL in Canada: Examples of current research. *CALICO Journal*, 28(3), 571-579.
- Chapelle, C., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301-315.
- Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3), 419-436.
- Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51-89.
- Chu, R. J. (2010). How family support and internet self-efficacy influence the effects of e-learning among higher aged adults - Analyses of gender and age differences. *Computers & Education*, 55(1), 255-264.
- Coe, R. M. (1994). Teaching genre as process. In A. Freedman & P. Medway (Eds.), *Learning and teaching genre* (pp. 157-169). Portsmouth, NH: Boynton Cook Publishers.
- Conti-Ramsden, C., Durkin, K., & Walker, A. J. (2010). Computer anxiety: A comparison of adolescents with and without a history of specific language impairment (SLI). *Computers & Education*, 54(1), 136-145.

- Cope, B., Kalantzis, M., McCarthy, S., Vojak, C., & Kline, S. (2011). Technology-mediated writing assessments: Principles and processes. *Computers and Composition*, 28(2), 79-96.
- Cotos, E. (2011). Potential of automated writing evaluation feedback. *CALICO Journal*, 28(2), 420-459.
- Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, 8(2), 73-83.
- Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72-93). Philadelphia: Multilingual Matters.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.
- Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal*, 28(1), 99-134.
- Donato, R., & McCormick, D. (1994). A sociocultural perspective on language learning strategies: The role of mediation. *The Modern Language Journal*, 78(4), 453-464.
- Dornyei, Z. (2003). Attitudes, orientations, and motivations in language learning: Advances in theory, research, and applications. *Language Learning*, 53(S1), 3-32.
- Education Testing Service (2008). Criterion brochure. Retrieved from:
http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf
- Education Testing Service (2011a). Criterion® online writing evaluation service.
 Retrieved from <http://www.ets.org/criterion>

- Education Testing Service (2011b). Frequently asked questions about ETS. Retrieved from: <http://www.ets.org/about/faq/>
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317-334.
- Fang, Y. (2010). Perceptions of the computer-assisted writing program among EFL college learners. *Educational Technology & Society*, 13(3), 246-256.
- Ferris, D. (1993). The design of an automatic analysis program for L2 text research: Necessity and feasibility. *Journal of Second Language Writing*, 2(2), 119-129.
- Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, 8(1), 1-11.
- Fox, J., Cheng, L., Berman, R. and Song, X., & Myles, J. (2006). Costs and benefits: English for academic purposes instruction in Canadian universities. *Carleton Papers in Applied Language Studies*, 23(1). Retrieved from http://www1.carleton.ca/slals/ccms/wp-content/ccms-files/1_Costs_and_Benefits_FOX_CHENG_BERMAN_SONG_MYLES.pdf
- Flick, U. (2009). *An introduction to qualitative research*. London: SAGE.
- Freedman, A., & Medway, P. (1994). New views of genre and their implications for education. In A. Freedman & P. Medway (Eds.), *Learning and teaching genre* (pp. 1-22). Portsmouth, NH: Boynton Cook Publishers.
- Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: E. Arnold.

- Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second language learning*. Rowley, MA: Newbury House.
- Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *Modern Language Journal*, 75(1), 74–101.
- Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal*, 93(s1), 719-740.
- Gentil, G. (2011). A biliteracy agenda for genre research. *Journal of Second Language Writing*, 20(1), 6-23.
- Gibbs, G. (2007). *Analyzing qualitative data*. London: SAGE Publications.
- Giles, H., & Johnson, P. (1987). Ethnolinguistic identity theory: A social psychological approach to language maintenance. *International Journal of the Sociology of Language*, 68, 69-99.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine Publishing Company.
- Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123-145.
- Hamp-Lyons, L. (1991). Reconstructing "academic writing proficiency". In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127-153). Norwood, NJ: Ablex Publishing.
- Harmer, J. (2004). *How to teach writing*. Harlow, England: Pearson Education Limited.
- Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145-159.

