Writing Development in a Study Abroad Context

Elisa Leslie Barquin

TESI DOCTORAL UPF / 2012

DIRECTOR DE LA TESI

Dra. Carmen Pérez-Vidal i Dra. Aurora Bel Gaya

DEPARTAMENT DE TRADUCCIÓ I CIENCIES DE LLENGUATGE

UNIVERSITAT POMPEU FABRA

Acknowledgements

First and foremost I would like to thank my two supervisors, Carmen Perez-Vidal and Aurora Bel-Gaya, for their trust in me and for their guidance and understanding throughout this process. Carmen, you are a wonderful motivator, and I would not have written this dissertation without your encouragement and generosity. I really appreciate all the support you've given me from beginning to end. Aurora, thank you for being so flexible, generous with your time, and for having patience with the many twists and turns I took before arriving at this point.

I would also like to sincerely thank everyone involved in the SALA project, especially the many researchers who volunteered their time to aid in data collection and the student participants who made the project possible. SALA has benefited a great deal over the years from a series of hard-working and competent research assistants: Pilar, Rebecca, Isabel, Margalida, and Teodora, this dissertation owes a debt to all of you. In particular, I would like to thank Pilar and Rebecca for taking the time to help me locate the necessary data, for providing answers and explanations to all my questions, and for generally making my life much easier at the beginning of this process.

Next I would like to thank our former department head, Enric Vallduvi, for his advice, good humor, and support over the years. I really appreciate all you've done for me, Enric. In a similar vein, I would like to thank Teun van Dijk for the various opportunities he has offered me during my time at the UPF, which have always been both challenging and rewarding experiences.

At the beginning of my degree I benefited from a departmental grant to spend a semester studying at the University of Edinburgh. I would like to thank Professor Antonella Sorace for welcoming me, including me in the many interesting projects and research groups under her supervision, and paving the way for an extremely stimulating period of study. I would also like to thank Dr. Albert Costa, of the UPF's Center of Brain and Cognition, for the time he dedicated to helping me with a previous research project, conducted in the wake of my stay in Edinburgh. The skills I acquired during that period improved my research overall and made writing this dissertation much less daunting.

On a more personal level: Holly and Erin, thank you so much for your help with this project; I am lucky to have friends who are so generous, responsible, and reliable (both statistically and more generally speaking).

Margalida, thank you for the many, many hours you have spent helping me think, commiserating with me, and motivating me to continue. You are a great friend, a valuable colleague, and I'm sure you will one day make an excellent PhD director. To Erica and Potato, thank you so much for sharing your home in Austin and providing the perfect writers' colony at a crucial stage. Tim, thank you for the positive energy and optimism you share with me every day, for making life extra fun, and for making sure I don't forget to enjoy it. Finally, to my parents, thank you for a lifetime of support, encouragement, and unconditional love.

Elisa Barquin
Barcelona, September 2012

Abstract

While study abroad at the university level has been shown to benefit many aspects of second language proficiency, little is known about how this affects participants' writing skills. The present study explores writing development in a group of 30 EFL learners over a period of 15-months and compares the progress made in two distinct learning contexts: study abroad (SA) and classroom instruction at home (AH). The learners' writing before and after each learning context is evaluated by trained raters and analyzed quantitatively, using an assortment of computational tools, to determine whether progress is made in the domains of complexity, accuracy, fluency, lexical diversity and sophistication, and cohesion. The learners' writing is also compared, in terms of both quality and characteristics, to the writing of 28 native speakers of English who wrote on the same topic under the same conditions. Results indicate the writing improves significantly after the SA context, and that learners make considerably more progress while abroad than during the AH context.

Resum

S'ha demostrat que les estades a l'estranger a nivell universitari són beneficioses en molts aspectes per a millorar la competència d'una segona llengua. Tanmateix, no se sap gaire sobre com afecten l'habilitat d'escriure dels participants. Aquest estudi investiga el desenvolupament de l'escriptura en un grup de 30 aprenents d'anglès com a llengua estrangera durant un període de 15 mesos. Alhora compara el progrés en dos contextos d'aprenentatge diferents: les estades a l'estranger i la instrucció a l'aula al país d'origen. S'avalua l'escriptura dels aprenents abans i després de cada context d'aprenentatge, d'una banda, per mitjà d'un grup d'avaluadors experts i, d'una altra, mitjançant un conjunt d'eines computacionals per a determinar si hi ha progrés en els dominis següents: complexitat, correcció, fluïdesa, diversitat i sofisticació lèxiques i cohesió. També es compara, en termes de qualitat i característiques, amb l'escriptura de 28 parlants nadius d'anglès que van escriure textos sobre el mateix tema i en les mateixes condicions. Els resultats indiquen que l'escriptura millora significativament després de l'estada a l'estranger i que els aprenents progressen més quan són a l'estranger que no pas en el context d'instrucció a l'aula al país d'origen.

Table of Contents

Table of Figures

Glossary of Abbreviations

AdCon: Additive connectives
AG1k: Advanced Guiraud 1000
ACTR: American Council of Teachers of Russian
CrefP: Anaphor overlap
AWP: AntWordProfiler 1.200
CrefA: Argument overlap
AH: At-home learning context
BNC: British National Corpus
CausCon: Causal connectives
C-M: Coh-Metrix
CEFR : Common European Framework of Reference for languages
CAF: Complexity, Accuracy, and Fluency
CELT: Comprehensive English Language Test for Learners of English
CLAN: Computer Langauge ANalysis program
Con: Connectives
CLIL: Content and Language Integrated Learning
CrefC: Content word overlap
DTCL: Department of Translation and Language Sciences
DC/S: Dependent clauses per sentence
EAP: English for Academic Purposes
ETS: Educational Testing Services
EFL: English as a Foreign Language
ESL: English as a Second Language
ESOL: English for Speakers of Other Languages
EFS: Error-free Sentence
EFT: Error-free T-unit
PROFILE: ESL Composition Profile (Jacobs et al., 1981)
ECP: ESL Composition Program (Jacobs et al 1981)
ECPE: Examination for the Certificate of Proficiency in English
L1: First/native language(s)
FI: Formal Instruction
ALLENCAM: *Grup d'Adquisició de Llengües a la Catalunya Multilingüe*
GI: Guiraud's Index
IM: Immersion learning context
IGP: Initial lexico-Grammatical Proficiency
IWL: Initial Writing Level
IELTS: International English Language Testing System
iBT: Internet-Based Test
ICC: Intra-class Correlation
L2SCA: L2 Syntactic Complexity Analyzer
LCP: Langauge Contact Profile
LFP : Lexical Frequency Profile

LGP: Lexico-grammatical Proficiency
LogCon: Logical connectives
MLC: Mean length of clause
MLS: Mean length of sentence
MELAB: Michigan English Language Assessment Battery
MLAT: Modern Language Aptitude Test
MLA: Modern Language Association
NAEP: National Assessment of Educational Progress
NES: Native English Speaking
NS: Native Speaker
NNS: Non-native Speaker
HyN: Noun Hyponymy
NP: Noun Phrase
SYNNP: Number of modifiers per noun phrase
OPI: Oral Proficiency Interview
POS: Part of Speech
PAU: Prova d'accès a la Universitat
Temp: Repetition of tense and aspect between adjacent sentences
RD: Resident Director (of SA program)
SLA: Second Language Acquisition
L2: Second/non-native language
S: Sentences
SD: Standard Deviation
SEM: Standard Error of Measurement
SALA: Stay Abroad and Language Acquisition
StrutA: Structural similarity of adjacent sentences
SA: Study Abroad
TempCon: Temporal connectives
TOEFL: Test of English as a Foreign Language
TWE: Test of Written English
TTR: Type/token ratio
UB: University of Barcelona
UCLES: University of Cambridge Local Examinations Syndicate
UIB: University of the Balearic Islands
UPF: University Pompeu Fabra
UWL: University Word List
HyV: Verb Hyponymy
W: Words
WAC: Writing Across the Curriculum

Introduction[1]

Every year hundreds of thousands of university students around the globe embark upon study-abroad programs, often taking a hiatus from their regular academic studies in order to immerse themselves in the language and culture of a foreign country. In the European context, study abroad exchanges of this nature receive considerable economic support through the ERASMUS program, which funds exchanges between European institutions of higher education and is a key element of the larger European aim to cultivate mobility within its borders and widespread multilingualism among its citizens.

In the 2010-2011 academic year, more than two hundred thousand European students from more than 32 different countries participated in the ERASMUS program, spending an average of 6-months studying abroad. Spain, where the present study was conducted, sent the largest number of students abroad (36, 183) and also received the most students (37, 433), reflecting the influence of ERASMUS on higher education and its importance in our context.[2]

One of the most common goals of study abroad participants is to develop or increase proficiency in a foreign language. Indeed, in a 2008 survey distributed to a representative group of 226 Spanish university students, authors Pineda-Herrero, Moreno-Andrés, and Belvis-Pons, found that language learning was the primary academic goal of Spanish ERASMUS participants. Indeed, this link between study abroad and language acquisition is one of the reasons why the ERASMUS program is so generously supported by the EU commission, which has declared multilingualism to be a major priority. While short-term study abroad periods such as those promoted by the ERASMUS program unquestionably have many social and cultural benefits, both for individuals and for the countries that participate, the linguistic benefits are the subject of a substantial and growing body of research in the field of second language acquisition (SLA).

Despite the popular belief that study abroad (SA) is a foolproof method for acquiring proficiency in a second or foreign language (FL)[3], empirical

[1] This research received financial support through HUM2004-05442-C02-01, HUM2007-66053-C02-01/02 and FFI2010-21483-C02-01/02 and ALLENCAM (SGR2005-01086/2009-140) from the Spanish Ministry of Education and the Catalan Government respectively.

[2] Source: "Erasmus-Facts, Figures & Trends" (European Commission, 2012).

[3] Although some researchers make distinctions between these terms based on considerations of multilingualism or context of acquisition, throughout this

research has shown that many factors may dictate the success of this endeavor, such as the length of stay, program characteristics, personality differences, and the degree of proficiency obtained prior to the SA, or 'initial level'. Furthermore, reviews of SA research (e.g. Collentine, 2009; DeKeyser, 2007; Freed, 1998) have demonstrated that the benefits of SA may be highly compartmentalized, affecting some skills and not others, or affecting only certain aspects of a given skill; for example, in studies of speech production there is substantial evidence that SA benefits fluency, but less evidence that it benefits accuracy in speech or phonology. Furthermore, such reviews have shown there is a dearth of research on the ways in which SA impacts language skills such as reading and writing, which tend to be associated with traditional classroom learning contexts but are an important component of overall proficiency, particularly at the stage of higher education when most SA occurs.

The question of whether short-term study abroad experiences will benefit the full range of linguistic skills, is one of practical importance since study abroad is frequently encouraged for language majors and for students who have an academic and professional need for high levels of competence in their FL. European and Spanish students who study abroad in English-speaking countries may assume that their time abroad will allow them to increase their scores on internationally-recognized proficiency exams, such as the IELTS (International English Language Testing System) or those produced by the University of Cambridge. Such exams generally test all four skills—reading, writing, listening, and speaking—and a lack of progress in any one area might leave test-takers with the impression that they have not made the expected progress. Test performance may contribute to the sense of disappointment that DeKeyser (2007) reports is common in SA participants, due to uninformed or unrealistic expectations about language acquisition while abroad.

Of the four skills, writing is often positioned as the most challenging to master. While this may be debated, research has found that writing skills tend to lag behind the other three skills in standardized testing contexts (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981) and that writing is often an obstacle when moving from upper intermediate to advanced levels of proficiency (Brown, Solovieva & Egget, 2010). Such findings most likely reflect the fact that writing in a FL depends upon not only language proficiency but upon writing competence more generally (Cumming, 1989). That is, in addition to the grammatical and lexical competence required for any language production, writing requires an

dissertation we use the terms second language (L2) and foreign language (FL) interchangeably, to describe any and all languages acquired after the native language(s).

understanding of genre and register, of the sociolinguistic requirements associated with each, and an awareness of the relationship between writer and reader (Flower, 1979). In a first language (L1), writing ability is closely linked to the amount and quality of formal education; writing skills are generally taught through explicit instruction in a classroom setting and cultivated through extensive practice. Given this link, one might surmise that learning to write in a FL must also take place in the classroom, and that study abroad participants who do not have explicit writing instruction are likely to return from their semester or year abroad with similar levels of writing ability as upon departure. This seemed to be confirmed by early self-report data cited in Meara (1994), who found that most of the 566 respondents he surveyed (British university students) did not feel their writing had improved during their year abroad; however actual empirical evidence on writing development while abroad is extremely scarce.

Adult language learners may bring vastly different degrees of prior knowledge and skill to the task of learning to write in a foreign language. University students participating in SA have presumably received writing instruction in their native language(s) throughout primary and secondary education; given this former instruction, and since writing is invariably included on exams assessing readiness for higher education, they are likely to have a baseline level of composing competence but may still be hampered by a lack of linguistic competence. For this population of learners, then, it may be that their FL writing skills will benefit more from extensive exposure to the target language, typical in a SA learning context, than from continued study in a traditional classroom setting. Despite the theoretical and practical interest of determining whether FL writing skills improve during SA, and how this compares to improvement at home, very few previous studies have examined this question and even fewer have approached it with methodological rigour and an awareness of FL writing ability in all its complexities.

The present study aims to fill this void, comparing the writing development that occurs in a 3-month SA period with the development that occurs during a previous period of EFL study at the home institution (AH). The overarching goal of the study is to determine whether the SA period is beneficial and what kinds of changes occur in each context. We consider development in a robust sample of 30 learners, using a repeated-measures design, and formulate four research questions in response to our review of the literature on study abroad, writing acquisition, and writing assessment.

The dissertation that follows is divided into two parts and eight different chapters, with the following design: The first three chapters constitute Part 1, and provide the theoretical background for the empirical study conducted in Part 2, which considers qualitative and quantitative changes in participants' writing after SA, in comparison to changes that occur AH.

Chapter 1 further highlights the practical importance of SA research and reviews the growing body of research on language acquisition in this context, particularly focusing on studies that compare SA and AH learning contexts. We then provide a critical review of the small handful of studies which have specifically focused on changes in writing ability after SA, and which have presented mixed results that highlight the need for further research on this specific skill.

Chapter 2 considers the nature of writing ability and explores the relationship between second language writing ability and second language proficiency more generally. We begin by exploring differences between writing and speech, and then review a handful of the influential cognitive models that have been proposed to describe the writing process. Finally, we consider ways in which second language writing has been found to differ from first language writing and review studies that have attempted to quantify the influences of linguistic competence and composing competence on the process and products of L2 writers.

Chapter 3 reviews methods of evaluating progress in writing, referring to both linguistic progress and progress in writing skill in general. This chapter is divided into two parts. The first part considers qualitative assessment of writing and draws upon both SLA and assessment research; we review the tools and procedures commonly used to obtain ratings of writing quality that achieve acceptable levels of validity and statistical reliability. The second part considers quantitative analysis of writing, reviewing the measures used and the theoretical rationales behind these choices. We consider measures in the domains complexity, accuracy, and fluency (CAF), which are often argued to be the three components that describe L2 proficiency, and also consider additional domains for quantitative analysis of writing, such as syntactic variety and cohesion.

The remaining chapters constitute Part 2 of this dissertation and present the design and findings of our empirical study on writing development in SA and AH contexts. Chapter 4 provides a brief introduction to the empirical study and presents our objectives and research questions. After a brief recap of the central issues covered in Part 1, we provide important information about the institutional and geographic context in which the study was conducted, including background on the large-scale research project (SALA) that is the source of our writing corpus. We next state the global objectives and outline the four specific research questions that were formulated to guide analysis.

Chapter 5 describes the design of the study, and gives detailed information on the participants, the learning contexts, data collection, and then the tools and procedures involved in both qualitative and quantitative analysis of participants' writing. First we describe the process of transcription and treatment of procedural issues that may have implications for analysis, such as spelling and punctuation. Next we describe the process of qualitative evaluation, discussing the rating scale and evaluation procedures used and reporting on intra- and inter-rater reliability. Finally we discuss the process of quantitative analysis, which involved both computational and manual coding of features associated with complexity, accuracy, fluency (CAF), lexical diversity and sophistication, and cohesion. We describe the methods used to analyze characteristics in different domains and the specific measures selected.

Chapter 6 presents the results of statistical analyses conducted to measure the changes in writing after SA and AH learning contexts. Results are organized around our four research questions. These results are then discussed in detail, and in relation to previous research, in Chapter 7. In Chapter 7 we also consider how the observed changes compare to self-report data and conduct some qualitative analysis, considering the longitudinal development of one individual participant in detail and looking at a handful of sample essays or extracts to improve our understanding of the changes at the group level. Finally, Chapter 8 aims to draw some conclusions from our study, to discuss the implications of our results, and to suggest topics for future research based on our findings.

Part I

Chapter 1

The Effects of Study Abroad on L2 Writing

This chapter reviews the trajectory of research on study abroad (SA) as a context of second language acquisition (SLA) and considers the ways in which SA experiences might affect L2 writing, which has been studied relatively less than other skills in this particular learning context. The comparatively little attention devoted to the effects of SA on writing may stem from the close association between writing and formal instruction (Carson, 2001), the notion that SA is not "designed" to benefit literacy skills (Ginsberg, 1992), or simply from the difficulty of assessing writing in comparison to other skills (Weigle, 2002), as we will explore in subsequent chapters. However the question of whether SA is likely to benefit L2 written competence is of practical interest because writing skills become particularly important at advanced stages of proficiency and in higher education, at precisely the moment when many learners choose (or are required) to study abroad. Furthermore, there is a theoretical benefit to considering the development of L2 proficiency in written as opposed to oral data, which has been the primary source of evidence on the linguistic benefits of SA. For example, several widely cited studies comparing the performance of learners in SA and at-home (AH) classroom contexts have suggested that spending time abroad does not facilitate the acquisition of lexico-grammatical competence (e.g., Collentine, 2004; DeKeyser, 1991b); however these studies have focused exclusively on oral production data. Since writers have more time to plan and monitor their production, and since writers are under greater pressure than speakers to produce grammatically accurate and complex language (Schoonen, 2009), analysis of written data may lead us to different conclusions about the effects of SA on lexico-grammatical abilities.

Similarly, the construct of fluency has been investigated extensively in SA research and has shown that SA tends to induce significant gains in oral fluency in terms of both objective measures (such as speech rate and utterance fluency) and subjective measures (as perceived by listeners) (e.g., Freed, 1995b; Segalowitz & Freed, 2004; Valls-Ferrer, 2011). While fluency in writing is somewhat harder to define, and relies more heavily on subjective judgments, measuring the benefits of SA on writing might illuminate the extent to which oral and written fluency are related and draw upon the same cognitive and linguistic resources.

The present chapter aims to provide a focused literature review of research on study abroad and to highlight the extent to which research has been lopsided in favor of oral production data. We pay particular attention to studies of fluency and lexico-grammatical competence (vocabulary growth, or complexity and accuracy in production data), abilities that are theoretically associated with writing competence, as we will discuss in Chapters 2 and 3, and which we are relevant to the empirical study carried out in this dissertation. The chapter is subdivided into five sections. Section 1.1 frames the importance of SA research, particularly in the European context, where SA programs under the umbrella of the ERASMUS program receive considerable political and institutional support largely because they are perceived to promote multilingualism. This section also considers differences between SA programs in Europe and North America, and highlights the need for more SA research. Section 1.2 discusses several methodological issues that became apparent in early SA research and should be kept in mind when conducting and evaluating empirical studies in SA contexts: the need for comparative data and the importance of initial level of proficiency. Section 1.3 reviews empirical evidence on how SA presumably benefits global L2 proficiency, as well as two specific aspects of L2 proficiency: fluency and lexico-grammatical competence, which have primarily been investigated using oral production data. Several studies that have considered these constructs in written data are also briefly touched upon in this section, and then are reviewed in considerably more detail in Section 1.4. In that section we take a closer look at the small handful of studies that have explicitly focused on written performance as a result of SA participation, considering the extent to which their findings are comparable and applicable to the population of interest in the present study: advanced-level EFL students in the European context. Finally, Section 1.5 summarizes some of the main points covered in this chapter and sets the stage for our discussions of writing ability and writing assessment in the subsequent theoretical chapters.

1.1 An introduction to study abroad research

1.1.1 SA in Europe: a path to multilingualism

In his review of SA research from a European perspective, Coleman (1998) points out that in order to understand the importance of SA research in Europe one must consider the political context in which foreign language learning is held to be a crucially important factor for enabling mobility, or the free movement of citizens between EU member states, "a central plank of EU policy" (p. 169). Multilingualism and linguistic diversity have been positioned as core European values since the early days of the European Union, and since the late 1950s there have been intergovernmental initiatives dedicated to promoting Europe's heritage of cultural and linguistic diversity. The importance of multilingualism in Europe has increased dramatically in the 21st century, as greater mobility and economic interdependence have increased the need for European citizens to cross borders and to communicate in languages other than their native tongues. 2001 was declared the "European Year of Languages", in the wake of the Council of Europe's 1997-2001 project "Language Policies for a Multilingual and Multicultural Europe", which included a range of activities aimed at helping governments increase multilingualism at the national level, and initiatives focused on improving (and funding) language learning and teaching in Europe. This project culminated in the launch of the Common European Framework of Reference for Languages (CEFR), which has since been widely promoted as a tool for standardizing language testing and a guide for developing effective methods and materials of language instruction[4].