- Hirvelä, A. (2005). Computer-based reading and writing across the curriculum - Two case studies of L2 writers. *Computers and Composition*, 22(3), 337-356.
- Huckin, T. (2004). Content analysis: What texts talk about. In C. Bazerman & P. Prior (Eds.), *What writing does and how it does it: An introduction to analysing texts and textual practices* (pp. 13-32). New York: LEA.
- Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, 7(3), 255-286.
- Hyland, F. (2000). ESL writers and feedback: Giving more autonomy to students. *Language Teaching Research*, 4(1), 33-54.
- Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System*, 31(2), 217-230.
- Hyland, K. (2003). *Second language writing*. New York: Cambridge University Press.
- Hyland, K. (2004). *Genre and second language writing*. Ann Arbor: University of Michigan Press.
- Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, 16(3), 148-164.
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83-101.
- Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly*, 30(4), 693-722.
- James, C. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing*, 11(3), 167-178.

- James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing*, 13(2), 80-92.
- Jamieson, J., Chapelle, C., & Preiss, S. (2004). Putting principles into practice. *ReCALL*, 16(2), 396-415.
- Kern, R. (2006). Perspectives on technology in learning and teaching languages. *TESOL Quarterly*, 40(1), 183-210.
- Kost, C. (2011). Investigating writing strategies and revision behavior in collaborative wiki projects. *CALICO Journal*, 28(3), 606-620.
- Kristensson, P. O. (2009). Five challenges for intelligent text entry methods. *AI Magazine*, 30(4), 85-94.
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, 1(4), 279-308.
- Kumaravadivelu, B. (1994). The postmethod condition: (E)merging strategies for second/foreign language teaching. *TESOL Quarterly*, 28(1), 27-48.
- Lee, H. K. (2004). A comparative study of ESL writers performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, 9(1), 4-26.
- Leki, I. (2000). Writing, literacy, and applied linguistics. *Annual Review of Applied Linguistic*, 20, 99-115.
- Macdonald, N., Frase, L., Gingrich, P., & Keenan, S. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, 30(1), 105-110.
- Martin, J. (2009). Genre and language learning: A social semiotic perspective. *Linguistics and Education*, 20(1), 10-21.

- Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, 53(1), 123-163.
- McDowall, T., & Jackling, B. (2006). The impact of computer-assisted learning on academic grades: An assessment of students' perceptions. *Accounting Education: an international journal*, 15(4), 377-389.
- Murday, K., Ushida, E., & Chenoweth, N. A. (2008). Learners' and teachers' perspectives on language online. *Computer Assisted Language Learning*, 21(2), 125-142.
- Murray, G. L. (1999). Exploring learners' CALL experiences: A reflection on method. *Computer Assisted Language Learning*, 12(3), 179-195.
- Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238-243.
- Peters, M., Weinberg, A., & Sarma, N. (2009). To like or not to like! Student perceptions of technological activities for learning French as a second language at five Canadian universities. *The Canadian Modern Language Review*, 65(5), 869-896.
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2), 103-134.
- Reid, J. (1994). Responding to ESL students' texts: The myths of appropriation. *TESOL Quarterly*, 28(2), 273-292.

- Ricento, T. (2002). Considerations of identity in L2 learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 895-910). New York: Routledge.
- Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces*, 68(1), 147-177.
- Rock, J. L. (2007). The impact of short-term use of Criterion on writing skills in ninth grade. Retrieved from: <http://www.ets.org/Media/Research/pdf/RR-07-07.pdf>
- Rose, D. (2008). Writing as linguistic mastery: The development of genre-based literacy pedagogy. In R. Beard, D. Myhill, J. Riley & M. Nystrand (Eds.), *The SAGE handbook of writing development* (pp. 1-19). London: SAGE Publications Ltd.
- Scheeler, M. C., & Lee, D. L. (2002). Using technology to deliver immediate corrective feedback to preservice teachers. *Journal of Behavioral Education*, 11(4), 231-241.
- Séror, J. (2005). Computers and qualitative data analysis: Paper, pens, and highlighters vs. screen, mouse, and keyboard. *TESOL Quarterly*, 39(2), 321-328.
- Shermis, M., Koch, C., Page, E., Keith, T., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, 62(1), 5-18.
- Shermis, M. D., Shneyderman, A., & Attali, Y. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy & Practice*, 15(1), 91-105.
- Sheyholislami, J. (2012). Linguistic minorities on the internet. In K. St.Amant & S. Kelsey (Eds.), *Computer-mediated communication across cultures: International interactions in online environments* (pp. 235-250). IGI Global.