Within this context of active promotion of multilingualism, numerous EU-funded projects and initiatives have labored to increase opportunities for European citizens to learn and practice foreign languages and have declared the goal: "that every person should be able to speak two foreign languages in addition to their mother tongue"[5]. Study abroad has often been presented as a means towards this goal; for example in the 1998 Council of Europe Recommendation that encouraged member states to "promote widespread plurilingualism[6]…by supporting the development of

[4] Source: http://www.coe.int/t/dg4/linguistic/Historique_EN.asp

[5] Source: http://ec.europa.eu/languages/languages-of-europe/languages-2010-and-beyond_en.htm

[6] Referred to as 'plurilingualism' in many documents in order to distinguish between language diversity at the individual level and at the community level. (In early EU parlance, "multilingualism" is used to describe geographical areas

links and exchanges with institutions and persons of education in other countries so as to offer to all the possibility of authentic experience of the language and culture of others”[7].

The assumed link between study abroad and multilingualism has influenced the increased funding and support for the ERASMUS project, an EU-initiative founded in 1987 which funds the organization of study abroad programs in higher education and facilitates joint curriculum development and the transfer of academic credits between institutions (Coleman, 1998). The popularity of ERASMUS exchanges began a steady increase in the early 21st century and the program was bolstered by the Action Plan for 2004-2006 laid out by the EU’s ‘Multilingualism Policy Unit’ (Beacco & Byram 2007). As of 2010 more than 2.2 million European students had participated in the ERASMUS program, via more than 4,000 institutions in 33 different countries, and the European Council aims to serve 3 million ERASMUS students by the end of 2012,[8] and has set the goal that by 2012 at least 20% of all graduates from European higher education will have spent at least some time studying abroad (European Commission, 2012).

Students who participate in ERASMUS exchanges spend between 3 to 12 months studying outside of their home country, taking advantage of subsidized tuition and access to travel grants and cost-of-living stipends. These exchanges are encouraged by many home institutions, particularly for students specializing in foreign languages and language-related degrees, and reflect the widely held assumption that during SA learners will register significant gains in the second language and that time abroad will move participants towards their goal of increased L2 proficiency. Indeed, an early study by Teichler (1997) indicated that 86% of 3000 ERASMUS students surveyed between 1988 and 1992 claimed that learning a foreign language had a strong influence on their decision to go abroad. Because ERASMUS students’ choice of destination is influenced by their previous degree of L2 proficiency and their academic and career goals, they are often more linguistically advanced than their American counterparts, as we will discuss in the next section.

where multiple languages are spoken while “plurilinguism is used to describe individuals with competence in multiple languages)

[7] Source: Rec(98)6E 17 March 1998

[8] http://ec.europa.eu/education/lifelong-learning-programme/doc80_en.htm

1.1.2 SA in Europe vs. North America

While research interest in other parts of the world has been growing in recent years, a significant portion of the research on study abroad has been carried out at North American universities, and thus many of the studies reported in the sections below are concerned with the linguistic development of American university students participating in optional "semester abroad" programs. Since this population, and their SA experiences, differ in important ways from those of European ERASMUS students (the population studied in this dissertation), it is important to keep in mind the ways in which these differences might affect language acquisition.

In 1998, James Coleman (1998) provided a thorough review of the differences between SA in American and European contexts and, while SA programs and SA research have increased in popularity in the past 15 years, the major differences seem to have held constant. Firstly, he points out that American students tend to take part in relatively sheltered, carefully organized, programs, in which they travel overseas as an intact cohort and participate in most academic and cultural activities as a group. American SA students are often only tangentially affiliated with any host university, and may take courses unrelated to their chosen degrees, such that they may face fewer academic requirements than in the US, and be less concerned with academic achievement (DeKeyser, 2007). In this respect, Ginsberg's (1992) observation that SA programs "are not designed" to benefit academic skills, like reading and writing, may hold true for American students, who may dedicate most of their time to linguistic and cultural pursuits (Kinginger, 2009). In contrast, European students are more likely to travel alone or in smaller groups and to enroll directly in various host universities, facilitated by the ERASMUS program. ERASMUS students tend to complete work directly related to their academic degree, alongside students from the host country, and receive academic credit at their home institution, such that they may remain somewhat more focused on academic achievement than their American counterparts.

Given that ERASMUS students must complete exams and coursework in the language of the host institution, they are generally required to have relatively high levels of proficiency in their selected L2 before embarking on their SA. Due to the greater value placed on language learning and multilingualism in Europe, and the fact that foreign language instruction is often obligatory and begins early in primary education, ERASMUS students have typically received many more years of formal instruction in their L2 than American university students. For example, Coleman (1998) reports that, at the time of his study, ERASMUS students studying in the UK had typically received at least 10 years of formal instruction in English prior to arrival, while British students of French had an average of 8 years of instruction before studying abroad in France, the most popular destination at the time. In contrast, many American exchange students have had no language instruction prior to the university, and only one or two years of university-level study before embarking on SA. While a handful of more competitive SA programs in the US may require higher levels of proficiency and facilitate more direct interaction with host universities, the majority do not, and very few studies in the American context report advanced levels of proficiency for their participants.

Finally, due to the greater value placed on multilingualism in Europe and due to the increasing mobility of European citizens, ERASMUS students may be more motivated than their American counterparts and more likely to perceive proficiency in a foreign language as a valuable or necessary asset (Coleman, 1998). This is particularly the case for EFL students studying abroad in English speaking countries (Kinginger, 2009), the case of participants in the present study. Within Europe, the UK is among the most popular destinations of ERASMUS students (European Commission 2012), reflecting the increasing numbers of students prioritizing English as their primary foreign language and striving to improve their proficiency. Due to the ever-increasing use of English as a global language (Crystal, 1997) and the fact that English skills, especially literacy skills, are becoming increasingly important for academic and professional success, particularly in Europe, EFL students in ERASMUS contexts may be particularly focused on the language learning task, and on more formal, academic competences (such as writing) that may be less relevant for other SA students. While the following sections will review research conducted in both European and non-European contexts, it is important to keep these differences in mind before jumping to generalize across in studies. In particular, the differences in pre-program preparation and motivation might lead to better or more uniform gains for ERASMUS students, for reasons that will become evident in the following sections.

1.1.3 Open questions about SLA in SA contexts

As recently as 1994, Paul Meara pointed out that in the UK context there was surprisingly little empirical research on the linguistic benefits, despite the fact that SA programs were obligatory for many language students, and made the provocative claim that “our current belief in the importance of a year abroad rests on some very flimsy, and largely anecdotal evidence”. He called attention to the economic importance of SA, in the light of the growing popularity of the ERASMUS program, pointing out that SA programs correspond to a "huge investment in human capital, not just for the country but also for the individual students" (p. 38). Across the Atlantic, where SA programs under the name of the "junior year abroad" have been popular for language students since the mid-20th century, there was similarly little research prior to the 1990s, and the only widely cited empirical study prior to the late 80s is a large-scale study by Carroll (1967) that was not explicitly focused on SA, but merely made incidental observations about SA benefits gleaned from a more general study on language proficiency in university students.

While research interest has increased dramatically in the nearly three decades since Meara's critique, there are still many open questions about the precise linguistic benefits of SA, particularly when the literature review is restricted to studies of ERASMUS students in Europe, where empirical studies of SA are even more scare despite the continued and increasing investment of public funds. Furthermore, there is still a disparity between the empirical evidence that has been gathered and the popular conception of SA experiences as a quasi-magical road to L2 proficiency, such that "the literature shows that many students come back from abroad with a certain level of disappointment about their progress" (DeKeyser, 2007, p. 208). That is, there is a widespread belief that SA is an optimal context in which to acquire a foreign language, and that SA participants will dramatically increase their proficiency after even a brief sojourn abroad (Rivers, 1998); in contrast, empirical research has shown that SA contexts may be optimal for the acquisition of certain L2 competencies but not others (Brecht, Davidson, & Ginsberg, 1995), and that gains may be conditioned by a wide range of individual and institutional characteristics, such as length of stay (Sasaki, 2011), type of residence (Rivers, 1998), personality differences (DeKeyser, 1990, 1991a), cognitive differences (Segalowitz & Freed, 2004), and initial levels of proficiency (Brecht et al., 1995), all of which might influence learners' opportunities to interact with native speakers and fully benefit from the SA experience.

The literature review in the following sections demonstrates that although a fair amount of empirical evidence has been gathered, particularly with regards to oral fluency, many open questions remain and more research is needed in order to help manage the expectations of language teachers and SA participants. In particular, more research is needed on how SA experiences affect participants' literacy skills, and particularly their writing, which has been studied less than other skills (see Llanes (2011) for a recent review of SA research on each of the 4 skills). While it is logical that oral production has been studied more than written production, since SA contexts were "designed" to benefit oral skills (Ginsberg, 1992) and thus both researchers and participants may have the highest expectations for these skills (DeKeyser, 2007), the full range of language skills are important in the academic contexts of instruction and assessment that SA participants return to, and thus the question of whether writing skills are likely to benefit from time spent abroad cannot be ignored. SA participants, particularly EFL students in the European context (the population of interest in the present study), may assume that time abroad will raise their overall proficiency and enable them to improve their scores on standardized tests, such as the Cambridge ESOL exams, IELTS, or the TOEFL, which are essential for many of the educational and employment opportunities these learners aspire to; however such tests invariably evaluate proficiency through the full range of L2 skills (reading, writing, listening, and speaking). It is thus important to document the extent to which SA might have a positive or negative impact on all skills assessed, since uneven development across skills may prevent students from advancing in overall proficiency and contribute to their reported sense of "disappointment" in their progress.

1.2 Methodological issues in study abroad research

As briefly mentioned above, empirical research on SA was quite scarce prior to the 1990s, despite the fact that SA participation had long since been actively encouraged, especially for language students. The body of research has been steadily growing since that time and advanced our understanding of language acquisition during SA considerably; however such research has also been improving gradually in quality and methodological rigor. Thus, although we must take the findings of early studies with a grain of salt, as they are often limited in terms of reliability or validity, the critiques leveled against them have proved crucial for moving the field forward and promoting the development of improved research designs. They have also illuminated the wide range of individual differences and independent variables that may limit the generalizability of findings across studies but which are important to keep in mind when comparing and conducting SA research. In this section we will briefly consider two important issues in SA research, the need for comparative data and the role of initial level, which are relevant to the design of the empirical study presented in the second half of this dissertation and which allow us to better interpret and critique the literature reviewed in Section 1.3.

1.2.1 Selection biases and the need for comparative data

As mentioned above, the earliest empirical study cited in reviews of SA research (e.g., DeKeyser, 2007; Freed, 1998) is that of Carroll (1967), who examined the test scores of a large sample of language majors (N=2,782) prior to their graduation from college, in order to evaluate the average attainment of foreign language proficiency. Carroll measured achievement in French, Spanish, German, Russian, and Italian, using the *MLA Foreign Language Proficiency Test for Teachers and Advanced Students* to assess performance in Speaking, Listening, Reading, and Writing. He reports that, across languages, the students who achieved the highest levels of proficiency were those who had spent time abroad, and that proficiency increased with the amount of time spent abroad, such that students who had completed summer programs or brief "tours" performed better than students who had never been abroad, and students who had spent an academic year abroad performed better than the other two groups. Carroll reports scores for the Listening test alone, but he claims that these are representative, and that for the other skill tests "the patterns of results are for the most part very similar to those for the Listening test" (p. 136). Carroll's study is widely cited in SA research and has been influential, since its publication, in maintaining the popular perception of SA as an optimal context for language acquisition or even a prerequisite for advanced foreign language proficiency. Notwithstanding this fact, his study serves to illustrate an important confound in much SA research, which is that there may be qualitative differences between students who choose (or are able to) study abroad and those who do not, and that these might influence ultimate attainment of language proficiency in ways unrelated to context. Given that SA was not the primary focus of Carroll's study, he reports little information on pre-SA proficiency or other variables that might have influenced results, making it impossible to determine the direction of the presumed cause and effect relationship between SA participation and ultimate attainment in terms of proficiency.

Authors such as DeKeyser (1990) and Meara (1994) have pointed out that SA research should theoretically be able to show that the amount of linguistic progress made during SA is significantly different from the amount of progress that might be made during a comparable period of classroom study. While this is a valid point, finding adequately comparable groups is no easy task, such that the results of many early studies—including several in Freed's (1995a) seminal collection of SA research, *Second Language Acquisition in a Study Abroad Context*—are confounded by the same selection biases apparent in Carroll's (1967) large-scale report. The issue of selection bias has been particularly problematic in American contexts, where SA participation is almost always completely voluntary, even for language majors (Coleman, 1998). Language majors are the most frequent subjects of SA research, and since they often choose to study abroad in order to improve proficiency, the students who choose to study abroad (or spend longer periods of time abroad) may be more motivated, more culturally open-minded, or simply more outgoing than those who forgo this option (Freed, 1998). Additionally, there may be socio-economic differences between students who are free to dedicate a year or semester to study abroad and those that must remain behind (perhaps needing to work to support their studies or care for family members). This selection bias alone might explain the greater proficiency achieved by SA participants in a number of comparative studies, and raises questions about some of the empirical data at our disposal. While most recent studies make a valiant effort to obtain comparable SA and AH groups – evaluating pre-program levels of proficiency, motivation, and even cognitive aptitude – this remains a persistent problem in SA research.

One noteworthy attempt to obviate this problem was put forth by Milton and Meara (1995) in their study of receptive vocabulary acquisition during SA. These authors advocated for a repeated-measures research design, in which the same participants are evaluated at different points in time (e.g., before and after SA and AH learning contexts) and thus serve as their own control group. In Milton and Meara's design, since participants are compared against their own past performance, most individual differences are held constant, so that differences in achievement may be presumed to result from changes in the learning context. While repeated-measures designs must consider ordering effects and the effects of changes in proficiency from one time to another, this may be the optimal design for analysis of SA, since participants may vary in so many ways that may potentially influence gains.

1.2.2 Individual differences and the role of 'initial level'

Another widely cited and influential early contribution to SA research came from a group of scholars supported by the American Council of Teachers of Russian and the National Foreign Language Center (ACTR/NFLC), who evaluated proficiency gains in 658 American university students who participated in 4-month SA programs at various institutions in the former Soviet Union between 1984 and 1990 (e.g., Brecht et al., 1995; Ginsberg, 1992). While the ACTR studies include several findings that may be specific to Russian and the cultural context of the former Soviet Union (such as a strong influence of gender on SA achievement), the project collectively constitutes one of the largest and most robust studies of language acquisition during SA and offered many important findings with regards to the individual differences that predict SA outcomes.

The ACTR studies measured changes in listening, reading and speaking, using the OPI (Oral Proficiency Interview) and standardized Listening and Reading comprehension tests developed by ETS (Educational Testing Service, the largest assessment body in the United States). All tests were administered just before and at the very end of SA participation. The primary goals were to evaluate the relationships between skills and skill gains, and to determine the best predictors of gains, looking at a wide range of pre-program measures. These measures included pre-program listening, reading and speaking scores, performance on an ACTR qualifying exam, which tested reading and grammatical knowledge, and on the Modern Language Aptitude Test (MLAT). Additionally Resident Directors (RDs) at each of the SA institutions rated participants on different individual characteristics related to motivation and attitudes. The most important finding of the study was that gains in all areas had strong negative correlations with initial levels, a phenomenon "consistent with a "normal" s-shaped learning curve" and that "as a consequence of these strong relationships looking at the effects of other variables…makes sense only with preprogram levels controlled" (Ginsberg, 1992, p. 12).

Once the effects of pre-program level were controlled, they explored predictors of gains for each skill separately. They found that the best predictor for both Listening and OPI gains was pre-program Reading ability (as well as scores on the ACTR qualifying exam, which tested both grammar and reading ability together). The effect of pre-program Listening scores on Reading gains was significant, though not as large as the influence of Reading on Listening; the only other variables showing an influence on Reading gains were ACRT qualifying test performance (which also tested reading) and MLAT scores. For oral proficiency, they found that OPI gains were predicted by pre-program levels of Reading comprehension, and scores on the qualifying grammar tests; Listening comprehension was not a significant predictor for the whole sample, however pre-program Listening scores became significant for discriminating between learners at higher levels (those who moved between scores of 1 and 2 on the OPI, but not those who moved between 0 and 1). Exploration of different individual characteristics and RD ratings indicated that features such as age, gender, and personality characteristics such as "Willingness to Use Russian" and "Taking Advantage of Cultural Opportunities" all significantly predicted gains in different areas: in particular, students who were rated highly on the cultural opportunities criterion made greater gains across the board.

Overall, the ACTR study influenced the trajectory of SA research in three important ways related to individual differences and initial level. Firstly, it illustrated the importance of individual characteristics and personality differences, and thus the advisability of controlling for factors such as gender, pre-program participation in SA, and affective variables. This finding further highlights the complications of selecting adequately comparable groups in cross-sectional studies and adds support to Milton and Meara's (1995) arguments for repeated-measures designs. Secondly, the ACTR study showed the extent to which development may be variable across skills, and that the influence of reading comprehension on listening and oral abilities appears to be unidirectional. This illustrates the desirability of investigating individual skills in a focused manner, and not inferring global proficiency gains from performance in certain specific areas. Meara's (1994) examination of self-report data from 586 British SA participants added further support in this direction, finding a significant difference across skills, with greater gains perceived for speaking than for other abilities. Finally, the ACTR study provided strong evidence that pre-program proficiency as an important effect on SA outcomes and thus that this factor should be controlled for in SA research. The expectation is that higher levels of preparation in reading and grammar (skills associated with formal instruction and experience in classroom learning) may facilitate gains in other domains.

The finding that pre-program Reading comprehension facilitated gains in Listening and Speaking led the authors of the ACTR study to conclude that "communication skills are most effectively built upon a solid grammar/reading base" (Brecht & Davidson, 1991, p. 16); this conclusion, in turn, laid the groundwork for the "threshold" hypothesis discussed in Collentine (2009). As Collentine describes, there is now a growing consensus that SA participants must have reached a certain 'threshold level' of proficiency prior to going abroad in order to take full advantage of the rich and plentiful input and opportunities for learning. Although this threshold level is not well defined—and the assumption is that once all participants have 'crossed' the threshold, the higher level students will gain relatively less, in accordance with the normal learning curve—the notion that students' gains during SA will be influenced by their initial level of proficiency has been widely accepted and confirmed in various empirical studies over the past decade, such as those by Segalowitz and Freed (2004) and O'Brien, Segalowitz, Freed, and Collentine (2007), who evaluated different cognitive-linguistic abilities (such as lexical access, and phonological memory) and found that the degree of SA improvement in oral proficiency (as measured by the OPI) was affected to at least some extent by prior competence in these areas.

The notion that language learners must reach a 'threshold level' of competence in certain domains in order to progress to more complex or challenging ones is certainly not restricted to SA contexts, but is a hypothesis that permeates many other domains of SLA, including L2 writing acquisition, as we will see in the next chapter. For example, various evidence-based theories suggest that L2 writers must reach a certain degree of L2 proficiency in order to make use of L1 knowledge about writing, and writing strategies learned in the L1 (Sasaki and Hirose, 1996; Manchón, 2009). In terms of SA research, we will simply note here that initial level is an important factor to consider when conducting and evaluating empirical studies, and that the findings relative to beginning or intermediate-level studies may not be generalized to more advanced students.

1.3 The effects of study abroad on linguistic competence

With the exception of the large-scale study by Carroll (1967), there is a paucity of studies that have examined global proficiency in terms of all 4 skills. In the British context, Meara (1994) considered improvement in all 4 skills, looking at questionnaire responses by 586 British language students on a national survey (the Nuffield Modern Languages Inquiry). Meara reports that perceived gains seemed to be quite uneven across skills, with the majority of participants indicating that their speaking and listening abilities had improved substantially but not their written skills; however the utility of these data is somewhat limited, as DeKeyser (2007, p. 209) warns that self-assessment data often has a "distressingly low correlation" with objective tests. Coleman (1998) reports that a longitudinal study by Alderson & Crashaw (1990) found that 17 ERASMUS students made substantial improvement in grammar, listening, reading, and writing on "well-established placement tests" (p. 188), after either one or two terms abroad; however since this study is unpublished it is impossible to evaluate the reliability of these findings.

Several studies, such as the previously mentioned ACTR studies (e.g., Brecht et al., 1995) argued that SA promotes global gains in proficiency but looked at only three of the four skills (listening, reading, and speaking), while an even larger number have attempted to gauge proficiency through listening and speaking alone (e.g., Allen & Herron, 2003; Dyson, 1988). The decision not to include writing in many such studies presumably stems from the greater difficulty and expense associated with writing assessment (Weigle, 2002); however it seems amiss to talk about global proficiency without assessing the full range of skills associated with this construct. Coleman (1996) aimed to quantify the effects of SA on the proficiency of ERASMUS students in a large-scale cross-sectional study in which he compared the performance of more than 7000 language students at 190 different British universities (N = 190), as part of his European Languages Proficiency Survey. He compared the performance of 1st and 2nd year students, who had studied only in formal instruction contexts, with 4th and 5th year students, who had completed a stay abroad, and found that post-SA performance was significantly better; however proficiency was measured using a C-test, a proxy for measure that lacks an oral/aural component and thus, like the other studies mentioned, provides only a partially complete picture. (Coleman's study, while similar to Carroll's (1967) study in certain respects, does not suffer from the selection bias issues since SA was obligatory for all students. It does, however, suffer from a lack of comparative data, since there is no evidence that the students would not have continued to improve over time in continued formal instruction).

Far more popular than studies of global proficiency, or attempts to quantify progress in multiple skills, have been smaller-scale, focused studies that have examined comparative gains in different, specific, dimensions of L2 competence. The vast majority of research has approached this task using oral production data, primarily relying on the role-play or interview portions of the OPI or independent tests modeled after the OPI. Such data has been used to explore gains in a wide range of sub-competencies associated with speaking ability, such as "fluency", as evaluated by both subjective and objective measures (e.g., Freed, 1995b; O'Brien et al., 2007; Segalowitz & Freed, 2004; Valls-Ferrer, 2011); phonological accuracy (e.g., Diaz-Campos, 2004; Mora, 2008); communicative competence (e.g., Lafford, 1995, 2004) narrative abilities (e.g., Collentine, 2004); and lexico-grammatical competence, looking at grammatical complexity or accuracy in isolation, or looking at the interplay of complexity, accuracy, and fluency together as an overall index of proficiency (e.g., Allen & Herron, 2003; Collentine, 2004; Isabelli & Nishida, 2005; Juan-Garau & Pérez-Vidal, 2007; Mora & Valls-Ferrer, 2012). A far smaller number of studies have explored some of these same competencies in written production data, using either argumentative or narrative essays (Freed, So, & Lazar, 2003; Perez-Vidal & Juan-Garau, 2009) while only one series of studies has used written data to explore the development of L2 writing ability in depth, looking at changes in both the writing process and products of learners in relation to SA experiences (Sasaki, 2004, 2007, 2011).

The sections below will review the findings for two specific domains of L2 competence associated with global L2 proficiency: "fluency" and lexico-grammatical competence, which are of particular interest in the present study. While neither of these constructs are simple to define, and thus have been investigated using a variety of different measures and methodological approaches, enough research has been conducted in both areas to give us some clear hypotheses about whether SA is clearly an optimal context of acquisition.

1.3.1 Gains in fluency

The development of oral fluency has been one of the primary interests of SA researchers over the past several decades, motivated by early theories by Stephen Krashen with regards to the differences between instructed vs. naturalistic language acquisition (e.g., Krashen, 1985). Krashen argued that L2 fluency was best obtained from naturalistic, implicit learning, and might be developed merely from exposure to comprehensible input in the target language, given the proper attitude and motivation on the part of the learner. This led researchers such as DeKeyser (1990, 1991a, 1991b) to speculate that SA, a stimulating, highly motivating environment where learners theoretically received constant input in the target language, would be particularly beneficial for fluency, in comparison to contexts of formal instruction (DeKeyser, 2007). While much of the research reviewed below indicates that certain aspects of fluency are, indeed, best acquired in SA contexts—in particular utterance fluency and communicative competence—it is unclear whether some of the other characteristics that influence perceived fluency, such as lexico-grammatical accuracy, may require formal instruction or simply derive from language experience in any context, whether in the classroom or at home. This section will review studies that have focused specifically on perceived fluency and utterance fluency, while studies that have investigated grammatical abilities associated with fluency will be discussed in Section 1.3.2.

One of the earliest studies to document oral fluency gains during SA was that of Lennon (1990), who collected oral production data from 4 German EFL learners before and after a 6-month period abroad, and documented significant improvement in terms of both perceived and utterance fluency. Although the lack of comparative data makes it difficult to determine whether the learners might have made similar improvement with 6-months of exposure to English at home, two comparative studies in Freed (1995a) suggested that Lennon’s findings were valid and that SA contexts are truly superior to AH contexts for promoting oral fluency. Freed (1995b) analyzed OPI data from 30 American students learning French as a second language before and after 16 weeks of SA (n=15) or AH formal study (n=15), evaluating speech samples in terms of both utterance fluency (using a selection of temporal measures and dysfluency indices) and perceived fluency (as evaluated by native speaker judges, using a 5-point scale). She found that in the post-test data the SA participants spoke more quickly and smoothly than the AH participants, with fewer clusters of dysfluencies and longer streams of continuous speech, and were perceived as more fluent by native speakers, confirming the results reported by Lennon (1990) for his smaller group of EFL learners. In the same volume, Lafford (1995) used the role-play portion of the OPI to measure development in both fluency and communicative strategies in 42 students over the course of a semester: 26 SA participants, studying in either Mexico or Spain, and 16 AH participants, studying at home in the classroom. She found that the SA participants, but not AH, participants demonstrated improved oral fluency, in terms of speech rate and self-correction, and greater communicative competence over the course of the semester. In the post-test they were able to use a broader range of communicative strategies, in particular for initiating, maintaining and ending conversations.

While both Freed and Lafford's early studies were under-descriptive in terms of the characteristics of participants and vulnerable to some of the critiques related to selection bias, their results were confirmed in more methodologically rigorous series of studies published in a 2004 special volume of *Studies in Second Language Acquisition* dedicated to comparing SLA in SA and AH contexts. In this volume, both Lafford (2004) and Segalowitz and Freed (2004) looked at data collected from 46 American undergraduates studying Spanish in two learning contexts, who were carefully selected for comparability: 26 students spent a semester studying abroad in Alicante, Spain (SA group), while 20 students completed a semester of formal study at the University of Colorado (AH group). In pre-tests administered at the beginning of the study, both groups scored in the low-intermediate range on the OPI, and in comparable ranges on the Spanish SATII test, though the AH group scored slightly higher on the latter. (This same data set was also used by Collentine (2004) to explore lexico-grammatical gains as we will discuss in the next section). Lafford (2004) did not consider utterance fluency in her follow-up study, but did confirm that SA participants appeared to have less need for conscious strategies to bridge communication gaps (due to a lack of L2 knowledge) and to resolve interactional problems in dialogue, which she attributed to the SA group's increased fluency and narrative abilities.

Segalowitz and Freed (2004) were interested in general oral proficiency gains, as measured by OPI scores, and a range of utterance fluency measures, and were also interested in seeing how these gains were influenced by "language contact" in AH and SA contexts, and by cognitive processing abilities. Language contact was evaluated using the Language Contact Profile (LCP), a measurement tool presented in the same volume (Freed, Dewey, Segalowitz, & Halter, 2004), while cognitive processing abilities considered speed and efficiency of lexical access, and attentional control. With regards to overall gains in oral proficiency and fluency, their results confirmed those of Freed (1995b), showing that the SA participants, but not the AH participants, improved significantly from the pre-test to the post-test in terms of OPI scores, length of turn, speech rate, and showed a reduction in the use of non-fluent fillers and pauses. Thus despite the fact that the SA and AH participants were highly comparable in all domains of oral proficiency on the pre-test, the SA group was significantly better than the AH group on the post-test. Notwithstanding these important results, their additional research questions highlighted the extent of variability that may be expected in SA gains and showed that participants' initial levels of oral abilities, as well as their cognitive processing abilities, had a significant effect on the extent to which they improved over time. Their consideration of initial level in relation to the LCP reports suggested that students with higher initial levels had more opportunities to interact with native speakers and receive input in the target language. Their findings thus add support to the threshold hypothesis discussed in Section 1.2.2 and suggest that it is relevant both because higher level students are more able to comprehend the input in their environment and also because they may be more likely to take advantage of opportunities for extracurricular and social activities and thus spend more time in contact with the target language; notwithstanding this fact, the earlier findings of the ACTR studies suggest that language contact is also influenced by personality differences, such as the ones evaluated through RD ratings.

While virtually all of the previously mentioned studies considered the development of American SA participants who, as previously noted, often take part in more sheltered programs and with an additional component of formal language instruction, a handful of recent studies examining development in Catalan-Spanish speaking EFL learners participating in ERASMUS exchanges shows similar evidence that SA is an optimal context for the development of oral fluency (e.g., Juan-Garau & Perez-Vidal, 2007; Mora & Valls-Ferrer, 2012; Trenchs-Parera, 2009; Valls-Ferrer, 2011). For example, Valls-Ferrer (2011) examined oral production data collected from 30 advanced EFL students (Catalan-Spanish bilinguals) using the repeated-measures design promoted by Milton and Meara (1995) to compare the gains made after comparable periods of FI and SA. She used pre- and post-test interview samples to consider changes in three dimensions of fluency: utterance fluency, using a range of temporal measures and dysfluency indices; perceived fluency, as evaluated by both native and non-native speaking judges; and rhythm, a relatively understudied feature associated with native-like speech patterns. She found that participants made significant gains in all three domains of fluency after the SA context, but made no gains during the previous period of formal instruction at home. In line with the threshold hypothesis and the results of the ACTR studies, she found that learners' initial levels of fluency played a role in their relative improvement and that pre-program fluency had a facilitating effect on SA gains. Using a sample of 19 learners from the same population, Trenchs-Parera (2009) focused on dysfluency phenomena and showed that after SA, but not after AH, participants showed a significant decrease in non-fluent disruptions, such as self-repetitions, pauses, and non-lexical fillers, moving closer to the norms registered for a control group of 10 native speakers measured using the same criteria.

1.3.1.2 Studies with written data

Although very few studies have collected written data from SA participants, they have all touched upon the notion of fluency to at least some extent. Perez-Vidal and Juan-Garau (2009), again looking at Spanish-Catalan EFL learners in the ERASMUS context, considered fluency simply in relation to the number of words per minute (a proxy measure we will discuss in more detail in Chapter 3) as one dimension of a study focused on a wider range of variables in the domains of complexity, accuracy, and fluency (CAF). They examined argumentative essays collected from 37 learners before and after periods of formal instruction at home and SA, using Milton and Meara's (1995) repeated-measures design. They found that words per minute increased significantly after the SA period, but not after the AH period, and concluded that SA was more beneficial than formal instruction for written fluency. In contrast, Freed, So, and Lazar (2003) also considered words per minute as a proxy for written fluency but found no comparative advantage for SA participants in the American context. They collected narrative essays from 30 American students of French, 15 of whom completed a semester of SA in France and 15 of whom completed a semester of classroom study, and considered a range of textual measures alongside qualitative evaluations of fluency.

Freed et al. (2003) also reconsidered oral fluency changes in this study (again using OPI interview samples), so that they could directly compare progress in the two modalities, and for both the essays and the oral data, they had native speaker judges subjectively evaluate "fluency" on a Likert scale and then indicate (in written descriptions, and via a checklist) the factors that they believed had influenced their judgments. Methods for the oral data replicate those reported in Freed (1995b). To evaluate written fluency, 5 native speaker (NS) judges (3 of whom had also evaluated oral fluency) were asked to rate essays on a scale from 1 to 7, (from "not at all fluent" to "very fluent") and told they were free to interpret fluency however they wished (p. 6). As in the earlier study, NS judges reliably detected differences in the oral data and perceived the SA participants to be more fluent than the AH participants in the post-test. In contrast, no advantage of SA was found for writing, and neither group appeared to make any improvement in written fluency over the course of the semester. Freed et al. report that the writing of the AH participants was perceived as more fluent than the writing of the SA participants both before *and* after the semester, and that neither group improved significantly (they report that the AH group's scores appeared to decline slightly, and that the SA group's scores increased slightly, but that changes were not statistically significant), concluding that "nothing seemed to suggest that the SA students' writing was more fluent as a result of their having spent time abroad" (p. 6). All judges said their ratings were influenced by grammatical accuracy, and some of the other factors mentioned were vocabulary (richness, word choice), organization, and complexity of thoughts. They found no significant differences between the AH and SA groups in terms of essay length, a quantitative measure often used as a proxy for fluency in writing, although they did indicate that the SA group's essays increased in length and that the AH group's essays decreased in length.

Finally, Sasaki (2004, 2007) considered fluency, among a range of other variables, in the writing of Japanese EFL students after SA experiences of varying lengths, and compared the development of SA participants with participants who remained at home. As in the previous studies (Freed et al., 2003; Juan-Garau & Perez-Vidal, 2009) fluency was measured as a function of the number of words produced (both overall, and in relation to time, since her participants were not given strict time limits). In Sasaki (2004) she followed a small group of 11 participants, 6 of whom spent time abroad in the US, and found that both the SA and AH groups improved their writing fluency, in terms of length and speed, over the course of 3.5 years. Her findings suggested that SA experiences can promote gains in writing fluency but that gains may be no more dramatic than the gains achieved during formal instruction, further highlighting the importance of obtaining comparative data. It is worth mentioning, however, that the participants in Sasaki's (2004) received intensive, process-writing instruction, in which they spent a great deal of time practicing writing, in both the AH and SA contexts. In a follow-up study, Sasaki (2007) examined fluency changes in 13 participants, 7 of whom spent between 4 and 9 months abroad. This time she found a comparative advantage for the SA participants, showing that students who had been abroad wrote longer essays at a faster rate by the end of the study, while the AH participants did not. In this follow-up study the AH participants did not receive the same intensive process-writing instruction, although they were enrolled in general EFL courses at their home institution. Both of Sasaki's studies, and two subsequent studies (Sasaki, 2009, 2011) have important consequences for understanding how and why SA might promote gains in written competence, and will be discussed in greater depth in Section 1.4.

1.3.2 Gains in lexico-grammatical competence

One of the more controversial areas of SA research has related to whether SA contexts are beneficial, or more beneficial than AH contexts, for the development of lexico-grammatical competence, which has primarily been investigated by considering improvement in lexico-grammatical accuracy and complexity in oral production data. DeKeyser's (1990, 1991a, 1991b) early work is often cited as evidence that SA and AH contexts do not lead to differential improvement in terms of grammar; however these studies did not actually compare the development of the two groups over time, or consider grammatical knowledge per se. Instead, DeKeyser was focused on determining whether SA contexts increased participants' tendencies to 'monitor' their linguistic output[9]. He collected baseline data from 12 students, 7 who spent a semester studying Spanish in Spain and 5 who spent a semester studying in the classroom. At the beginning of the term he tested both groups' knowledge of concrete grammatical features taught in intermediate-level Spanish classrooms, using a discrete-point test, and also collected oral production data. He then considered the extent to which participants speech was grammatically accurate with regards to structures that they already knew, and thus showed evidence of successful monitoring. He found that the AH and SA participants were not significantly different in terms of monitoring behavior at the beginning of the term, though he makes no claims about overall differences in grammatical knowledge; given that the SA participants had either completed or placed out of the Spanish classes that the AH participants were enrolled in, we may assume that they performed better overall. He then collected oral production data from the SA group at two different points over the course of their semester abroad and found no changes, indicating that they were still similar to the AH participants despite having spent time in Spain. Thus while DeKeyser's early studies indicate that SA contexts do not lead to an increased ability to monitor grammatical accuracy in speech, they should not be taken as evidence that SA does not lead to increased grammatical competence, since he makes

[9] The theoretical interest of this question derived from a debate initiated by Stephen Krashen, in a series of studies throughout the late 1970s and 1980s, over the differences between naturalistic and instructed learning (see DeKeyser, 1990 for an extensive review of the debate surrounding monitoring and its role in SLA).

no claims about their accuracy with regards to features not measured on the discrete-point test, and did not administer this test at later points in the study.

1.3.2.1 Receptive vocabulary gains

One of the most robust findings in SA research has been in the domain of receptive vocabulary acquisition. In their influential early paper, Milton and Meara (1995) used measured vocabulary growth in 53 European students studying English in the UK, estimating learners' vocabulary sizes based on their knowledge of words in different frequency band levels. All participants had high levels of English having studied English for at least 6 years prior to higher education. Using their repeated-measures design, Milton and Meara evaluated learners' vocabulary sizes before after a 6-month period of SA and a previous 6-month period of AH study in their countries of origin, and found that participants acquired new vocabulary on average five times faster during the SA than at home. While they did note that there were considerable individual differences, "with some subjects showing huge increases in vocabulary, while others show much more modest gains" (p. 23-24), their results suggest a robust advantage for SA contexts in terms of vocabulary growth. Their study showed that more advanced learners made relatively more moderate gains, which may have stemmed from the normal learning curve discussed in Ginsberg (1992) but which the authors recognize may also have resulted from a ceiling effect associated with the test. That is, since the vocabulary test used measured vocabulary size only up to 10,000 words, and may not have been sufficient for the most advanced participants.

Receptive vocabulary growth was reexamined with improved instruments in Ife, Vives-Boix, and Meara (2000), who collected data from 36 intermediate and advanced British undergraduates who studied abroad in Spain for 1 or 2 semesters. They used a translation task as well as a word association task designed to assess lexical organization (and thus to avoid the ceiling effect associated with tests of vocabulary size). Their study did not include a control group; however they did compare the progress made by participants who spent 1 semester abroad with those who spent 2 semesters abroad, and also considered the effect of initial level when evaluating gains. They found that both the intermediate and advanced learners made considerable lexical progress during their time abroad, and that the time spent abroad had an exponential effect on vocabulary growth for both groups. Specifically learners who spent 2 semesters abroad improved their vocabularies up to 3 times more than learners who spend 1 semester abroad.

Finally, in a more recent study, Dewey (2008) examined vocabulary acquisition in 56 intermediate-level learners studying Japanese in three contexts, SA (n=20), AH (n=22), and intensive domestic immersion (IM) (n=13), for 9-13 weeks. Groups were carefully controlled to ensure comparability in terms of initial level. He used three tests designed to measure the breadth and depth of participants' receptive vocabularies before and after each treatment and found that the SA group outperformed the AH group at the end of the study, receiving significantly higher scores on all three vocabulary tests (the IM group fell somewhere in the middle, performing similarly to the SA group on 2 of the 3 tests, but significantly worse on the third test).

1.3.2.2 Lexico-grammatical competence in speech

Studies that have examined the development of lexico-grammatical competence in oral production have reported mixed results. Collentine (2004), working with the same subjects evaluated in Segalowitz and Freed (2004)—46 American learners of Spanish—considered lexico-grammatical competence demonstrated in interview segments of the OPI, collected before and after AH and SA contexts. He considered lexico-grammatical accuracy (looking at 17 specific morphosyntactic features) and also looked at lexical knowledge in a more general sense, by calculating lexical diversity and sophistication. In the post-test samples he found that the two groups performed similarly in terms of overall accuracy; however the AH group performed significantly better than the SA group on a handful of specific features, indicating that the AH context was more beneficial for grammatical accuracy than the SA context. In order to evaluate whether the decrease in accuracy in the SA group was due to qualitative differences in their discourse, he conducted post-hoc analyses considering features associated with "narrative ability", tagging characteristics associated with narrative discourse in Biber's (1988) taxonomy, such as past-tense verbs and third-person morphology, and then calculating a "narrative score" for each sample. He found that the SA group produced more narrative discourse than the AH group, and showed a statistically significant increase over time, not seen in the AH group. Collentine suggested that the SA group's "apparent disregard for accuracy" (p. 240) should thus be interpreted in relation to their production of more complex and discourse-appropriate speech. In terms of lexical abilities, he found that the AH group again outperformed the SA group, showing greater gains in their production of new word types and producing significantly more unique adjectives (though not more unique words in general) in the post-test. He also looked at the 'semantic density' of speech samples, again using Biber's taxonomy to code features associated with informationally rich discourse, such as nouns and adjectives, and then calculating an "informational-richness" score. He found no significant difference between groups for this feature, although he noted that because the SA group demonstrated greater fluency and produced more speech in the given time frame, their speech was more semantically dense overall.

In contrast to Collentine's (2004) findings, several studies that have examined the acquisition of specific morphosyntactic features have suggested an advantage for SA over AH contexts. Isabelli and Nishida (2005) for example, examined the development of the Spanish subjunctive in subordinate clauses, measuring the frequency and accuracy of production in an oral interview task based on the OPI. The authors considered longitudinal development in a group of 29 SA participants who spent a year studying abroad in Barcelona and had completed 4 semesters of university level Spanish, or the equivalent, prior to their stay. Data was collected at 3 different times: Month 0 (prior to the SA); Month 4 of the SA; and Month 9 of the SA. The authors also collected comparative, cross-sectional, data from two groups of AH learners: 16 students at the end of their 5th semester of university-level Spanish, and 16 students at the end of their 6th semester of university level Spanish. Isabelli and Nishida found no differences in the use of the subjunctive when comparing the AH learners in the 5th or 6th semesters of formal study. In contrast, they found that SA participants showed significant improvement in the use of the subjunctive from Month 0 to Month 4, and then showed continued, though more moderate improvement, from Month 4 to Month 9.

Howard (2001) also examined the development of specific grammatical features—past tense verb forms—in 18 Irish learners of French with advanced levels of proficiency. Oral production data was collected from 3 groups, 2 groups had comparable amount of formal instruction but 1 group had SA experience while the other group did not. A 3rd "control" group had no SA experience but had a further year of formal instruction. The group sizes were too small for statistical analysis and group selection may have been confounded by selection bias issues; however Howard reports that learners with SA experience used past tense forms more accurately, and in a wider range of contexts, than learners who had comparable amounts of formal instruction. The difference between Howard and Isabelli and Nishida's (2005) findings, on the one hand, and Collentine's (2004) findings, on the other, may be related to any number of methodological differences between the two studies and the number of factors considered; however they may also have been related to the differences in proficiency levels, a factor of interest in the present study. That is, while the former studies considered advanced learners, with more developed syntactic abilities, the latter study evaluated learners with intermediate levels of proficiency, who may not have reached the threshold level of syntactic knowledge that would have enabled them to take advantage of the SA experience for grammatical development.