- Silva, T. (1990). Second language composition instruction: developments, issues, and directions in ESL. In B. Kroll (Ed.) *Second language writing: Research insights for the classroom* (pp. 11-23). New York: Cambridge University Press.
- Stracke, E. (2007). A road to understanding: A qualitative study into why learners drop out of a blended language learning (BLL) environment. *ReCALL*, 19(1), 57-78.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Todman, J. (2000). Gender differences in computer anxiety among university entrants since 1992. *Computers & Education*, 34(1), 27-35.
- Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327-369.
- Truscott, J. (1999). The case for “The case against grammar correction in L2 writing classes”: A response to Ferris. *Journal of Second Language Writing*, 8(2), 111-122.
- Truscott, J., & Hsu, A.Y. (2008). Error correction, revision, and learning. *Journal of Second Language Writing*, 17(4), 292-305.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind - A quarterly review of psychology and philosophy*, 59(236), 433-460.
- Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, 28(2), 97-111.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157-180.

- Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, 9(1), 27-55.
- Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335-353.
- Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.
- White, E. M. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass Publishers.
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291-300.
- Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283.

Appendices

Appendix A - Detailed Description of Criterion

Criterion is an online computer program which provides guidance for and evaluation of essays written by both native English speakers and ESL (English as a Second Language) students (Education Testing Service, 2011a). The program is expressly designed to help students learn writing skills. The Criterion program was developed by the company ETS (Education Testing Services). According to their website (Education Testing Service, 2011b), ETS administers more than 50 million achievement and admissions tests each year. This includes TOEFL (Test Of English as a Foreign Language) — one of the most widely used English language proficiency tests (used by many colleges, universities and employers). ETS employs a large number of people to evaluate the written essay part of the TOEFL test. Originally, each essay was evaluated by two different people. Now, the company uses just one human plus the AWE computer program called e-rater® to do the evaluations (Enright & Quinlan, 2010). The Criterion program is actually composed of two parts — e-rater to provide an evaluation of student essays and Critique which provides feedback to students (Burstein, Chodorow & Leacock, 2004). Figure 9 illustrates the numerous features which are used by e-rater to evaluate a student essay. It is interesting to note that while ESL teachers today tend to emphasize Organization and Development in an essay, Figure 9 shows that the e-rater system devotes a very small percentage of its work to evaluating these (probably because of the difficulty involved). This is an important issue to consider when looking at the perceptions of Criterion.

Figure 9 The decomposition of e-rater into features (Enright & Quinlan, 2010)

The e-rater software makes use of a database of thousands of previously evaluated essays. All of these essays are sorted based on the prompt they are answering and their

holistic score from 1 to 6. When e-rater evaluates a new essay, it compares (using its numerous features) the new essay to all the existing essays for that prompt and decides to which holistic score it most closely corresponds. The early versions of e-rater relied on this database to evaluate essays and therefore would only work with prompts for which it already had a large database. The newer version of e-rater (version 2.0, released in 2006) also evaluates essays for completely new prompts (Attali & Burstein, 2006). This allows teachers using Criterion to create their own prompts for their class. ETS makes various claims about the validity of the rating portion of Criterion; however, the company does make it clear that Criterion, by itself, is not sufficient to teach writing. They point out that Criterion is "intended to be used under the instruction of a classroom teacher" (Burstein et al., 2004, p. 27).

The Critique portion of Criterion provides feedback to the students in five different areas: grammar, usage, style, mechanics, and organization and development. After a student submits their essay online, the program immediately provides a holistic score and feedback in the five categories as shown in Figure 10. The student can select the "View Grammar Results" link to see a graphical representation of the grammar results as well as the highlighted errors within the text (shown in Figure 11). Mousing over the highlighted text gives more details about the error. Similar screens of feedback are available for Usage, Mechanics, and Style. The student can also select the Organization and Development tab to see a colour coded breakdown of the organization of the essay (shown in Figure 12) including parts such as Thesis statement, Main Ideas, Conclusion, etc. Mousing over the highlighted text gives more details about that part of the text.

ETS Criterion **Writing Feedback Center**

Overall Score: **4 out of 6** [View Score Analysis](#)

Look carefully at the grammar, usage, and mechanics feedback for this essay and check the "Writer's Handbook" for each feedback message. Working on the grammar, usage, and mechanics in the essay may help you improve it.