In two studies of advanced-level EFL learners in the European context, Juan-Garau and Perez-Vidal (2007) and Mora and Valls-Ferrer (2012) considered improvement in grammatical accuracy in the oral production of Spanish-Catalan speakers, both within larger studies considering the interplay of complexity, accuracy, and fluency measures, and both using the repeated-measures design favored by Milton and Meara (1995) to compare progress in periods of FI and SA, respectively. Juan-Garau and Perez-Vidal collected data from 12 participants using a role-play activity and considered accuracy in terms of the overall production of grammatical and lexical errors, as well as the relative number of errors per clause. Although statistical analysis was limited by the small sample size, and improvement was not significant in either context, their data showed that participants showed improved accuracy after the SA context, with decreased numbers of both grammatical and lexical errors in their speech, but not after the FI context. Mora and Valls-Ferrer examined accuracy in oral interview data collected from 30 learners, measuring the proportion of errors per 'AS-unit' (Analysis of speech units) and the proportion of error-free AS-units. They found that participants' speech improved in accuracy over time and after each context, but that gains were relatively much greater after the SA context, and that participants showed a significant increase in the percentage of error-free units after the SA context, but not after the FI context. Both of these studies suggest that more advanced learners, such as is often the case of ERASMUS students, and particularly for EFL learners in this context, tend to show improved grammatical accuracy in speech. Taken together with the other studies in this domain, it seems likely that the threshold effect is particularly relevant for the development of lexico-grammatical abilities, and that only learners with relatively advanced initial levels of syntactic competence may be expected to make further progress during SA contexts.

1.3.2.2 Studies with written data

Both Freed et al. (2003) and Perez-Vidal and Juan-Garau (2009) considered features of accuracy and complexity in their previously mentioned studies of written production data. Freed et al., in comparing the narrative essays of their 30 SA and AH learners, considered the length of T-units, a measure associated with syntactic complexity, and 3 textual measures of accuracy: error-free T-units; correct noun-adjective agreement, subject-verb agreement, and past tense usage. They found no differences between the SA and AH groups for any of these measures in either pre- or post-test essays. These results are in accordance with the results reported for fluency, which did not show any advantage of SA over AH contexts, or show evidence of development for either group over the course of a semester, since they note that the 5 NS judges all claimed that features of grammatical accuracy had influenced their fluency ratings, and that 4 of the 5 judges also mentioned issues concerning vocabulary (richness, and word choice).

Perez-Vidal and Juan-Garau (2009) also considered changes in accuracy, measured as the number of errors per word, and changes in 3 measures of syntactic complexity: clause length, dependent clauses per clause, and the ‘coordination index’, or the proportion of coordinate clauses to independent clauses (these and other CAF measures, are reviewed in detail in Chapter 3). Overall, they found that there were no significant changes in the essays of their 37 participants, after either context, and no evidence of qualitative differences between contexts. They also examined lexical diversity, by means of a Guiraud’s Index (a type/token ratio that attempts to minimize the effects of text length), and found that participants used a wider variety of lexis in their essays after the SA, but not the AH, context. These findings suggest that the SA may have facilitated lexical acquisition, in line with the findings reported for receptive vocabulary growth. Miyuki Sasaki’s (2004, 2007, 2009, 2011) studies do not examine changes in lexico-grammatical competence in isolation, although in each of these studies she shows that SA experiences have a positive effect on overall essay quality, which she measures using a popular analytic scale by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981), which asks raters to consider features of accuracy and complexity in their judgments.

1.4 Research on writing in SA contexts

As illustrated in Section 1.3, studies of global changes in proficiency, and of changes in fluency and lexico-grammatical competence, have been heavily lopsided in favor of oral production. As mentioned, writing performance has largely been excluded from large-scale empirical studies and self-report data has suggested that SA participants do not feel that their writing benefits as much as other skills. In Meara’s (1994) analysis of the Nuffield questionnaire data, he found when participants were asked to evaluate their improvement in different skills on a scale from 1 to 5 (with 1 indicating the least improvement and 5 indicating the most improvement): “Most respondents felt that their ability to speak the language and to understand the spoken language had improved a lot: 75% of the respondents rated their improvement in spoken language ability at 4 or 5; and 87% rated their improvement in listening skills at the same level.

For reading, however, only 49% of the respondents rated their improvement this highly, and for writing skills, the number of respondents who claim to have improved a lot drops to 33%. In fact, over 30% of respondents rated their improvement in writing skills as low or negligible (p. 34). Meara further notes that only a third of the 586 were required to do any written work while abroad, and that home institutions largely did not required L2 written work upon return, extrapolating that: "the overall impression one gets from these figures is that writing in the foreign language is not a priority either for the students, their teachers at home, or for those responsible for them during their period abroad" (p. 34).

In the years since Meara's observations of the self-report data, only a small handful of studies have made an attempt to evaluate his findings with empirical research on written development. Indeed, the previously mentioned studies by Freed, So, and Lazar (2003), Perez-Vidal and Juan-Garau (2009), and Sasaki (2004, 2007, 2009, 2011) are the only published studies that have considered written production data after SA and AH contexts, and their results have been somewhat mixed, which is unsurprising due the differences in the populations studied and the methods employed.

Freed, So, and Lazar's (2003) study is the one most frequently cited in the literature, and is used to argue that SA does not benefit written fluency in the same way as it benefits oral fluency, the conclusions arrived at by the authors themselves; while this study constituted an important contribution to the field, as the first to examine empirical data on writing development, it is methodologically limited in many respects, and thus their findings must be taken with a grain of salt. Firstly, the sample size is quite small for meaningful between-groups comparisons and the range of proficiency levels is reportedly quite varied (ranging from "a few months to 9 years"), yet no information is given on how levels of proficiency were distributed between the two groups. Secondly, the authors do not report either inter- or intra-rater reliability for writing fluency scores, which one suspects would have been quite low, given the vagueness of the scale used and the variety of features mentioned in the follow-up questionnaires. Finally, several details—such as the increase in text length and lexical density after SA and not AH—suggest that the SA context did have a positive effect that was measurable in writing, despite the authors' conclusions that SA was not conducive to "fluency" in writing in the same way as seen for speech; however the lack of important information (such as the conditions of the writing task, or whether quantitative measures considered the effect of text length) makes it impossible to determine whether these positive effects are reliable. (The authors indicate that "more detailed versions" of their research was presented in earlier papers—Freed, So and Lazar, 1998;

Freed, Lazar, and So, 1999—however these are neither published nor publicly available for evaluation).

Juan-Garau and Perez-Vidal's (2009) study in the European context showed that fluency and lexical diversity improved significantly after the SA context and not after the AH context, which suggested that the SA context was more beneficial to the writing of their more advanced learners. These interesting findings contrasted with those of Freed, So, and Lazar (2003), and set the stage for further research exploring the nature of quantitative changes in L2 writing, laying the groundwork for the present dissertation. In particular, the two studies raise the question of the extent to which quantitative changes in fluency might be indicative of global improvement in L2 writing, as perceived by trained readers. While text length is often a good indicator of text quality, and is a common index of 'fluency' in writing, Freed et al. reported that although their SA participants wrote longer texts than their AH peers, the native speaking judges actually perceived them to be less fluent. It would be interesting to know whether the increased fluency observed after the SA context in Perez-Vidal and Juan-Garau's study might be evidence of improvement in the overall quality of their texts, or whether their texts would be perceived as more fluent by trained raters. Additionally, although Perez-Vidal and Juan-Gaurau did not find significant differences in the domains of accuracy and complexity, the increased proficiency observed in other domains raises the question of whether a larger or more focused range of measures might unearth further signs of development: for example, while the authors did not observe significant changes in subordination or coordination, this might have been due to the fact that their learners had advanced levels of proficiency and improvement in the domain of complexity in advanced-level writing has been argued to mainly affect clausal complexity or phrasal elaboration, as opposed to subordination or coordination (Norris & Ortega, 2009), as we will discuss in Chapter 3.

The only studies to date that have looked at the effects of SA on writing beyond the level of textual characteristics were the four studies carried out by Miyuki Sasaki, in the Japanese context, in a series of longitudinal studies with overlapping research goals and evolving hypotheses (2004, 2007, 2009, 2011). Her studies are highly complex and methodologically rigorous and thus deserve detailed attention, although the extent to which her results may be applicable to different populations, such as the ERASMUS students in this study, are unclear. The four studies built upon her earlier work aimed at creating a comprehensive model of second language writing for Japanese EFL learners (Hirose & Sasaki, 1994; Sasaki & Hirose, 1996; Sasaki, 2000, 2002), which are reviewed in the next chapter. Her participants were all Japanese undergraduate students majoring in British and American Studies, who had studied English for 6 years prior to beginning their degree and had received little instruction in L2 writing prior to entering the university. Sasaki (2004) followed 11 participants for 3.5 years, collecting written data at 4 different points and evaluated changes in writing quality, as measured by an analytic scale, and fluency, as measured by text length as a function of time. During their freshman year all participants had two terms of intensive process-writing instruction, and in the remaining years all had regular EFL classes at their Japanese university. After their freshman year, 6 of the participants completed SA programs of varying lengths (2-8 months) in the US or Canada, and reportedly took both EFL and composition courses while abroad. Sasaki found that both the AH and SA significantly improved their L2 writing from beginning to the end of the study, and that there were no significant differences between groups, although the small sample size limited statistical analysis. Sasaki (2007) looked at 13 participants from the same population, 7 of whom spent time abroad in the US or Canada, with stays ranging from 4-9 months. In contrast to the first study, the participants did not have any writing-specific instruction in their freshman year. Statistical analysis was again limited, but in this study she found that the SA group significantly improved their mean composition scores over time, while the AH group did not, such that the SA groups' scores were significantly higher than the AH group's scores at the end of the study.

The differences between these two studies suggest that the writing progress made by both groups of participants in the 2004 study, and by the SA participants in the 2007 study, was primarily due not to the context of learning per se, but to the amount of formal instruction received. In interviews collected after each period of data collection, Sasaki's participants all attributed their writing progress to writing classes received either at home (in the first study) or abroad. The SA participants in Sasaki (2004) reported that during their time abroad they "were forced to write much and often in English" (p. 556). This report of extensive writing practice during SA, even for the 2 participants who spent only 2 months abroad, stands in contrast to Meara's (1994) report indicating that British SA participants did very little written work while abroad, and reflects one of the particularities that make it difficult to generalize Sasaki's results to other contexts. That is, since Sasaki's participants studied in North America, where there is a long tradition of "writing across the curriculum" (WAC) (Fulwiler & Young, 1982) and where process-writing is considered a fundamental part of secondary and higher education, their SA experiences seem to have included a great deal of writing practice and focused instruction. While there may be more focus on writing today than in Meara's (1994) report, writing-intensive composition courses are still a rarity in European universities, and so SA participants in European contexts, whether American or ERASMUS students, may not benefit to the extent that Sasaki's reportedly did.

While the results of Sasaki's early studies suggested that any differences in writing products may have been due to the amount of formal instruction received, whether at home or abroad, she did notice qualitative differences in the attitudes and motivation of SA participants in the follow-up interviews and verbal protocol data which she collected along with writing products. She also noticed differences between the participants who had been on shorter SA programs and those who had longer stays. These led her to investigate both length of stay and motivational factors in more detail in her subsequent studies. Sasaki (2011) is a follow-up of Sasaki (2009), using a larger sample but investigating the same variables, and in both studied she considers changes in writing quality (again using ratings on an analytic scale) and changes in motivation, as a function of the length of SA. Sasaki (2011) reports data from 37 participants followed over 3.5 years, who were assessed at 4 times. All participants were assessed at the beginning and end of their freshman year, after which 9 participants remained at home and the remaining 26 had SA experiences of varying lengths: 9 spent 1.5 to 2 months abroad; 7 spent 4 months abroad; and 12 spent 8-10 months abroad. As in Sasaki (2004) all participants had writing instruction during their freshman year in Japan, but after this AH instruction consisted only of regular EFL classes and were not writing intensive; the SA participants again received more writing instruction during their time abroad. She found that the composition scores of all 3 SA groups improved from beginning to the end of the study, but that the AH group did not. The AH group made progress in their freshman year but their scores gradually declined to pre-freshman year levels, again suggesting the importance of focused writing instruction.

Sasaki (2011) did report some interesting results with regards to motivation which indicate that SA alone may have a positive effect on writing ability, since students were more motivated to practice their writing after spending time in the target language community, and better able to create "imagined communities" of English speakers, which increased their motivation to write well. These changes in motivation appeared to last even after the participants had returned from the SA, but to be correlated with the length of their stay. Sasaki reports that in their junior and senior years in Japan all participants had fewer EFL classes, and did not receive any focused writing instruction. The two groups with the longer stays were the only groups to show significant improvement into their junior year, and the group with the longest stay (8-11 months) was the only group that continued to improve into their senior year. In the follow-up interviews evaluating motivation, the students with longer SA stays referenced social and academic communities formed while abroad and demonstrated that they may have internalized the communicative purpose of writing. Sasaki also notes that the students with the longest stays made demonstrated "intrinsic motivation" to write (not directly related to academic or professional success) and a desire to maintain the linguistic proficiency they had gained while abroad. The lasting effects of the SA experience on participants, and the continued improvement seen in participants with longer stays despite the lack of continued instruction upon their return to Japan, indicates that SA may be beneficial to written skills for reasons independent from the explicit instruction reportedly received, because spending time in the target language community gave participants an increased appreciation of the communicative purpose of writing and had lasting effects on motivation. Notwithstanding the suggestive nature of Sasaki's findings, it is important to point out that her final study suffers from a selection bias which, while recognized, may complicate interpretation of results. That is, she reports that the longer SA programs were competitive, and that spots were awarded to participants with the highest scores on the TOEFL exam at the end of their freshman year. Since all participants began their degrees with similar levels of proficiency, the spots in the longer SA programs went to those students who had worked the hardest to increase their proficiency during their freshman year, and thus may have been more intrinsically motivated to begin with.

1.5 Summary: the effects of SA on L2[10] writing

As we have seen in this chapter, there is relatively little research on writing development during SA contexts, in comparison to research on oral skills, and it remains an open question whether we might expect ERASMUS learners to improve their writing proficiency during a period of time abroad when there is no focused writing instruction during that period, as is often the case. Given that ERASMUS learners, particularly EFL students, as is the case of the participants in the present study, may have academic and professional goals that require high overall

[10] In the case of the participants in Perez-Vidal and Juan-Garau's (2009) study, drawn from the same population studied in the present dissertation, the majority of participants were early bilinguals with two native languages (Spanish and Catalan), and thus English might more appropriately be labeled as an "L3" (Cenoz & Jessner, 2000). There is a growing body of research on the ways in which L3 acquisition might differ from L2 acquisition (e.g., Rivers & Golonka, 2009) and, while a full review is beyond the scope of this dissertation, we will reconsider the bilingualism of our participants in Part II when we present the empirical study.

proficiency, of which writing is a component, and specifically depend upon their L2 writing skills, this question is of both practical and theoretical interest. The first study to address writing development in SA contexts (Freed, So, & Lazar, 2003) suggested that SA contexts were not beneficial to this skill; however the participants in this study were North American exchange students who were relatively less advanced that typical EFL learners in the European context, and the methodological problems discussed in the previous section raise the question of whether their results should be generalized to other studies. Sasaki's four studies offer evidence that SA experiences, particularly longer ones, may have lasting effects on motivation (which in turn has lasting effects on written performance); however her studies all consider Japanese learners in North American universities who receive process-writing instruction while abroad. It remains unclear if we should expect similar gains from participants in ERASMUS programs who do not experience the intensive writing instruction reported in Sasaki, and whether mere exposure to the language and to an English-speaking community might have a positive effect on writing proficiency and on motivation. Increased motivation to write might certainly explain the fact that the ERASMUS students in Perez-Vidal and Juan-Garau (2009) wrote significantly longer essays after their 3-month stay abroad, and raises the question of whether their participants' essays were also of higher quality or might have differed in accuracy and complexity if a larger range of measures were explored (this question is picked up in the empirical study presented in Part II of this dissertation). As previously mentioned, one of the reasons why writing has been left out of large-scale studies of SA development (e.g., Brecht et al., 1991) is because of the relative costs associated with assessing writing in valid and reliable ways. This difficulty in part from the very nature of writing ability, a complex construct that draws upon a range of cognitive and linguistic resources, as we will explore in detail in the next chapter.

Chapter 2

The Nature of L2 Writing Ability

This chapter explores the construct of writing ability in general and then considers the relationship between writing ability and linguistic proficiency when writing in a second or foreign language. It is divided into four main sections. Section 2.1 provides an introduction to the construct of writing ability, exploring the link between writing and formal education and some of the relevant differences between writing and speech. Section 2.2 then discusses cognitively oriented conceptions of writing ability and presents several influential models that have been proposed to account for the many internal and external factors that come into play during the writing process. The complexity of the writing process, and the multi-disciplinary nature of writing research, is such that no single model can be expected to adequately describe the performance of all writers in all contexts; however the available models have done much to advance our present understanding of writing as a recursive, cognitively demanding, problem-solving activity, and this understanding is necessary in order to develop and evaluate hypotheses about how L2 writers might differ from L1 writers. Section 2.3 explores differences between L1 and L2 writing, reviewing studies that have compared L1 and L2 writing performance from both process- and product-oriented perspectives. Particular attention is given to studies that have expressly considered the relationship between L2 proficiency and writing expertise, disentangling two constructs that are often conflated in L2 writing assessment. Finally, Section 2.4 provides a brief conclusion reflecting on the relationship between L2 composing competence and L2 proficiency, in light of the literature reviewed, and paving the way for our discussion of writing assessment in Chapter 3.

2.1 Introduction: the nature of writing ability

This section introduces the construct of writing ability by discussing the link between writing and formal education and describing some of the salient differences between writing and speech. The goal is to help us appreciate the complexity of the writing process and contextualize the cognitive models described in section 2.2.

2.1.1 Writing ability and formal education

The ability to write is intimately linked to formal education (Carson, 2001; Cummins, 1979; Grabe & Kaplan, 1996). While the goals and structure of writing instruction may vary from culture to culture (see Purves, 1992), children typically learn the basics of their language's orthographic system in the first years of schooling and then continue to develop writing skills for the remainder of their education; in language and writing classrooms students progress, with variable success, from simple narratives to increasingly complex genres, such as those involving argumentation and persuasion (Deane et al., 2008). Providing students with written competence across a variety of 'academic' genres is often a central focus of primary and secondary education, particularly in the English-speaking world, where the ability to write well is closely linked to academic and professional success (Grabe & Kaplan, 1996; Weigle, 2002).

One of the best ways to appreciate the link between writing ability and formal education is to consider how writing differs from speaking in terms of the length of the acquisition process and in terms of variability in the L1 population. Firstly, virtually everyone, barring disability or extreme social deprivation, learns to speak their first language(s) in early childhood, even in the absence of explicit instruction; in contrast, writing must be explicitly taught, usually in academic settings, such that members of the L1 community who never attend school may never learn how to write their language (Sperling, 1996)[11]. Secondly, the degree to which members of an L1 community master the spoken language is relatively uniform. Although there are certainly differences between native speakers in terms of fluency, eloquence, or rhetorical skill (Segalowitz, 2011)—some due to internal, cognitive differences, and others due to

[11] This comparison is reminiscent of Jim Cummins' (1979) early distinction between BICS (Basic Interpersonal Communication Skills) and CALPS (cognitive academic language proficiency), as two fundamental aspects of language proficiency, with speaking falling into the former category and writing falling into the latter.

formal instruction—all native speakers master the morphosyntactic and pragmatic conventions required to process and produce grammatically accurate and complex speech. That is, they become, by definition, maximally proficient speakers of their language. In contrast, the degree to which writing ability is mastered will largely depend on the amount and quality of received instruction and practice (Purves, 1992).

Distinctions between expert and novice writers, and between strong and weak novices, are apparent at all grade levels and even into higher education and the professional world, such that a relatively small portion of fluent speakers will also become "expert" writers, skilled across a range of genres and registers (McNamara, 2010). For example, in the U.S. context, a 2007 study conducted by the National Assessment of Educational Progress (NAEP) assessed the writing ability of a representative sample of 12th graders (27,900 students in 660 public and private schools nationwide) across narrative, informative, and persuasive writing tasks; this study found that only 26% of 12th graders scored at levels above "Sufficient" and that only 5% scored in the highest category "Excellent", across tasks. When these scores were translated into achievement levels based on national standards, only 24% of 12th graders were considered at or above "proficient" in terms of writing ability, representing solid academic performance.[12]

Much of the writing research reviewed in this chapter has been motivated by the observed variability in writing attainment and has aimed to improve writing instruction, and literacy outcomes, by focusing on the differences between expert and novice writers, and identifying strategies associated with good writing that may be explicitly taught in classroom settings. For example, the findings that expert writers tend to make global plans for their writing, instead of simply thinking about what comes next (Bereiter & Scardamalia, 1987); that expert writers revise more at the global level, instead of the word and sentence level (Hayes, Flower, Schriver, Stratman, & Carey, 1987), and that expert writers are more aware of their real or imagined audiences than novice writers (Flower, 1979), have led to recommendations for instructional techniques that explicitly train students to make global plans (for example, by outlining) or to focus on the intended reader (Sperling, 2001).

The variability in written competence has also led many researchers to consider the effects of linguistic, cultural, and ethnic diversity, and to draw attention to the socio-cultural component of writing. The link between writing and formal education becomes particularly important

[12] http://nces.ed.gov/nationsreportcard/writing/

when one considers theories of writing as a social act, and the notion that acquiring written competence entails being socialized into a specific community of practice, or a 'discourse community' (see Grabe & Kaplan, 1996; Johns, 1990). When we speak of writing ability in acquisition and assessment contexts, we are generally speaking of the type of writing that is valued in academia, and writing expertise has been described as a 'key to entry' in the academic discourse community (Weigle, 2002, p. 17). The traditional modes and genres taught in writing classrooms are viewed as heavily conventionalized, contextually determined, structures that are imbued with a complex set of customs and expectations.

One of the goals of formal education is to familiarize students with the genres valued in the academic discourse community, and to arm them with a repertoire of linguistic and discourse knowledge that will allow them to participate in academic arenas where writing is used both to transmit and transform knowledge (Bereiter and Scardamalia, 1988). It is thus important to keep in mind that academic writing ability is derived, in large part, from exposure to academically valued genres and from explicit feedback and practice that raises students' awareness of the expectations and requirements of specific tasks or assignments. A number of researchers have argued that certain aspects of academic discourse can be transferred from speech, but only when students have been socialized in 'standard English' and exposed to oral genres with greater requirements of elaboration and explicitation, such that certain ESL learners and members of the L1 community may be at a relative disadvantage and require even more explicit intervention and instruction than 'mainstream' students (see Sperling's 1996 review article for an extensive discussion of this issue and a review of the specific findings). The link between writing ability and formal education, the variability in L1 writing, and the socially-determined expectations for written products will be revisited in Chapter 3 in our discussion of writing assessment.

2.1.2 The relationship between writing and speech

While we have already pointed out several differences between writing and speech with regards to timing and ultimate attainment, there are many other differences at the socio-cultural, cognitive, and textual levels, and the connection between writing and speaking has attracted the attention of teachers, linguists, psychologists, and anthropologists interested in language acquisition and development, and in language production as a social phenomenon (Grabe & Kaplan, 1996; Sperling, 1996, 2001). On the one hand, certain strains of writing research have focused on similarities between writing and speech, the extent to which spoken language can foster written development, and what students need to 'learn and unlearn about language' when they shift from speaking to writing (Sperling, 2001, p. 9). On the other hand, theoretical discussions of writing ability, particularly from SLA or assessment perspectives, have often found it more useful to highlight the differences between writing and speech as a means of contextualizing the relative complexity of the writing process (e.g., Schoonen et al., 2009).

Early research primarily focused on the differences between writing and speech at the textual level and demonstrated that writing is typically more syntactically complex and lexically sophisticated than speech and requires more elaboration, concision, and grammatical accuracy (see discussions in Grabe & Kaplan, 1996; Weigle, 2002). A considerable body of research on register variation, initiated by Douglas Biber in 1988 with his volume *Variation Across Speech and Writing*, has used multi-dimensional analysis to identify a set of factors that reliably distinguish between oral and written text types across a range of genres, registers, and even languages: the literate dimension, for example, is associated with a high frequency of noun phrases and prepositional phrases, fewer verbs, subordinate and dependent clauses, and has been characterized as "phrasal" as opposed to "clausal", while the oral dimension is characterized by high frequency of verbs, adverbs, pronouns, and greater elaboration at the clausal level (Biber, 2012). These findings become particularly relevant for quantitative analysis of texts, as we will explore in the next chapter, particularly from the perspective of CAF, where measures associated with mature or proficient speech have often been used indiscriminately to evaluate L2 writing. In terms of illuminating our understanding of the nature of writing ability, however, it is important to keep in mind that the textual differences between writing and speech—for example at the levels of accuracy or syntactic complexity—stem from differences in the cognitive demands of each mode of language production and in the social contexts in which writing is required and valued (Weigle, 2002).

At the cognitive level, the most important difference between writing and speech derives from the relative isolation of writing, or the need to communicate a message in the absence of a physically present interlocutor (Sperling, 2001; Weigle, 2002). That is, while both writing and speech are modes of communication, speakers tend to communicate with interlocutors who are physically present, while writers communicate in relative isolation. Of course there are cases along the 'speaking and writing continuum' (Sperling, 2001, p. 8) where this maxim is violated: for example, in televised or recorded speech, an interlocutor may not be physically present, while in a note-taking or dictation session, a writer may have ongoing interactions with speakers. In general, however, writing is considered a solitary endeavor and yet this does not mean that the writer does not need to attend to the needs of their real or imagined readers. Indeed, as mentioned in the previous section, an important part of learning to write entails learning how to anticipate the needs and expectations of the reader (Bereiter & Scardamalia, 1987; Sperling, 1996); this ability is so important that the transition from novice to expert writer has been conceptualized by some researchers as a move from "writer-based" to "reader-based" prose (Flower, 1979; Johns, 1990). Experimental research throughout the 1980s, reviewed in Sperling (1996), demonstrated that students who were taught how to consider their audiences wrote better than those who were not, and that specifying a target audience routinely had significant effects on writing performance. In contrast to speech, where a speaker can rely on their interlocutor to provide cues and to help them shape the message as it unfolds, a writer bears the entire burden for shaping the message as they attempt to communicate with their intended or imagined readers. Therefore, although writers are not under pressure to maintain the flow of conversation and typically have more time and cognitive resources free to plan their utterances and retrieve informational and linguistic content from long-term memory (Weigle, 2002), they face the considerable challenge of needing to monitor their language for clarity and coherence, in the absence of explicit feedback.

Finally, because of the academic contexts in which writing is typically used, the highly conventionalized nature of written language, and the greater time available to writers to produce and revise their utterances, writing is generally less tolerant of errors, redundancy, or imprecision (Schoonen et al., 2009; Weigle, 2002). This means that writers must dedicate time and attention to linguistic and discursive considerations that are less relevant in speech, and learn how to effectively manage their limited cognitive resources (i.e., working memory) to attend to information at multiple levels. As Weigle (2002) describes it: "a writer must devote a considerable amount of cognitive energy simultaneously managing several different kinds of information: information about the writing topic, information about the audience, and information about the acceptable forms of written texts" (p. 18). The cognitive demands of writing, and the pressure to produce syntactically and lexically complex and accurate language while considering the global demands of genre and discourse often proves particularly challenging for L2 writers, who have an imperfect command of the linguistic code, as we will discuss in Section 2.3, and may explain why writing skills have often been found to "lag behind" speaking skills and other aspects of L2 proficiency (Cummins, 1979; Jacobs et al., 1981; McNamara, 2010)

2.2 Models of the writing process

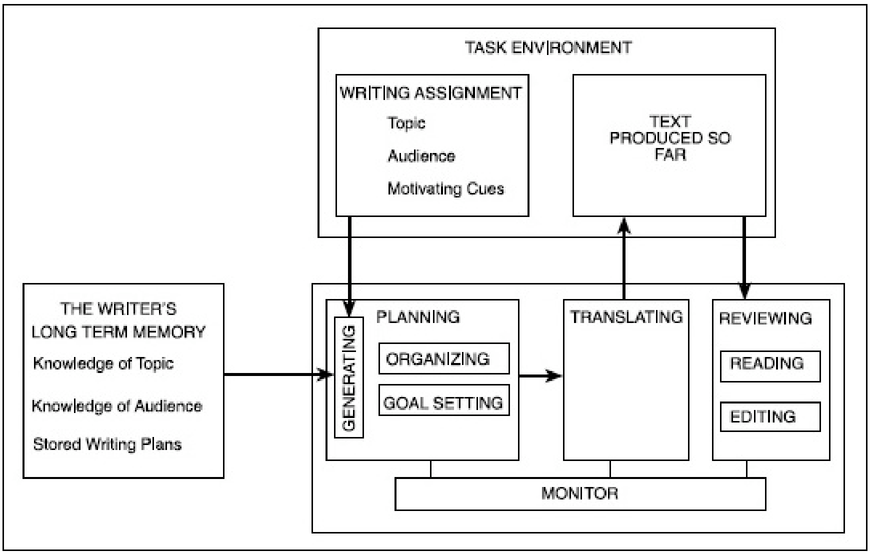
Over the past three decades researchers have attempted to understand the writing process through models or blueprints, which have led to an understanding of writing as a contextually-bound, problem-solving activity that draws upon internal and external resources and invokes a series of interactive cognitive processes. This section reviews some of the most influential models and their attempts to explain how texts are produced and how expert or skilled writers typically differ from novice or unskilled writers. Psychological theories about the writing process can be traced back to the late 70s and early 80s, when researchers began to gather data using think-aloud protocols and related methodologies and concluded that writing entails more than simply verbalizing pre-conceived ideas, and that expert writing, in particular, often involves inventing or discovering new ideas during the writing process as content is tailored to the intended or imagined reader (Bereiter & Scardamalia, 1987). Continuing research in the 1990s also began to highlight the cognitive demands of writing, and particularly the role of working memory, and to consider how the needs to manage limited cognitive resources may influence writing processes and strategies at both expert and novice levels (e.g., Hayes, 1996; Kellogg, 1996). Several influential models derived from each of these stages of research are considered in the following sections.

2.2.1 Early models: writing as a recursive, problem-solving task

The most influential early models of writing were inspired by theories of problem-solving and the "cognitive revolution" in educational research in the early 80s, which posited problem-solving and critical reasoning skills as a central concern of formal education across the curriculum (Sperling, 2001). Early models thus attempted to identify the cognitive processes used to grapple with problems at different stages of writing: when generating appropriate content; when translating pre-verbal ideas into words; when attempting to give those words a coherent structure that may be understood by an imagined reader; or even when attending to mechanical or surface level features such as punctuation and paragraphing. One of the earliest and most enduring models of the writing process to emerge from this perspective was that of Hayes and Flower (1980), which posited the composing process as a goal-directed, problem-solving activity in which writers create and modify their goals during the act of writing, influenced by the task demands and the text in progress (see Sperling, 2001).

The Hayes-Flower (1980) model, shown in Figure 2.1, presents the writing process as consisting of three main cognitive activities: *planning*, *translating*, and *reviewing*.[13] These activities are collectively regulated by the *monitor*, which represents the meta-cognitive control mechanism that helps the writer proceed through the writing process and allocate appropriate amounts of time and energy to each of the individual sub-processes. Planning involves generating and organizing ideas as well as setting goals at the global and/or local levels; translating involves turning these conceptual plans into written language, or converting pre-verbal ideas into words; reviewing involves reading the text produced thus far and modifying or editing it as problems are detected. Each of these activities and the functioning of the monitor are constrained by information obtained from both the *task environment* (the topic, audience, instructions, and the evolving text) and the writer's *long-term memory*, where they store their knowledge of the topic, of the audience, and "stored writing plans", developed through past experience.

Figure 2.1. The Hayes-Flower (1980) writing model



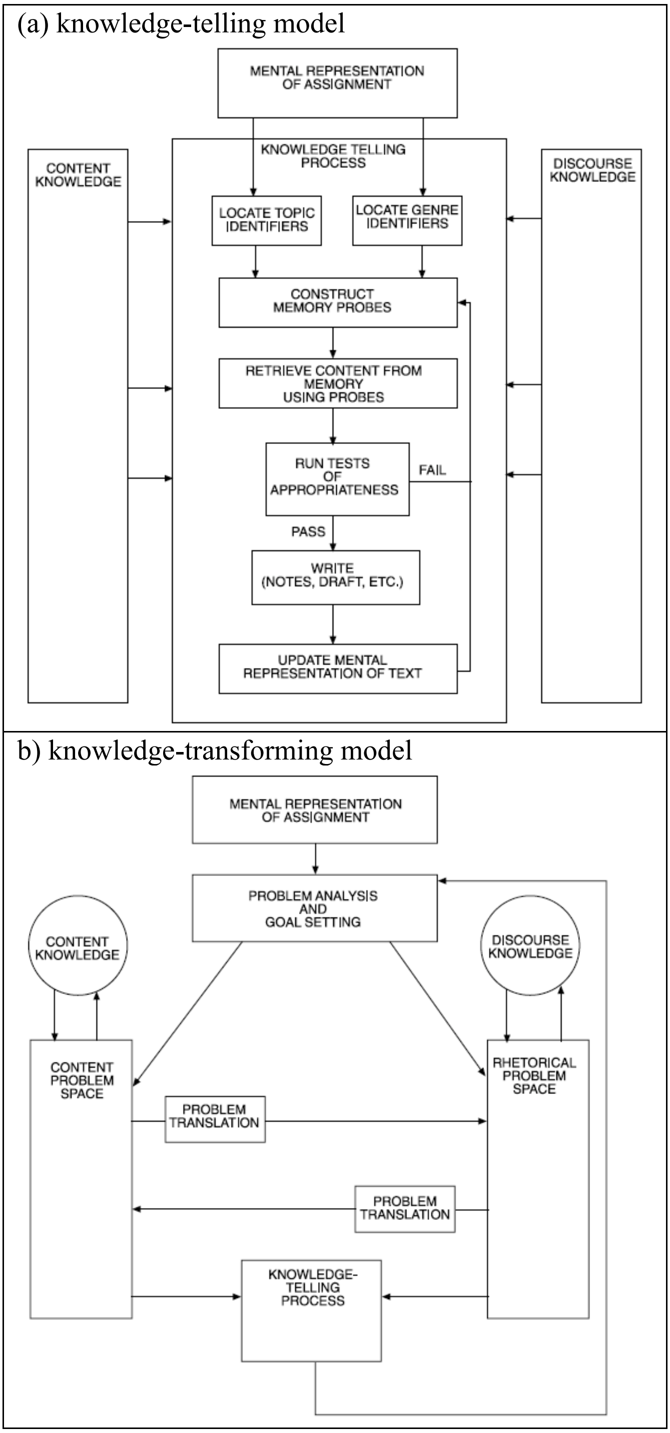
[13] Note that *translating*, in this and subsequent models discussed in this section, refers to the act of translating pre-verbal, conceptual content into language, and not to the act of translating from one language into another. In studies of the cognitive processes of L2 writers, such as those carried out by Rosa Manchón and colleagues at the University of Murcia (see Manchón, 2009), the translating process has alternately been described as *formulating*, most likely to avoid any confusion when multiple languages are involved.

The Hayes-Flower (1980) model was particularly influential as it marked a shift away from thinking of writing as a linear sequence, instead positing it as a set of recursive activities that may be initiated and interrupted throughout the writing process. That is, writers do not necessarily proceed linearly through the stages of planning, translating, and reviewing, but may interrupt one activity at any point to begin or return to another, as various content and rhetorical problems are detected and addressed. Thus planning may occur before, during, or after the production of a given section; translating may be interrupted at any point by revision or further planning. While later research investigating the temporal dimension of writing (see Manchón, 2009) has shown that different activities are more or less likely to occur at different points in the writing process (e.g., more planning tends to occur towards the beginning of the writing process than at the end; more revision occurs towards the end of the process than at the beginning) it has also confirmed that all processes are accessible and may be called upon at any point.

Hayes and Flower and their colleagues used this original model (with minor revisions, e.g., Flower and Hayes, 1981) throughout the 1980s as a lens through which to interpret the differences between expert and novice writers, or skilled and unskilled writers, and to develop a theory of writing expertise (Hayes & Flower, 1986). In their studies Flower and Hayes used composing-aloud protocols to glean that expert writers construct more elaborate goals than novice writers and that they continue to develop and modify their goals during the writing process. While novice writers' goals tended to relate entirely to content, and led them to generate content in response to the topic alone, expert writers tended to also set rhetorical goals, relating to the most effective or appropriate means of expression, and were influenced by their perceived audience. As a consequence, expert writers were found to revise more extensively than novices, and in a more ongoing fashion, as they evaluated their own texts in relation to these rhetorical goals (Hayes, Flower, Schriver, Stratman, & Carey, 1987).

Influenced by these observations, Bereiter and Scardamalia (1987) argued that the multi-faceted problem-solving approach used by expert writers to generate texts with ongoing attention to the topic, audience, and task demands is fundamentally different from the less strategic content-generating approach of novice writers, and that these approaches are best represented by two different models: the *knowledge-telling model* and the *knowledge-transforming model* (see Figure 2.2).

Figure 2.2. Bereiter and Scardamalia's (1987) dual models of writing


(a) knowledge-telling model
MENTAL REPRESENTATION OF ASSIGNMENT
CONTENT KNOWLEDGE
KNOWLEDGE TELLING PROCESS
DISCOURSE KNOWLEDGE
LOCATE TOPIC IDENTIFIERS
LOCATE GENRE IDENTIFIERS
CONSTRUCT MEMORY PROBES
RETRIEVE CONTENT FROM MEMORY USING PROBES
RUN TESTS OF APPROPRIATENESS
FAIL
PASS
WRITE (NOTES, DRAFT, ETC.)
UPDATE MENTAL REPRESENTATION OF TEXT
b) knowledge-transforming model
MENTAL REPRESENTATION OF ASSIGNMENT
PROBLEM ANALYSIS AND GOAL SETTING
CONTENT KNOWLEDGE
DISCOURSE KNOWLEDGE
CONTENT PROBLEM SPACE
RHETORICAL PROBLEM SPACE
PROBLEM TRANSLATION
PROBLEM TRANSLATION
KNOWLEDGE-TELLING PROCESS

Bereiter and Scardamalia argued that the fundamental difference between the writing processes of expert and novice writers concerns the extent to which the retrieval of content is strategically controlled in relation to the writer's rhetorical goals (Galbraith, 2009). In the process of knowledge-telling, novice writers retrieve content from long-term memory, in relation to their understanding of the topic and genre identified with the task, and the organization of their texts stems from associate relationships between this content. Essentially, knowledge-telling is a process that parallels speech production and does not require any more planning or goal setting than ordinary conversation, such that any speaker of the language with a sufficient grasp of the linguistic and orthographic system has access to this approach (Weigle, 2002). The novice writer's primary challenge is generating sufficient content in relation to a given topic without the benefit of a conversation partner or the element of interaction that guides content-production in speech. Expert writers, in contrast, approach the writing task via problem analysis and goal setting, and the generation of content is related to the content and rhetorical goals that they have set with a consideration of the task and the imagined audience. The first step of the knowledge transforming process involves *problem analysis and goal setting*, during which the writer develops a representation of the rhetorical or communicative problem to be solved. As they produce content, drawing on the knowledge-telling process and information stored in long-term memory, they evaluate the content produced in relation to their goals. As they tailor their ideas to the given task and audience, their understanding of the topic may deepen and evolve, which is why this process is referred to as one of "knowledge transformation". As in the Hayes-Flower (1980) model, the cognitive activities involved in writing are presented as cyclical and recursive, and primarily as a problem-solving endeavor; however while knowledge-telling relates simply to retrieving and producing content, knowledge-transforming requires an awareness of the reader and an ongoing process of evaluation and the modification of ideas. Fittingly, Bereiter and Scardamalia (1987) found that expert writers tend to develop more elaborate plans prior to writing and tend to modify and revise their writing far more extensively as they consider their communicative goals.

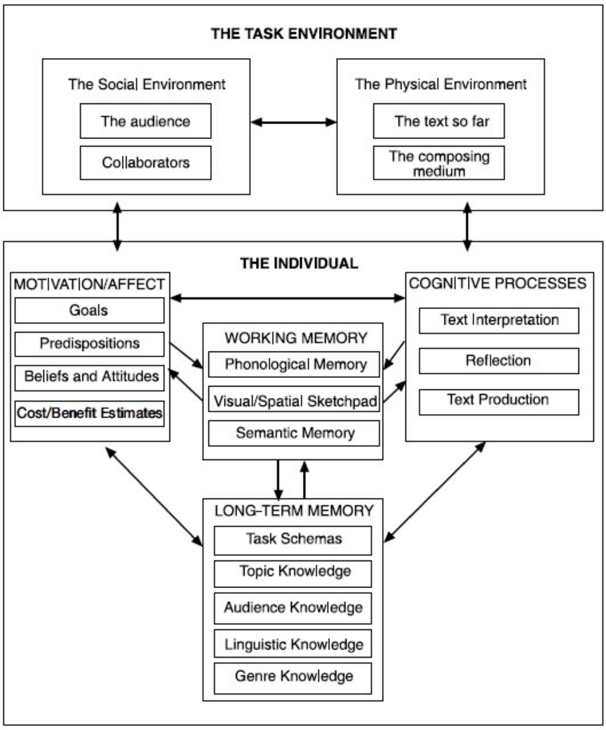
The knowledge-telling model is, of course, subsumed within the knowledge-transforming model, and expert writers are presumed to have "flexible access" to both approaches, "using whichever is appropriate to task demands" (Bereiter, Burtis, & Scardamalia, 1988, p. 262), while novice writers follow the less effortful knowledge-telling model in all circumstances, and do not tailor the content retrieved from long-term memory to the rhetorical demands required by the audience. As Weigle (2002) points out, Bereiter and Scardamalia's models are useful both for understanding the qualitative differences in the processes of skilled and unskilled writers and also for understanding the potential effects of task difficulty on writing performance. That is, two writers of differing writing expertise might perform similarly on a familiar genre of minimal cognitive complexity, such as a personal narrative, which might be completed adequately through the knowledge-telling process alone; however less familiar genres and more demanding tasks will only be completed successfully by writers who engage with the problem-solving processes explicated in the knowledge-transforming model. Theoretically, expert writers have access to both approaches and may activate either as required by the task demands whereas inexpert writers are unable to follow the more complex and effortful knowledge-transforming approach. Therefore when attempting to differentiate between skilled and unskilled writers at more advanced stages of development, such as writers in higher education contexts, it may be preferable to choose a task that demands a higher level of engagement, an issue that we will revisit in Chapter 3.