Trail Feedback Analysis

Grammar:	6 errors	View Grammar results
Usage:	6 errors	View Usage results
Mechanics:	12 errors	View Mechanics results
Style:	2 comments	View Style comments
Organization & Development:		View Organization & Development comments

Figure 10 Immediate feedback in Criterion (<http://www.ets.org/criterion/ell/demo/>)

Figure 11 Grammar feedback in Criterion (<http://www.ets.org/criterion/ell/demo/>)

Trail Feedback Analysis Menu

Organization & Development

Use the color key on the left to identify each element in your essay. To view elements one by one, select **Show Individual Elements**.

[View Question](#)

All too often many feel the necessity to do what others do, be it choosing cars, clothes, colors, restaurants, refreshments, and recreation. The most obtrusive with Urb is this part of the essay your thesis? The purpose of a thesis is to organize, predict, control, and define your essay. Look in the **Writer's Handbook** for ways to improve your thesis.

☐ Show individual elements
☒ Show all elements

[View Score Analysis](#)
[Print Combined](#)
[Close Report](#)

Remember, for more information, click on the **Writer's Handbook** link for each feedback message.

Figure 12 Organization and Development feedback in Criterion

Appendix B - Sample of Transcription Data

Table 3 below shows an excerpt from the transcription of the oral interviews. The first column (Interview Number) indicates that this is the interview with Student B. The second column shows the time (in minutes and seconds) since the beginning of the interview at which the dialogue took place. The third column represents the speaker (either the interviewer or the student) while the fourth column (Text) shows what was said. The fifth column (Video Notes) is a description of what was occurring on the computer screen as the dialogue took place. The sixth column shows the first level (open) coding of that piece of dialogue. The final column contains the second level of coding (axial) assigned to that piece of dialogue.

Table 3 Sample excerpt of transcription data

Interview Number	Time-stamp	Speaker	Text	Video Notes	Coding Level 1 (OPEN)	Coding Level 2 (AXIAL)
B	2:18	Interviewer	What kind of mistakes?... just click on grammar...	Student clicks on "Subject-verb agreement". Essay appears showing the phrase "When in Rome do as the Romans do" with "Rome do" highlighted in blue.	Finding a highlighted error that isn't quite an error.	FDBCK10.2.1, TECH8.1
B	2:22	Student B	(long pause)... Just... ohhh... Oh yes...so in here, I should write "does"...	Student mouses over "Rome do". Pop-up displays "The subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb."	Student thinks that the verb should change to "does" which is incorrect - it actually just needs a comma.	FDBCK10.2.1, TECH8.1

Interview Number	Time-stamp	Speaker	Text	Video Notes	Coding Level 1 (OPEN)	Coding Level 2 (AXIAL)
B	2:42	Interviewer	Ummm... "When in Rome, do as the Romans do". Not really... I think that's correct... So... so that... that's a mistake that's... that's not really a mistake... The program's wrong			
B	2:55	Student B	Ohhhh		Indicating mild surprise that the program might be wrong.	UNDER5.2.2
B	2:58	Interviewer	And the program says they may... MAY not agree...			
B	3:01	Student B	Oh yeah..			

Appendix C - Student Written Questionnaire

Please fill in the following anonymous questionnaire about the Criterion writing program.

1. Your Age (in years): _____
2. Your Gender: ☐ Male ☐ Female
3. Which ESL writing level are you currently enrolled in? ☐ Level 7 ☐ Level 8
4. How many years have you lived in Canada? _____
5. Outside of school work, about how many hours *per day* do you spend on the computer?

6. How many essays have you written using the Criterion online program?

- | | Strongly
Agree | Agree | Neutral | Disagree | Strongly
Disagree |
|---|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 7. I feel satisfied with Criterion's overall holistic scores. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8. I find Criterion's Make a Plan feature to be useful. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9. I find Criterion's Writer's Handbook feature to be useful. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10. I feel satisfied with Criterion's Grammar scores and feedback. | ☐ | ☐ | ☐ | ☐ | ☐ |

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
11. Criterion's Grammar feedback was helpful to me.	☐	☐	☐	☐	☐
12. I feel satisfied with Criterion's Usage scores and feedback.	☐	☐	☐	☐	☐
13. Criterion's Usage feedback was helpful to me.	☐	☐	☐	☐	☐
14. I feel satisfied with Criterion's Mechanics scores and feedback.	☐	☐	☐	☐	☐
15. Criterion's Mechanics feedback was helpful to me.	☐	☐	☐	☐	☐
16. I feel satisfied with Criterion's Style scores and feedback.	☐	☐	☐	☐	☐
17. Criterion's Style feedback was helpful to me.	☐	☐	☐	☐	☐
18. I feel satisfied with Criterion's Organization and Development scores and feedback.	☐	☐	☐	☐	☐
19. Criterion's Organization and Development feedback was helpful to me.	☐	☐	☐	☐	☐
20. The college should continue to use the Criterion program.	☐	☐	☐	☐	☐

Appendix D - Student Interview Questions

The following questions are to be asked in a video recorded interview with selected students at the college enrolled in Level 7 or Level 8 ESL who are using the online Criterion writing program. The interview will be video recorded in the computer Language Lab with each student logged into the Criterion program.