2.2.2 Revised models: the role of working memory

In Levy and Ransdell's 1996 volume *The Science of Writing* two influential studies continued the trajectory of cognitively-oriented research framing writing as a complex and recursive problem-solving activity, but began to pay attention to the role of working memory and to recognize that working memory is essential for the functioning of the different cognitive processes involved in writing (Becker, 2006; Galbraith, 2009). The concept of working memory in psychology is often used to describe the limitations we face when performing certain tasks, such as holding words, numbers, or visual information in memory in order to transcribe or repeat them; working memory is limited in terms of both the quantity of information it can hold as well as the amount of time it can hold it (Hayes, 2006). Both of the studies in Levy and Ransdell's volume, Kellogg (1996) and Hayes (1996) adopted the model of working memory proposed by Baddeley and colleagues (Baddeley, 1986), which posited that working memory has separate stores for visual-spatial information (*the visuo-spatial sketchpad*) and for verbal information (*the phonological loop*), and that both of these stores are managed by an overriding *central executive* function. Kellogg's study linked these different components of working memory to different activities involved in the writing process, arguing for example that that the visuo-spatial sketchpad is primarily active during the processes of planning and editing, when the writer is concerned with the organization of the text, and is not relevant during text production. While Kellogg's hypotheses were influential to the field, laboratory research has since shown that interference with working memory may negatively impact writing activities in ways not accounted for by his model, and supports a more general involvement of working memory in the writing process (Hayes, 2006). For example, Hayes (2006) reports that articulatory suppression, which affects the ability to store verbal information in the phonological loop, has been found to negatively impact writers' ability to correct errors in their texts during the editing process, although Kellogg (1996) had proposed that the phonological loop is not involved in the editing process. These findings seem to support the more general model proposed by Hayes (1996), which presents working memory, as a whole, as one of the internal variables that guide and influence the entire set of recursive cognitive activities that comprise the writing process.

Hayes (1996) model (see Figure 2.3), improves upon the earlier Hayes-Flower (1980) and, with its inclusion of both working memory and motivation/affect as influences pertaining to the individual, comprises one of the best available models of the writing process. As in the earlier model, Hayes organizes all potential influences on the writing process, but in this model they are categorized as pertaining to one of two categories: *the individual* or *the task environment*, with greater focus on the former.

Figure 2.3. The Hayes (1996) writing model



Under the umbrella of the task environment, he includes both the social environment (the audience/reader and any collaborators) and the physical environment (the text in progress and the medium of writing). Hayes (1996) highlights the importance of the social environment and of the real or imagined interaction between writer and reader in light of evidence that writers behave differently when writing for familiar audiences (e.g., friends or teachers) than when writing for strangers, and the growing understanding of writing as a socially situated, communicative act. The physical environment includes the text in progress, as in the Hayes-Flower model, but also acknowledges the potential effect of the *composing medium* (handwriting or word processing). While the Hayes model was developed before word processing was the norm, researchers had already begun to investigate the cognitive differences between writing with a pen and paper and writing on a computer, where the mouse and copy/paste functions allow the writer to jump around within a text and reorganize elements more freely and the composing medium has since been found to affect virtually all aspects of the writing process (Deane et al., 2008).

The internal influences on the writing process include the writer's working memory, as discussed above, and the writer's long-term memory, including their understanding of the topic, task, genre, and their linguistic and sociolinguistic knowledge, both of which have reciprocal relationships with the writer's cognitive activities and with motivation/affect. The cognitive processes included are *text interpretation, reflection,* and *text production,* which come into play during planning and editing as well as drafting, and Hayes (1996) descriptions of these processes emphasize the important role of reading at each stage, particularly during planning and revision. Text interpretation and reflection involve reading both the prompt and any available instructions, for the purposes of planning, as well as reading the emerging text in order to generate more content and in order to evaluate it and to detect possible problems. At each stage, all three cognitive processes interact with memory (both working memory and long-term memory) and with motivation/affect. The recognition that affective variables influence writers' goals and the amount of effort they are willing to expend on a given task represents an important improvement upon earlier models, and is particularly important to keep in mind when considering the processes of L2 writers. L2 writers' motivation may be affected by their linguistic and sociolinguistic competence, by their perceived relationship to the target language community, as we say in the research by Sasaki (2009, 2011) discussed in Chapter 1, or by the relatively greater effort required to produce text in L2, an issue that we will address in the following section.

2.3 Writing in a second language

Thus far in this chapter we have focused on the construct of writing ability without differentiating between first and second language writers. Early studies of the L2 writing process conducted throughout the 1980s, inspired by the findings and methodological approaches of L1 research, generally emphasized the similarities between first and second language writing (Jones & Tetroe, 1987; Krapels, 1990). Krapels (1990) synthesized the results of more than a dozen of these early process-oriented studies, which invariably used think-aloud protocols to analyze the writing behavior of carefully selected individuals or small groups of participants—generally ESL students in US university contexts, designated as either "skilled" or "unskilled" writers based on their performance on university placement exams or in ESL classrooms. The majority of these small-scale studies found that L2 writers mirrored the behavior of their L1 counterparts: skilled L2 writers showed evidence of global planning and revision, goal-setting, and other characteristics of the "knowledge-transforming" approach proposed by Bereiter and Scardamalia (1987); unskilled L2 writers, like unskilled L1 writers, spent less time planning and revising, and appeared to focus most of their attention on generating content, or on the process of "knowledge-telling". These findings were obtained even in studies where the skilled and unskilled writers had "advanced" levels of L2 proficiency (e.g., Zamel, 1983), leading researchers to argue that a lack of competence in L2 writing was the result not of a lack of proficiency, but of a lack of 'composing competence' or 'metaknowledge'—that is, a lack of awareness of what 'good writing' entails (e.g., audience awareness, discourse organization). Early researchers argued that this 'composing competence' could be transferred across languages, such that first language writing competence is reflected in second language writing performance, and there are no inherent, qualitative differences between the two processes (Krapels, 1990, p. 49).

Notwithstanding the predominant findings of this first era of L2 writing research, scholars like Krapels (1990) and Silva (1993) were quick to point out the dangers of drawing firm conclusions from small-scale and case-study data. In a 1993 meta-analysis of 72 comparative studies of L1 and L2 writers in EFL contexts, Tony Silva highlighted a number of differences that were observed in both process- and product-oriented research. He noted that the research reviewed collectively indicated that L2 composing is "more difficult and less effective" than L1 composing (p. 661), highlighting findings that second language writers expend more effort while writing, particularly during the process of formulation, and yet are routinely judged as less successful, obtaining lower evaluations on holistically scales, and produce texts that are less sophisticated and accurate with regards to syntax, lexis, and discourse organization. Silva pointed out that, despite the desire to gloss over differences, at least some degree of performance loss may be expected in L2 and that although their may be broad similarities between the writing processes across languages, L2 writers face more and greater challenges, whether due to gaps in linguistic or sociolinguistic knowledge or due to more limited cognitive resources (e.g., Jones & Tetroe, 1987; Raimes, 1985). In the nearly two decades since Silva's review, the body of research on second language writing has increased in quantity and methodological rigor and several robust studies with larger groups of participants have aimed to isolate and quantify the differences between L1 and L2 writing and the extent to which these are constrained by linguistic proficiency. A selection of these studies is reviewed in the following section.

2.3.1 Qualitative differences

Despite the overall similarities between L1 and L2 writing and the fact that early research established that many skills may be transferred between the two, later research has aimed to document the ways in which second language writing may differ from first language writing, honing in on specific sub-processes, such as formulation, planning, or revision, or focusing on overarching cognitive activities like the allocation of resources or problem-solving techniques. By definition, second language writers have more limited linguistic resources, which may make the process of formulation more effortful and cognitively demanding, leaving fewer cognitive resources available for "higher order" processes like planning and revision (Galbraith, 2009; Manchón, 2009). An early process-oriented study by Jones and Tetroe (1987) was among the first to suggest that composing in a second language uses greater "cognitive capacity" than composing in a first language and to observe that attention to linguistic issues reduces the amount of time available for planning and attention to higher-order concerns. Jones and Tetroe used a within-writers design to study the composing processes of 6 Spanish-speaking graduate students writing in L1 and L2 and noted performance loss in L2 despite the relatively advanced proficiency-levels of their participants, which they ascribed to these more limited cognitive resources. Since Jones and Tetroe's study, others have highlighted the ways in which cognitive and linguistic limitations may lead to qualitative differences between L1 and L2 writers, often by focusing on specific components of the writing process, and in particular on formulation, planning, and revision.

2.3.1.1 Formulation

Intuitively, the process of formulation, or converting pre-linguistic ideas into words, is one of the areas in which second language writers are expected to have greater difficulty (Galbraith, 2009). In one of the first studies to explicitly quantify this difficulty, using a within-writers design, Chenoweth and Hayes (2001) analyzed the process of formulation by focusing on the production of sentence parts, or specifically on the production of interrupted "bursts" of language while writing. They used think-aloud protocols to analyze the behavior of a group of 13 writers writing in L1 and L2, considering differences between languages and also considering the effect of proficiency on L2 performance. Their writers were all undergraduate university students, native speakers of English, and had either 2 or 4 semesters of classroom instruction in their L2 (French or German). Chenoweth and Hayes then calculated the number of length of *P-bursts* and *R-bursts* in participants' writing in each language. *P-bursts* were defined as chunks of text delimited by pauses but followed by continued language production; *R-bursts* consisted of chunks of text delimited by pauses followed by revision of the language produced, as opposed to production of new material. Chenoweth and Hayes hypothesized that the length of a P-burst depends on the capacity of the "translator" an internal variable limited by the writer's linguistic resources and working memory. Thus the length of a P-burst indicates the amount of language that a writer is able to produce before the capacity limits of the translator are reached. They compared the performance of each participant in L1 and L2 and also compared the performance of students with higher and lower levels of L2 proficiency. As hypothesized, they found that the length of P-bursts in L2 was shorter than in L1, and that greater linguistic proficiency in the L2 led to longer bursts. They also found that writers produced a higher proportion of R-bursts in L2 than in L1, suggesting that they revised more of the language they produced. Together these findings confirm the intuitive assumption that the formulation process is more effortful in L2 than in L1: L2 writers are less able to translate complete thoughts into words, due to limitations in working memory, and are forced to revise more of their language, due to limitations in both memory and in linguistic resources.

While Chenoweth and Hayes (2001) focused exclusively on the writing process, and did not consider the quality of the writing produced, it is easy to imagine how less fluent formulation might negatively impact text quality. Galbraith (2009) points out that the inability to hold longer chunks of language in working memory might may it difficult for writers to keep track of their ideas and express them adequately, even if they possess a coherent understanding of the topic, at the pre-linguistic level, and have an awareness of the task demands. That is, the overall coherence and complexity of the ideas a writer is able to express may depend on their ability to quickly transcribe those ideas into written language, so that they may be combined with other ideas and evaluated in relation to content and rhetorical goals. Given these observations, the fact that even relatively experienced L2 writers produce significantly shorter P-bursts in L2 than in L1 may go a long way to explaining the deficiencies in final products observed by Silva (1993).

In two of more than a dozen empirical studies carried out by a research group based at the University of Murcia (henceforth, the Murcia Research Group), Roca de Larios, Marin, and Murphy (2001) and Roca de Larios, Manchón, and Murphy (2006) also used a within-writers design to consider differences in the formulation process in L1 and L2, but focused more specifically on the nature of the problems writers encounter during formulation. Like Chenoweth and Hayes (2001), they also considered how these differences were manifested at different levels of proficiency. Their participants were 21 Spanish-speaking students at 3 levels of proficiency (and education): 7 high school students with 3 years of English instruction; 7 undergraduate students with 6 years of English instruction; and 7 recent university graduates (English majors) with 9 years of English instruction. In both studies they had participants write argumentative essays in both L1 and L2 while composing aloud and used the protocol data to analyze problem-solving behavior during the process of formulation. In Roca de Larios et al. (2001), the authors observed simply that fluent formulation was more common than problem-solving formulation in both L2 and L1, but that the proportion of fluent formulation was significantly larger in L1, suggesting that learners faced many more problems in the L2 and that grappling with these problems was a major source of their reduced fluency. In Roca de Larios et al. (2006) the authors delved into the nature of problem-solving in more detail, by coding the problems faced as either 'compensatory'—when strategies were used to compensate for a lack of linguistic resources—or 'upgrading'—when strategies were used to improve the lexical, stylistic or rhetorical features of the text. They found that the greater density of problems in L2, which affected writers at all 3 levels of proficiency, was primarily due to different numbers of compensatory problems. That is, compensatory problems were "virtually nonexistent" in the L1 data, but were prevalent in the L2 data, and the number of upgrading problems was similar across languages. They concluded that the formulation process was more effortful in L2 because the writers were faced with more than twice the number of problems in L2 and had to divide their time and attention between solving both linguistic and stylistic problems. With regards to proficiency, Roca de Larios et al. (2006) found that although all writers spent a similar amount of time on problem-solving in L2, the distribution of these problems changed in relation to their level: as proficiency increased the amount of time spent on compensatory problems decreased while the amount of time spent on upgrading problems decreased. This difference was particularly dramatic between the highest and lowest level learners: the lowest level learners spent twice as much time on compensatory problems as on upgrading problems, while the highest level learners spent nine times more time on upgrading problems as on compensatory problems.

2.3.1.2 Planning

Early process-oriented research suggested that writers typically plan less in L2 than in L1 (Krapels, 1990; Silva, 1993), potentially because they require more time and resources for formulation. In another study conducted by the Murcia Research Group—using the same data set described above—Manchón and Roca de Larios (2007) explored qualitative differences in the planning behavior of their writers in relation to both the language used (L1 or L2) and relative proficiency in the L2 condition. They found that, unlike formulation, there did not appear to be qualitative differences in planning behavior across languages (similar amounts of time were spent planning in L1 and L2) but that behavior did change in relation to proficiency, and that the lower level participants showed some evidence of performance loss when writing in L1 possibly due to a decrease in planning and to a greater focus on topic as opposed to organization or the structuring of ideas. These findings suggest that planning behavior might not vary simply due to language background, but instead depends upon writing expertise, or the amount of instruction and practice received in either language (remember that participants not only had different levels of proficiency but different levels of education). Manchón (2009) reports on some of the differences in planning behavior that emerged when participants at different levels were examined separately. She specifically reports that the lowest level participants dedicated similar amounts of time to planning in both L1 and L2, indicating that like the learners observed in Stevenson (2005), their writing skills in L1 may have been relatively undeveloped and thus that basic strategies were easily transferred across languages. The mid-level participants planned more in L1 than L2, suggesting that although they had more advanced strategies in L1, their linguistic proficiency restricted the extent to which they could transfer strategies to the L2. Finally, the advanced participants actually planned more in the L2 than in L1, indicating that their writing proficiency in L2 may have surpassed their writing proficiency in L1, though they were still more advanced than the other participants across languages. These findings will be addressed below when we consider how both L1 competence and L2 proficiency may constrain L2 writing behavior.

2.3.1.3 Revision

Stevenson, Schoonen, and de Glopper (2006), building off of Stevenson (2005), looked at the revision processes of 22 adolescent writers writing in Dutch (L1) and English (L2), collecting writing samples in each language (using argumentative essays) and analyzing revision behavior using both keystroke logging software and think-aloud protocols. They focused on 'online revision', or revision that occurs during the writing process as opposed to after a complete draft has been produced, and considered whether it affected surface linguistic elements or features of content. Revisions were thus coded for orientation (whether they were linguistic or conceptual), domain (the size of the unit revised, at the level of word, clause, or above clause), location (the place in the text), and action (whether the revision consisted of an addition, deletion, or substitution). The results showed that for each of the four dimensions there were qualitative differences between L1 and L2. With regards to orientation, the authors observed there were more linguistic revisions in L2 than in L1; however this did not lead to a corresponding decrease in content-related revisions, and the amount of conceptual revisions in L2 and L1 was highly similar. Similarly, although L2 writers made more revisions below the word and clause levels, more immediate revisions, and more substitutions and deletions, the frequency of the other revisions processes was similar across languages. Overall, Stevenson et al.'s results suggested that the greater frequency of linguistic revision in FL did not lead to the inhibition of other revising processes. That is, the writers appeared to spend more time solving language problems in their L2, but this did not seem to affect their ability to solve content-related problems. Notwithstanding these findings, the authors stipulate that the adolescent writers observed did not make many higher-level revisions in either L1 or L2. That is, perhaps because they were novice writers and did not exhibit many of the more sophisticated strategies associated with high-level writing ability, they had less difficulty maintaining their writing behavior across languages. Furthermore, the analysis considered frequencies alone, and did not attempt to quantify the amount of time spent on different revision processes, a factor that Roca de Larios et al. (2006) and others have found to be important for describing the differences between L1 and L2 writers. Manchón (2009) reports that in the data collected by the Murcia Research Group, for which revision was analyzed as a function of time as opposed to merely the frequency of episodes, the amount of time spent on revision increased linearly with the proficiency level of participants, as did the proportion of revision aimed at the elaboration and clarification of ideas, and at solving discourse and stylistic problems, as opposed to revision aimed at correcting linguistic errors.

2.3.1.4 Allocation of resources

Stevenson (2005) examined the relationship between fluent “bursts” of language and the writing process, similarly to Chenoweth and Hayes (2001) but also considering the ways in which fluency affected each of the different writing sub-processes (planning, formulating, and reviewing) and the allocation of resources among these processes. Stevenson measured fluent bursts using keystroke-logging software and quantified fluency as a measure of the average number of words produced between pauses of 2-seconds or more, although she did not differentiate between P-bursts and R-bursts, or those followed by revision. She also used think-aloud protocols to analyze the amount of time writers spent on the different writing sub-processes (planning, formulating, and reviewing) in both L1 and L2, and considered these data in relation to fluency and in relation to the perceived quality of their written products, as measured by holistic evaluations. As hypothesized, she found that the writers produced more content in L1 than in L2, and that they paid more attention to linguistic processing in L2 than in L1, with more localized reading and more strategies used to solve language problems. She also found that the learners’ conceptual processes seemed to be inhibited as a result of this attention to linguistic problems. In particular, participants spent less time planning in L2 than in L1, and as a result their L2 essays were perceived as less “rhetorically well-developed”. They also spent more time reading localized structures in their L2 essays, and less time on global reading of the entire text, suggesting that there was a “narrowing of focus” when writing in the L2, and that writers dedicated a great deal of time to reading and rereading specific clauses or sentences in order to arrive at an acceptable formulation. Although the writers were less fluent in L2 than in L1, Stevenson did not find any significant relationship between the degree of fluency and the quality of texts, in either language, in contrast to expectations. When taken alongside the findings reported for revision (e.g., Stevenson et al., 2006), it becomes apparent that some inhibition of conceptual processing can occur in L2 writing, but that this may be apparent only when one considers the amount of time, as opposed to simply the frequency, spent on different writing processes. The observation that qualitative differences between L1 and L2 writing may emerge only when the temporal dimension is considered is in line with the principle findings of the Murcia Research Group (see Manchón, 2009).

Manchón (2009) synthesizes the findings gathered by the Murcia Research group in terms of how language background (L1 or L2) and proficiency together seemed to affect the allocation of resources during the writing process, which became apparent only when the temporal dimension was considered (the timing of different writing processes as a function of the total amount of time spent writing). In particular, Manchón reports that the amount of time devoted to the "optional" processes of planning and revision[14] varied with regards to proficiency, as did the extent to which the writing process was recursive as opposed to linear in nature. The lower level learners tended to exhibit a more linear process in which planning episodes were concentrated at the beginning of writing, formulation in the middle and revision at the end, with formulation taking up the greatest amount of time by far. In contrast, the more advanced learners dedicated relatively more time to planning and revision and these activities were more evenly distributed throughout the writing process (though the greatest concentration of episodes still occurred at the beginning and the end, respectively). Manchón points out their results suggest that "as proficiency grows, a more balanced allocation of attentional resources to different processes takes place" (p. 107) and argues that the more advanced writers were able to "strategically decide what attentional resources to devote to which composing activities at any particular point in the writing process" (p. 108). In contrast, the less advanced writers had to focus the majority of their effort on formulation and had fewer resources available to attend to planning and revision while they were generating text.

2.3.1.5 The role of proficiency

Together, recent studies of the L2 writing process suggest that L2 writing strategies *may* be transferred across languages but that L2 writing is clearly a more challenging process, and that writers, especially at lower levels of proficiency, must spend more time on formulation and problem-solving, particularly in order to compensate for linguistic deficits. For example, the adolescent writers in Stevenson et al. (2006) and the lower level learners in the Murcia studies performed similarly across languages, and did not demonstrate writing expertise in either language: they spent little time on global planning or "higher-order" concerns and instead focused most of their attention on formulating, on problem-solving and revision related to language use. While the two groups of more advanced writers in the Murcia study claimed to be concerned with the organization of ideas, as opposed to merely content, and demonstrated an awareness of what 'good writing' entails, the differences between these two groups

[14] Manchón argues that formulation, or the act of generating text, is the only non-optional writing activity, and that this may consume virtually all of the time and attention of novice writers (2009: p. 120).

indicated that writing expertise may only be transferred across languages once a certain level of proficiency (or writing expertise) has been reached. That is, only the highest level group showed evidence of 'expert' strategies, such as larger proportions of global planning or problem-solving related to upgrading concerns, across languages. It seems likely that the mid-level group, despite greater awareness of the requirements of good writing, were less proficient in their L2, and that this lack of linguistic resources forced them to spend more time on formulation and left them with fewer resources available to put their writing expertise to good use.