The interview will start with a brief introduction and explanation about the purpose of the interview (as part of a study of the effectiveness of the Criterion program). The student will be reminded that the interview is confidential. Nothing about the student will be communicated to the teacher. The student's name will not be included in any published information about this study. The interview should be no more than 30 minutes.

1. Which ESL level are you currently enrolled in? _____

2. How long have you lived in Canada? _____

3. How many essays have you written using the Criterion online program?

4. Are you satisfied with the marks you have received from the program?

5. Please provide examples of corrections that the program suggested for you.

6. Did you disagree with any of the corrections that the program suggested for you? [If possible, show examples on screen]

7. Compared to your normal way of writing an essay (i.e. without Criterion), how much time does it take to write an essay using Criterion? Does it take a longer time, a shorter time, or about the same amount of time? (Not elapsed time, but the time spent working on the essay)

8. How good a job does the program do of evaluating your spelling? [Display Spelling Feedback page on screen]

9. How good a job does the program do of evaluating your grammar? [Display Grammar Feedback page on screen]?

10. How good a job does the program do of evaluating your essay content? [Display Content Feedback page on screen]

11. How good a job does the program do of evaluating your writing style? [Display Content Feedback page on screen]

12. What things have you learned from the Criterion program?

13. What suggestions would you make to someone else who is about to start using the program? [Demonstrate on screen, as necessary]

14. What, if any, help for using the program did you receive in advance? [Demonstrate on screen, as necessary]

15. What, if any, additional help do you think should be provided for others using the program? [Demonstrate on screen, as necessary]

16. Can you provide some examples of things about the program that frustrated you? [Demonstrate on screen, as necessary]

17. Can you provide some examples of things about the program that impressed you? [Demonstrate on screen, as necessary]

18. Do you think that the college should continue to use the program? Why or why not?

19. Do you have anything else you would like to tell me about your experience with Criterion? [Demonstrate on screen, as necessary]

Appendix E - Teacher Interview Questions

The following questions are to be asked in a private oral interview with teachers at the college teaching Level 7 or Level 8 ESL using the online Criterion writing program.

The interview will start with a brief introduction and explanation about the purpose of the interview (as part of a study of the effectiveness of the Criterion program). The teacher will be reminded that the interview is confidential. The teacher's name will not be included in any published information about this study. The interview should be no more than 30 minutes.

1. Which ESL level are you currently teaching? _____

2. Are you comfortable using computers? _____

3. Outside of school work, about how many hours per day do you spend on the computer?

4. Have you written any essays using the Criterion online program? If so, how many?

5. Are you satisfied with the marks that the program provides?

6. Please provide examples of useful corrections that you have observed the program using.

7. Did you disagree with any of the corrections that the program has suggested? Please give examples.

8. Have you used the program's summarization of student essays to help you focus on certain aspects of writing in your classes? Please give examples.

9. Have you provided feedback directly to your students through the program's utilities? Please give examples.

10. Compared to students' normal way of writing an essay (i.e. without Criterion), how much time does it take to write an essay using Criterion? Does it take a longer time, a shorter time, or about the same amount of time? (Not elapsed time, but the time spent working on the essay)

11. How good a job does the program do of evaluating students' grammar?

12. How good a job does the program do of evaluating students' usage?

13. How good a job does the program do of evaluating students' mechanics?

14. How good a job does the program do of evaluating students' writing style?

15. How good a job does the program do of evaluating students' organization and development?

16. What suggestions would you make to someone else who is about to start using the program?

17. Can you provide some examples of things about the program that frustrated you?

18. Can you provide some examples of things about the program that impressed you?

19. Do you think that the college should continue to use the program? Why or why not?

20. Do you have anything else you would like to tell me about your experience with Criterion?