2.3.2 Linguistic proficiency vs. writing expertise

The studies reviewed in the previous section suggest that differences in L1 and L2 writing may depend upon the extent to which linguistic proficiency allows writers to demonstrate their writing expertise across languages. They also suggest that writing expertise plays a role (in that 'unskilled' writers perform similarly across languages, but skilled writers are constrained by proficiency), an issue that has been explored more explicitly by studies that have grappled with the extent to which both of these variables influence L2 written products. One of the earliest such studies was conducted by Alister Cumming in 1989. Cumming compared the writing processes and products of 23 French-speaking university students in their L2 (English) on writing tasks of varying cognitive complexity (a letter, an argumentative essay, and a summary task). He considered learners with three levels of "writing expertise", as determined by self-reports, their performance in L1 (based on holistically evaluated essays), and whether or not they had professional writing experience in French. 10 of the participants were classified as "basic" writers, 8 were classified as "average student" writers, and 5 were classified as "expert" writers. Approximately half of each group was classified as having intermediate English proficiency, while the other half had advanced English proficiency (assessed by university faculty via oral interviews). Cumming analyzed the written products (evaluated with an analytic scale) and the writing processes (using think-aloud protocols) to compare groups and also to explore the relationship between writing expertise and L2 proficiency. He found that both factors accounted for large portions of variance in the quality of written products and in the problem-solving strategies the students used while composing, and that there were no significant interactions between writing proficiency and L2 proficiency, suggesting that these two factors made different contributions to the writing process and written products.

Cumming (1989) reports that the expert writers behaved like expert writers, regardless of proficiency: they made greater use of heuristic search strategies for evaluating and solving problems, attended to multiple aspects of their writing when making different decisions, and produced texts that were evaluated as significantly more effective in terms of content and discourse organization. The experts' writing strategies appeared to hold constant across the different tasks, but were more prevalent in the more cognitively demanding task (argumentative writing), suggesting that they had access to multiple approaches and consistent with the knowledge-telling/knowledge-transforming dichotomy proposed by Bereiter and Scardamalia (1987). He found that the basic writers, in contrast to the expert writers, did not vary their use of strategies across tasks and did appear to create mental models that guided their writing, such that they were always relying on the "what next" strategy (Bereiter & Scardamalia, 1987). Therefore, although the basic writers "were able to produce phrases and sentences with evident fluency in their second language, they had great difficulties conceiving how these phrases would cohere strategically in their overall discourse" (p. 121).

In contrast to writing expertise, which led to qualitative differences in writing processes and text quality, Cumming (1989) found that language proficiency had a purely additive effect: that is, while greater proficiency led participants to improved writing performance and higher-quality texts, it was not found to cause qualitative differences in the thinking or decision making processes of writers. His findings suggested that writing expertise is a central cognitive ability, and that second language proficiency facilitated performance but did not lead to qualitative changes. That is, their L2 proficiency did not determine the principal characteristics of their writing performance, though it had a facilitating effect. He found that writing expertise in L2 closely mapped onto their writing expertise in L1 and that they appeared to transfer the strategies from one language to another, regardless of proficiency. Finally, he found that 'monitor overuse' or the extensive monitoring of second language usage in the domains of grammar, spelling, and punctuation, appeared to have a negative impact on writing performance, although he primarily drew these conclusions from the performance of one participant who was described as a *monitor overuser* (p. 123). He argued that writing expertise is a specifically developed intelligence with unique cognitive characteristics that can be applied across languages and that second language proficiency, in contrast, is specific to each language. Both seem to contribute different elements to second language writing performance. Because the level of L2 proficiency did not have a significant effect on decision-making processes, he argued that writing expertise is "as easily attained in a first or second language" (p. 125). He argued that the instructional implications were such that expert writers might not need instruction in writing in their second language, but would best be served by practice and activities aimed to increase proficiency, while basic or average writers should receive process-writing instruction similar to that aimed at L1 learners.

Sasaki and Hirose (1996) also analyzed the relationship between proficiency and writing expertise (measured via L1 writing ability as well as writing "metaknowledge"), with the goal of developing a model for L2 writing that addressed both factors. They studied L2 writing ability in 70 Japanese EFL students, controlling for educational and cultural background, and attempted to investigate the different factors that influenced the quality of L2 expository writing. All participants wrote argumentative essays in both L1 and L2, evaluated with an analytic scale, and answered questions about the writing process on a discrete points test. On this test, participants were asked about concepts such as topic sentence, unity, coherence, the typical organization of English expository writing, and had to identify the most well-organized paragraph from among 5 alternatives and justify their choice. They considered L1 essay scores, proficiency level (score on the *Comprehensive English Language Test for Learners of English,* or CELT) and metaknowledge score in regression analysis in order to determine which variables best accounted for variation in the L2 essay scores of participants. They found that L2 proficiency explained 52% of variability; L1 writing ability explained 18%, and metaknowledge explained 11%; and that CELT Total, Japanese Composition Total, and Metaknowledge together explained 54.5% of the English Composition Total score variance. Although all of the variables investigated appeared to help explain differences in writing performance, only the CELT total made a significant unique contribution to explaining English composition scores (the other two variables made very little unique contribution—1.5% for Japanese scores, .3% for metaknowledge score), indicating that the correlation between English scores and Japanese scores/metaknowledge largely overlapped with the correlation between the English score and the CELT total.

Overall, Sasaki and Hirose's (1996) results indicated that L2 proficiency plays a major role in explaining L2 writing ability. They also observed a difference between these data and data from a pilot study with higher proficiency writers (Hirose & Sasaki, 1994), which led them to argue that "L1 writing ability might gain greater explanatory power only after students' L2 proficiency has surpassed a certain level" (p.156) and that "the lesser explanatory power of L1 writing ability in the present study suggest that it may not be so powerful in explaining L2 writing ability when the two languages have different rhetorical conventions" (p.156).

Their finding that L2 proficiency and L2 writing ability are related differs from previous studies (e.g., Cumming, 1989), which might be due to the fact that participants had developed both skills primarily in academic contexts and thus were both related to aptitude for academic achievement, as the authors themselves note. Sasaki and Hirose used the results of their study to propose a path diagram type of model to explain Japanese students' EFL writing ability (Figure 2.4), in which composing competence is postulated as a higher-order factor affecting both L1 and L2 writing ability.

Figure 2.4. Model of EFL writing from Sasaki & Hirose (1996), p. 161.

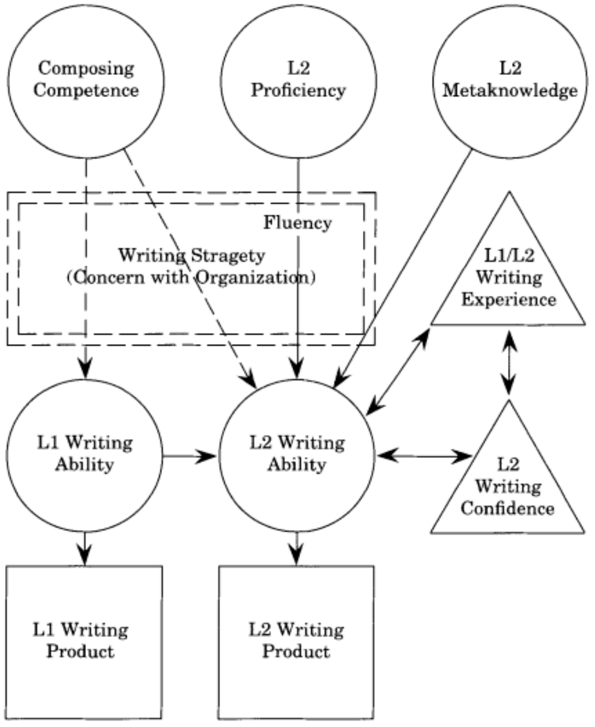


Figure 2. Path diagram illustrating an explanatory model of EFL writing. The measured variables are enclosed in squares, and unmeasured latent factors are enclosed in circles. Latent background factors are enclosed in triangles. The unidirectional arrows indicate one-way causal relations. The straight lines in the arrows are based on the results of the present study, and the broken lines are based on speculation. The process features are represented by broken line rectangles (see Sasaki, 1993).

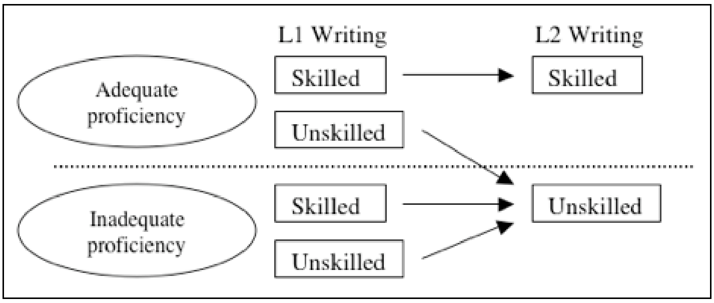
2.4 Summary: The Nature of L2 Writing Ability

As we have seen in this chapter, studies of the L2 writing process suggest that the nature of writing in a second language is similar to writing in a first language, but that it is constrained by second language proficiency, and there is an undefined 'threshold level' of proficiency required in order for learners to transfer their knowledge and expertise across languages (or to make use of metacognitive knowledge about writing acquired in L2). Process-oriented research has demonstrated that while many aspects of the L1 and L2 writing processes are qualitatively similar, the L2 writer must dedicate more time and attention to dealing with local, language-specific problems (Roca de Larios et al., 2006; Stevenson, 2005), and this may inhibit their attention to conceptual problem-solving. These greater challenges demand more of their time and consume resources that might otherwise be available for planning, revising, and attending to rhetorical concerns and the communicative context (Manchón, 2009). Once higher levels of proficiency are reached, however, skilled second language writers make use of the same strategies observed for skilled first language writers (Cumming, 1989) indicating that the two processes are qualitatively similar despite the latter being more effortful.

Research focused on comparing differences across L2 writers of different proficiency levels has shown that the relative difficulty of composing in an L2 decreases in tandem with linguistic proficiency (Chenoweth & Hayes, 2001; Manchón, 2009) but is still constrained by writing expertise in the L1 (Cumming, 1989; Sasaki & Hirose, 1996). The latter two studies, in attempting to partial out the relative influences of writing expertise and linguistic proficiency on L2 products arrived at different conclusions: Cumming (1989) found that writing expertise was relatively more predictive than linguistic proficiency, while Sasaki and Hirose found that linguistic proficiency was more predictive, although both factors played a role. Sasaki and Hirose report that a previous study with higher proficiency writers pointed to a greater role for writing expertise, however, and both groups of participants in Cummings study had relatively high levels of proficiency. Together these studies suggest that linguistic proficiency may be a primary determinant at lower levels (also see Schoonen et al., 2003). They point to a relationship in which writing ability may be transferred only once learners are over a certain basic threshold, as represented in Figure 2.5.

As indicated, once second language writers have obtained sufficient linguistic proficiency, their writing will reflect the writing expertise gained in both L1 and L2 with regards to both the process (Manchón, 2009) and the quality of their texts (Cumming, 1989). In the absence of adequate metacognitive knowledge and writing expertise, however, no amount of L2 proficiency will allow learners to perform like skilled writers, particularly on cognitively-complex tasks that require allocation of resources to planning, revision, and complex problem-solving.

Figure 2.5 Relationship between proficiency and writing skill



In sum, our exploration of second language writing ability in this chapter has given us a better understanding of the many internal and external influences on writing performance and the different factors that might play a role in the extent to which writing skills improve over time, which in turn allows us to improve the methodology of the empirical study conducted in Part II and add theoretical interest to the question of whether SA experiences are likely to benefit participants' writing skills. For example, our review of the differences between writing and speech have helped draw attention to the ways in which these two modes of production vary and why analysis of written progress must consider a different set of features than analysis of speaking progress, particularly when assessing development through textual features. The L1 writing models such as those of Bereiter and Scardamalia (1987) also suggest that evaluating writing progress in advanced-level learners (such as those in higher education contexts, like the participants in the present study), must make use of tasks that are sufficiently complex if one wishes to evaluate written expertise and the extent to which writers attend to higher order content and rhetorical problems. Finally, our understanding of the role that L2 proficiency plays in the L2 writing process allows us to hypothesize that SA experiences will be beneficial to L2 writing even in the absence of formal instruction, despite the mixed results gathered in previous research, and inspires us to use more finely grained analyses to evaluate the location and nature of this improvement. Armed with this greater understanding, in the following chapter we now turn to the practical questions involved in L2 writing assessment, the selection of tools and instruments and the best practices for ensuring the validity and reliability of writing assessment in research contexts.

Chapter 3

Evaluating Progress in L2 Writing

In any empirical study of writing development one is forced to confront a special set of challenges related to the difficulty of assessing writing (routinely recognized as the hardest skill to assess) and of assessing language development in writing. Polio (2001) outlines the set of characteristics that SLA researchers have deemed relevant for measuring changes in L2 writers' texts: overall quality, linguistic accuracy, syntactic complexity, lexical features, content, mechanics, coherence and discourse functions, fluency and revision. Each of these features comes with its own set of open theoretical questions and methodological challenges, as we will see in the present chapter, which is divided into two main parts, described below.

In Section 3.1 we will discuss methods of evaluating the overall quality of a piece of writing, drawing primarily on research conducted in the context of large-scale assessment (i.e., standardized testing). Although SLA research and assessment research do not always coincide in their theoretical interests or objectives, research conducted in the context of large-scale proficiency tests (such as those administered by ETS or UCLES[15]), has the advantage of sample sizes and financial resources rarely seen in SLA research, and as a result has made important headway in improving the reliability and validity of writing assessment. One of the

[15] ETS (Educational Testing Service) and UCLES (University of Cambridge Local Examinations Syndicate) represent the two largest assessment agencies in the US and UK, respectively.

primary difficulties in evaluating the overall quality of a piece of writing lies in the apparent subjectivity of this process, given that there is simply no objective definition of writing quality in either L1 or L2 (Kroll, 1990). That said, experienced teachers and readers of L2 texts have been found to share a common set of intuitions and criteria that allow them to consistently distinguish between "good" and "bad" writing (or between "skilled" and "unskilled" writers), focusing on a combination of syntactic, lexical and discourse features, on the appropriateness and coherence of content, and the writer's awareness of both purpose and audience (Haswell, 2005; Jacobs et al., 1981). While these intuitive methods may be acceptable in classroom contexts, they are clearly inappropriate in "high-stakes" tests where evaluations of performance must be well-justified. Large-scale proficiency tests are considered "high-stakes" tests, in that the decisions made based on test performance have important consequences on participants' lives (Weigle, 2002). In the case of the TOEFL, for example, (the *Test Of English as a Foreign Language*, administered by ETS) a test-taker's score impacts their ability to access North American universities, which may in turn affect their career options and future earning power. The fact that these tests have an important gate-keeping function has kept them under public and institutional scrutiny, which has forced them to develop and maintain high standards of reliability and validity (Polio & Williams, 2009), from which SLA researchers may benefit. That is, researchers interested in measuring written progress, in analyzing differences between writers, or in studying the relationship between writing quality and linguistic phenomena, must have access to more objective and justifiable measures, for which they can turn to the findings of large-scale testing. In Section 3.1, we will discuss these findings, which relate to procedures for obtaining valid and reliable scores of writing quality, attending to factors that impact the performance of both test-takers and raters, and methods of reducing unwanted variance in performance, so that test scores approximate "true" scores of writing ability.

In Section 3.2 we will focus on quantitative analysis of L2 writing, which relies on objective measures of text-based characteristics to evaluate learners' writing in various domains, such as complexity, accuracy, fluency, or cohesion. We will discuss the theoretical links between these features and the constructs of writing ability and/or language proficiency, and also consider the theoretical and practical questions in each domain that influence the precise measures selected and the methods of operationalizing them. Much of the research reviewed in this section draws on theories of 'CAF': that is, theories that second language acquisition is multi-componential in nature, and can primarily be explained through the competing demands of Complexity, Accuracy, and

Fluency (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009). We consider that variables in the domains of CAF, while important and useful for measuring SLA progress, must be situated within the larger context of writing ability and must be supplemented by analysis of other characteristics associated with writing ability, such as reader awareness and discourse organization, as expressed through the constructs of coherence and cohesion (see McNamara, Crossley & McCarthy, 2010).

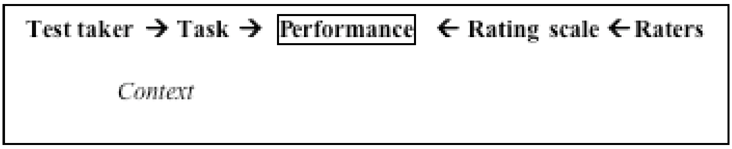
3.1 Assessing Writing Quality

3.1.1 Introduction

There appears to be a consensus in current assessment research that the most valid method of evaluating writing ability is through 'direct' tests of writing (Hamp-Lyons, 2003). Direct tests refer to tests that elicit actual writing samples from participants, which are then evaluated by one or more trained raters; they are referred to as 'direct' to contrast them with 'indirect' tests, or those that evaluate learners' knowledge about writing via multiple-choice, cloze, or other discrete-point question types. While the agreement that actually evaluating writing samples is the best method of evaluating writing ability may seem intuitive or obvious, the fact is that indirect testing was the preferred approach for a substantial portion of the 20th century and was widely supported by psychometric research in which statistical reliability is highly valued (Grabe & Kaplan, 1996). Until the 1970s, indirect testing was the norm and scholars argued that evaluations of writing samples were simply too subjective and unreliable to be appropriate for high-stakes tests, such as university entrance exams. Although indirect tests did produce highly reliable results, researchers gradually began to recognize that validity deserved as much consideration as reliability, and that any valid measure of writing ability needed to engage with theories of writing as a communicative skill (Hamp-Lyons, 1990). Influenced by the L1 writing research reviewed in the previous chapter and by the growing awareness that writing ability entails the ability to set rhetorical goals and tailor a text to a real or imagined reader, assessment researchers focused on improving methods of eliciting and evaluating writing samples that could be scored reliably. Slowly, as reliability improved, the greater validity of direct testing led for the wider acceptance of this practice. The result of this process, and of the concern for both validity and reliability, has been a set of procedures and recommendations that have evolved over the past several decades and allow us to argue that scores on direct tests, if properly designed, are good approximations of "true" writing skill (Weigle, 2002).

Much of the research on direct writing assessment has focused on identifying and classifying the many variables that impact either reliability or validity—two key aspects of assessment discussed in more detail below—and this classification is used to help orient our discussion in the remainder of this section. These factors are usually organized as relating to either the *participants* in the assessment process (test-takers and raters) or to the *texts* involved in the assessment process (the tasks/prompts, rating scales), as represented in Figure 3.1.

Figure 3.1 Factors that influence test performance (based on McNamara, 1996)



The majority of the research reviewed in this section covers issues related to task design and rating scales, as these are the factors that the researcher has the most direct control over, though considerations of the raters and test-takers are discussed to the extent that they affect these two processes. Issues involved in task design include the selection of topics that are contextually appropriate, or the use of prompts that provide adequate levels of specificity without overly influencing the writer, thus allowing all test-takers to perform to the best of their ability. Issues surrounding rating scales include comparisons between holistic and analytic (or multiple-trait) scoring, criterion validity, and rater training. At all stages of writing assessment two key concerns are reliability and validity, and these concepts will be referenced repeatedly throughout this chapter. In the words of Groot (1990): reliability asks: "To what extent do differences in scores between learners reflect differences in ability, rather than other factors?", while validity refers to appropriateness, asking "Do the test scores indeed reflect the ability the test is intended to measure?" (p. 11). Validity includes reliability (a test cannot be valid without being reliable) but also captures other, more inherent characteristics of assessment. Hamp-Lyons (1990) outlines four types of validity that must be considered in writing assessment: face validity, content validity, criterion validity, and construct validity.

Face validity refers to the way a test appears to an "intelligent outsider" (Hamp-Lyons 1990, p. 70), such as professors and administrators involved in the higher education system. Face validity was one reason why indirect tests fell out of favor in these contexts, as there was skepticism that one could properly evaluate writing ability without actually making students write. *Content validity* refers to the extent to which the test includes content that is appropriate and relevant to test-takers, given the purpose and context of the test. *Criterion validity* refers to the relationship between test scores and other measures associated with the same construct. It includes *concurrent validity*, or correlations between writing scores and other measures of proficiency collected at the same point in time, as well as *predictive validity*, or correlations with future performance. For example, the written component of the TOEFL exam, the TWE (Test of Written English) would be considered to have high predictive validity if scores correlated highly with the grades test-takers received in freshman year composition courses. Finally, *construct validity* is, according to Hamp-Lyons, the most important type of validity and refers to the extent to which the test reflects "the psychological reality of behavior in the area being tested" (p.71). That is, it reflects the actual knowledge and abilities of interest on the test, as opposed to other factors. All measures and methods used in writing assessment must be questioned in terms of construct validity as a starting point.

3.1.2 Task design

One of the foremost concerns of writing assessment research has been the design of writing tasks that allow learners to perform to the best of their ability and that distinguish between learners based on writing ability alone (as opposed to differences in personality, topical knowledge, affective variables, etc). While the influence of these unwanted sources of variance may be reduced via the design or selection of a well-validated rating scale and with sufficient rater training, task design is a crucial first step in this process. Task design must engage with the perceived purpose of the test (i.e., what will the test be used for) and with the characteristics of the test-takers, their expected level of proficiency, and their goals with regard to their second language (Weigle, 2002). While researchers generally agree that measuring writing ability with a single task has many limitations (Polio & Williams, 2009), and that it is undoubtedly preferable to obtain multiple writing samples, the practical considerations and high costs associated with qualitative evaluations of texts are such that in many contexts resources severely limit the number of tasks that may be used (Weigle, 2002). When this is the case, the first step is to decide upon the appropriate discourse mode or communicative purpose, with a consideration of the features indicated above.

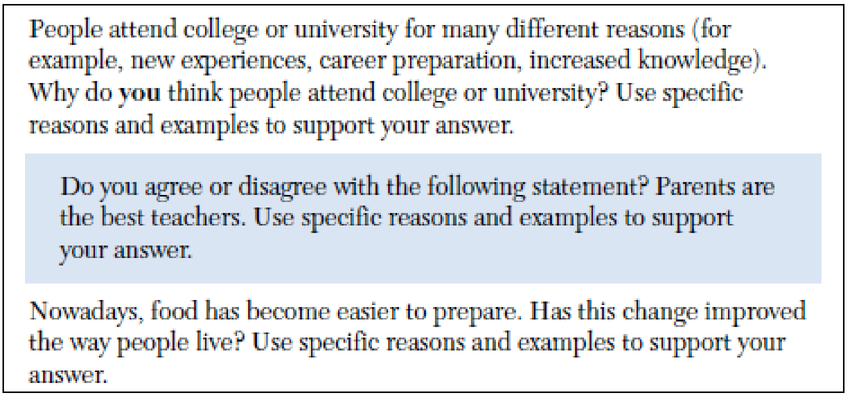
In a test designed for advanced proficiency learners in academic contexts, the most appropriate discourse mode is often the argumentative or persuasive essay. In the categorization of text types identified in Vähäpassi's (1982) model of writing discourse (as reviewed by Sara

Weigle in her 2002 volume *Assessing Writing*), argumentative and persuasive essays are categorized as ‘Type III’ texts, or those that require the writer to invent or generate new information (as opposed to Type I texts, such as dictations or copied material, or Type II texts, such as instructions or summaries, which simply require the writer to reproduce or organize previously available information). Type III texts are considered the most cognitively demanding, and include the types of writing that take on greater importance near the end of compulsory education and throughout higher education (Deane et al., 2008). As Weigle notes, Type III texts like argumentative essays are the types of writing “seen as most critical in academic writing for first-language writers, and for second-language writers in academic settings” (p. 10). For this reason, large-scale assessment designed to screen applicants for their readiness to enter into English-speaking universities, like the TWE, invariably rely on argumentative essays among their essay prompts.

Once the genre and rhetorical mode has been specified, the next concern relates to topic selection, which must engage with the concepts of content and construct validity and has a profound influence on the overall validity of any writing test (Polio & Williams, 2009). As briefly mentioned above, content validity in writing assessment refers to the requirement that test-takers be asked to write about topics that are relevant in the context they hope to enter (e.g., North-American universities, in the case of the TOEFL). Attention to this construct should ensure that the topics selected will not unfairly benefit or disadvantage certain groups of learners based on domain-specific knowledge. That is, a writing test designed to screen applicants to a graduate program in art history might justifiably ask test-takers to write a response critiquing a piece of art; however a writing test designed for general entry into undergraduate or graduate study should avoid topics that require extensive background knowledge of any subject. For this reason, the topics used on general proficiency tests like the TOEFL or IELTS (*International English Language Testing System*) tend to relate to general questions about society or culture, about which opinions may be formed without any specialized education or training. Figure 3.2 shows three typical prompts on the TOEFL exam, all published on ETS’s website for the reference of 2012 test-takers.

In addition to selecting content that avoids domain-specific knowledge and assumes that the majority of test-takers will have sufficient familiarity and experience to develop opinions and produce texts of sufficient length and representative quality, an additional goal in topic selection is to choose topics that test-takers may personally relate to, so that they will be motivated to write. As Hayes (1996) model of writing makes clear, motivation has an important influence on the writing process. If a test-taker feels that the prompt is irrelevant, insufficiently clear, or is too ambiguous or obscure, they may not feel motivated to approach the task with the energy and commitment required to perform well. The first prompt in Figure 3.2 shows the advantages of designing a test for a population of test-takers that is relatively homogenous in terms of their goals and educational backgrounds. That is, because the TOEFL is designed for applicants to North-American colleges and universities, those who wrote this prompt may confidently assume that all test-takers will be able to provide an opinion on the desire for a university education. In contrast, if the test were designed to evaluate the language abilities of recent immigrants hoping to access work opportunities in a new country, as is the case of the General Training version of the IELTS exam, for example (Green, 2004), this topic would be inappropriate from the perspective of content validity.

Figure 3.2 Example prompts from the TOEFL exam[16]



People attend college or university for many different reasons (for example, new experiences, career preparation, increased knowledge). Why do **you** think people attend college or university? Use specific reasons and examples to support your answer.

Do you agree or disagree with the following statement? Parents are the best teachers. Use specific reasons and examples to support your answer.

Nowadays, food has become easier to prepare. Has this change improved the way people live? Use specific reasons and examples to support your answer.

Once the topic has been selected, the specific wording of the prompt must also be taken into account (Weigle, 2002). When writing a prompt, the test designer must consider the amount of specification, with regard to audience style, etc., and the amount of information to provide learners. On the one hand, prompts that are overly general and do not specify the genre, mode, or audience, might lead some test-takers to misinterpret the task demands (for example, they may opt to write a personal essay or anecdote when a more objective analysis is expected), and extreme variation might make it more difficult for raters to apply the same rating scale across texts. On the other hand, prompts that are overly specific or overload test-takers with instructions and expectations may 'pigeonhole' their responses and discourage the type of creativity and 'learner-initiated' behavior that enhances motivation and encourages test-takers to perform

[16] Taken from www.ets.org/Media/Tests/TOEFL/pdf/989563wt.pdf

to the best of their ability (Hamp-Lyons, 1990). Some empirical research is available to support these various decisions. For example, O'Loughlin and Wigglesworth (2003) examined how the presentation of information on the IELTS Academic Writing Task 1 influenced written performance and found that students wrote more complex texts when they were given less information.

Some of the same concerns come into play when deciding whether to give test-takers a choice of writing prompts. While a choice of prompt may increase the odds that all test-takers will find a topic that they are sufficiently knowledgeable about (and motivated to respond to), the use of multiple prompts requires considerably greater effort and expense at the validation stage (Weigle, 2002). That is, extensive validation procedures must be conducted to ensure that no one prompt is more likely to receive positive evaluations than another. In research contexts where the same group of learners is followed over multiple data collection points, variation between prompts may have a clear effect on performance and invalidate the conclusions one hopes to draw about language ability. For example, in Sasaki (2011) she mentions one of her participants claimed they had performed better at later data collection times because "the third- and fourth-year compositions were simply easier to write" (p. 92). Arguably, outside of large-scale testing situations where financial resources allow for extensive pre-test validation of different topics, maintaining the same topic constant across testing sessions may be a more reliable method of comparing performance. This is particularly the case in studies of language development from a CAF perspective, since research indicates that variation in topics (particularly across genres) may affect the quality and linguistic characteristics of texts (e.g., Reid, 1990; Tedick, 1990; Yang, Lu & Weigle, 2012). Yang, Lu and Weigle, for example, compared syntactic complexity indices in the essays of 191 ESL graduate students who each wrote essays on two argumentative essay prompts: the first asked participants to reflect on the importance placed on personal appearance, while the second asked participants on ways of ensuring future success. They found that participants made greater use of clausal elaboration when writing about the future topic, and made greater use of subordination when writing about the appearance topic.

Finally, a handful of other factors in task design also relate directly to practical considerations, but require acknowledgement of how they might influence performance. The factors include the amount of time allotted for writing, whether to specify a minimum or maximum text length, or the preferred medium of composing (i.e., handwritten or word processed). The question of time, for example, requires a clear cost-benefit analysis and should, like all other decisions, consider the purpose and context of the test. While most of the writing that we do in "real life" situations is not timed, and involves the use of additional resources, writing in academic context often occurs in classroom or exam situations where learners are forced to express themselves in writing in limited time and without external aids. On large-scale proficiency tests, time is limited for practical reasons, and the typical amount of time given ranges from 30 minutes to 2 hours, depending on the nature of the test and the number of tasks (Weigle, 2002). The goal is to provide learners with sufficient time to generate a coherent response to a given prompt, but to ensure that their response will be of a manageable length for raters to read and evaluate quickly. The justification for restricting learners to the minimum acceptable amount of time (30 minutes) is that this helps to reduce extreme amounts of variability in text length, which may have an unwanted influence on scoring. Time constraints, as opposed to word limits, seem to be the preferred manner of moderating differences in text length, and instructions to learners typically focus on encouraging them to complete the task, as opposed to simply producing a given number of words. In research contexts time limits are practical as well, and may also help in the analysis of different theoretical constructs. For example, if one is interested in examining the interplay between complexity, accuracy, and fluency, based on theories that these three domains compete for limited attentional resources, restricting time may render the differences more salient; if one is interested in descriptions of the L2 process, restricting time may be key given that the temporal dimension has been shown to be an important source of information about the characteristics and specific challenges faced by second language writers (as reviewed in Manchón, 2009), as we discussed in the previous chapter.

Finally, as regards the medium of composing, the decision to have participants handwrite or word-process their texts must again weigh questions of reliability and validity against practical concerns and the available research. On the one hand, it may be more valid in 2012 to have participants compose their texts on the computer, since the vast majority of 'real-world' writing now involves the use of a word-processor. That said, attempting to simulate real-world conditions raises a whole host of other issues: e.g., in the 'real-world' not only do we generally word process our texts, but we have access to spell-checkers, dictionaries and online resources that can be used to supplement gaps in linguistic or topical knowledge. Furthermore, for test-takers hoping to enter higher education contexts, handwriting may be more valid when one considers the continued use of bluebook exams that require in-class writing. Regardless of the choice of medium, test designers must consider research as to how this variable may influence performance. For example, research has shown that readers may react differently to handwritten versus word-processed prompts: specifically, they may be more lenient towards spelling and mechanical errors in handwritten texts, where they are less salient (Powers et al., 1994). On the other hand, other studies have shown that handwriting may create different biases, and that test-takers with neater and more legible handwriting may be evaluated more favorably than their peers irrespective of the actual content of their texts (Jacobs et al., 1981). Again, as with all factors in task design, decisions related to composing medium must consider the larger context and the practical considerations associated with administering and scoring the writing test.

3.1.3 Raters and Rating scales

In the previous section we considered the ways in which task design may influence performance and impact the validity of writing assessment. Once writing samples have been collected, ideally under conditions that allow test-takers to perform to the best of their ability and within the boundaries of a clearly delineated task, the next challenge in writing assessment relates to selecting a method of evaluating learners' scripts (i.e., a rating scale), and then ensuring that the scale is applied consistently (i.e., through rater training and reliability checking). In this section we will evaluate the recommendations and procedures aimed at increasing the reliability of ratings and the overall validity of subjective evaluations of writing quality.

3.1.3.1 Holistic vs. analytic rating scales

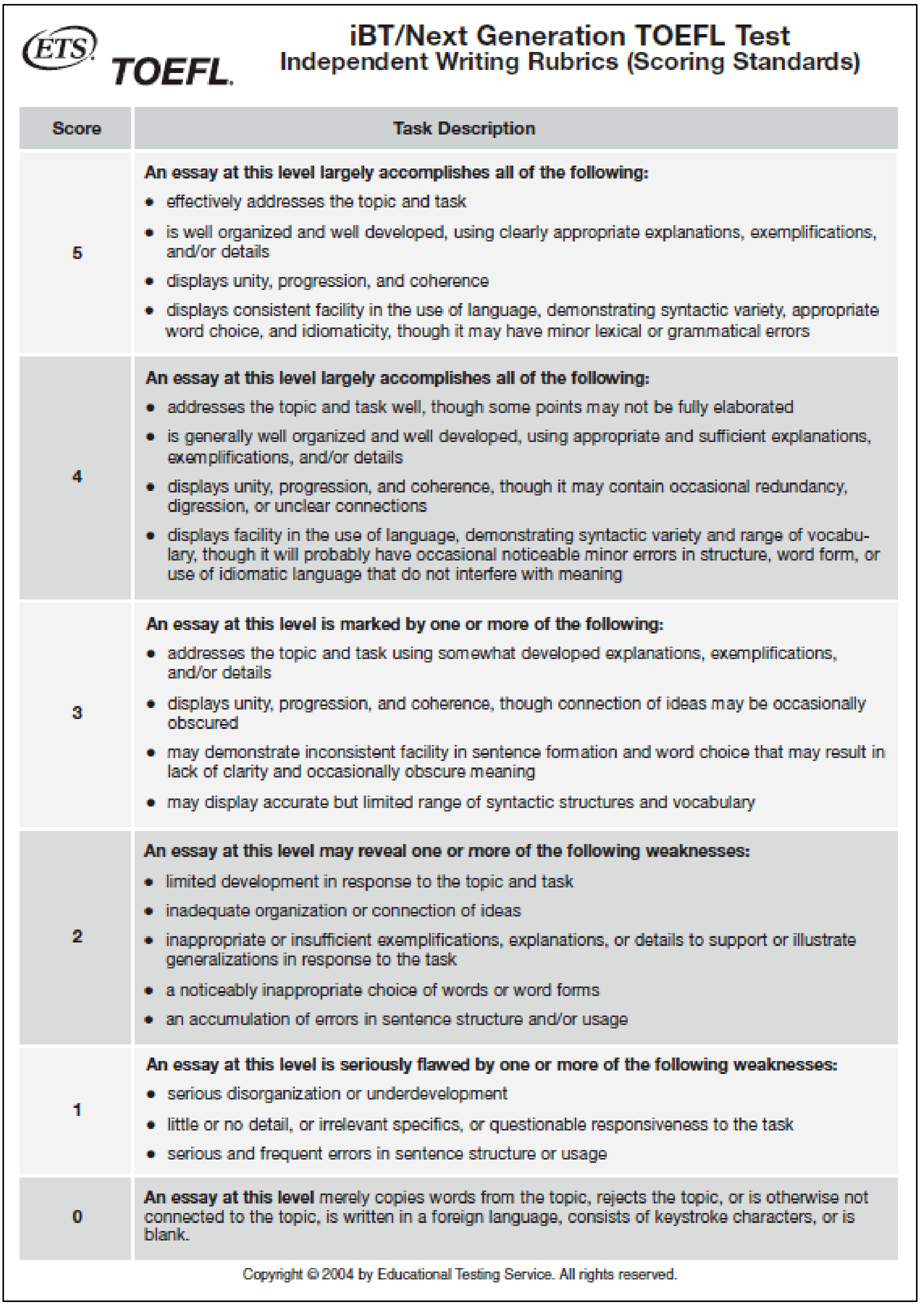
In general sense, all rating scales provide "an operational definition of a linguistic construct such as proficiency" (Davies et al., 1999, p. 153), and are instrumental in criterion-referenced testing. There are two main types of rating scales relied upon in large-scale writing assessment: holistic scales and analytic scales. While there are other types of scales used for different purposes—for example, multiple-trait and diagnostic scales (see Knoch, 2009)—the two main types are the most widely used and relevant to our purposes. Holistic scales ask raters to read a piece of writing once for a "general impression" and then to provide a single integrated score representing the overall quality. Common holistic scales include the 6-band scale used on the TOEFL Internet Based Test (iBT), shown in Figure 3.3, or the Michigan English Language Assessment Battery (MELAB). Analytic scales ask raters to separately evaluate different aspects of a piece of writing, usually through multiple readings, and report scores for different components or features, such as content, organization, or grammatical accuracy. These component scores may then be analyzed separately or added together to arrive at a global measure of quality similar to that obtained through holistic evaluation. Analytic scales are less common in large-scale assessment as they are typically more time-consuming (and therefore more costly) to use; however they are widely used in classroom and research settings as they provide more information about writing quality.

Each type of scale has various advantages and disadvantages based on ease of administration and usefulness for test-administrators or researchers, which Sara Weigle conveniently organized into the contrastive table seen below (Table 3.1). Figures 3.3 and 3.4 present two influential scales which are representative examples of each type.

Table 3.1 Adapted from Weigle (2002, p. 121): "A comparison of holistic and analytic scales on six qualities of test usefulness"

Quality	Holistic Scale	Analytic Scale
Reliability	Lower than analytic but still acceptable	Higher than holistic
Construct Validity	Holistic scale assumes that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects such as length and handwriting	Analytic scales more appropriate for L2 writers as different aspects of writing ability develop at different rates.
Practicality	Relatively fast and easy	Time-consuming; expensive
Impact	Single score may mask an uneven writing profile and may be misleading for placement	More scales provide useful diagnostic information for placement and/or instruction; more useful for rater training
Authenticity	White (1995) argues that reading holistically is a more natural process than reading analytically	Raters may read holistically and adjust analytic scores to match holistic impression
Interactiveness	n/a	n/a

Figure 3.3 Holistic scale used on the TOEFL iBT

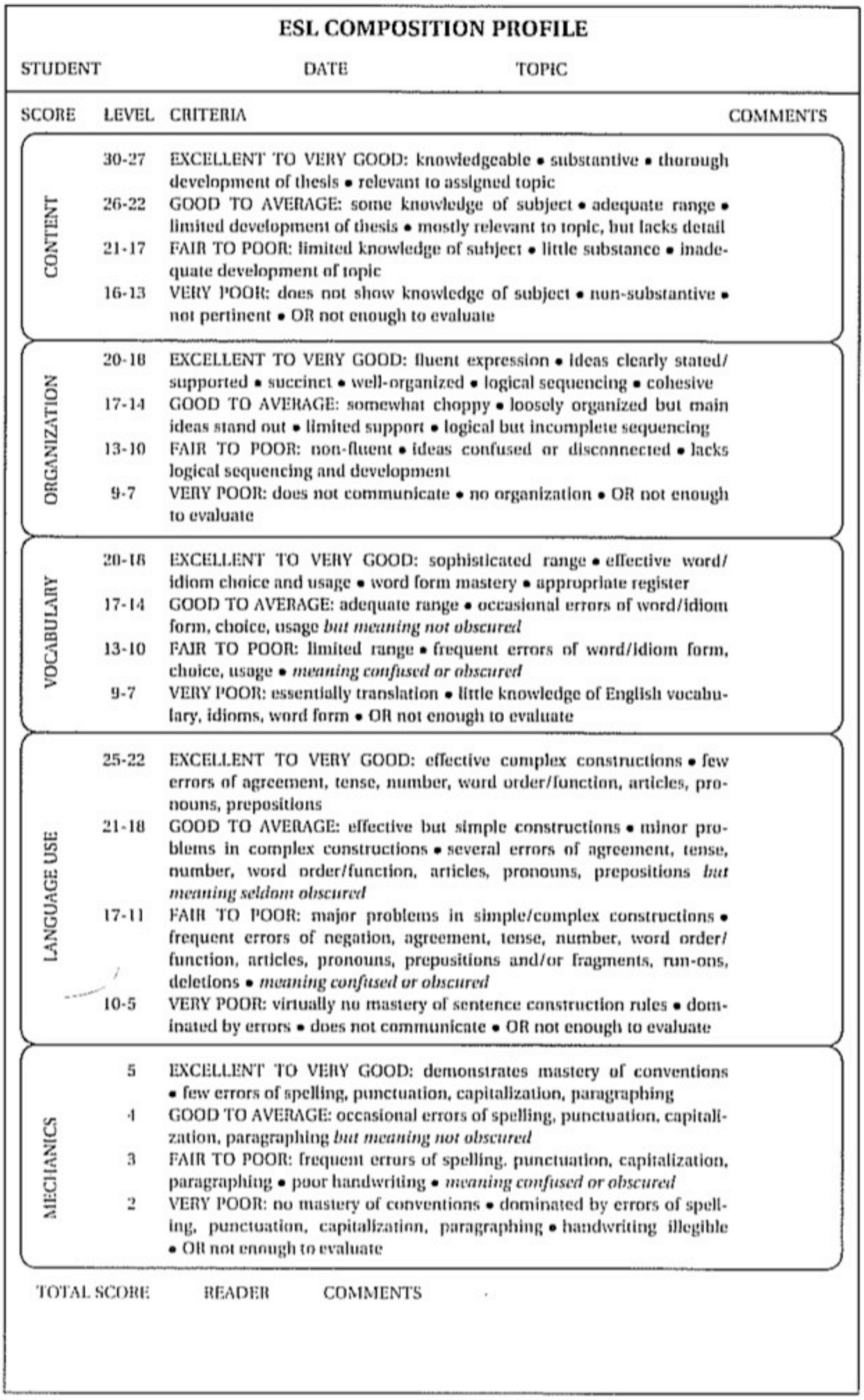


ETS TOEFL

iBT/Next Generation TOEFL Test
Independent Writing Rubrics (Scoring Standards)

Score	Task Description
5	**An essay at this level largely accomplishes all of the following:** • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	**An essay at this level largely accomplishes all of the following:** • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	**An essay at this level is marked by one or more of the following:** • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • may display accurate but limited range of syntactic structures and vocabulary
2	**An essay at this level may reveal one or more of the following weaknesses:** • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage
1	**An essay at this level is seriously flawed by one or more of the following weaknesses:** • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage
0	**An essay at this level** merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Figure 3.4 Analytic scale : The ESL Composition PROFILE (Jacobs et al., 1981, p. 30).



ESL COMPOSITION PROFILE

STUDENT DATE TOPIC

SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive	
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	
VOCABULARY	20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register	
	17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*	
	13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*	
	9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate	
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions	
	21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*	
	17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*	
	10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate	
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing	
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*	
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*	
	2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate	

TOTAL SCORE READER COMMENTS

The primary advantage of holistic scoring is that it is seen as the most efficient approach—given that raters only need to read each script once they may work more quickly—which is why it is more common on large-scale tests. Other advantages are that holistic scoring asks raters to focus on the positive elements of the text (e.g., what the writing does well) as opposed to its deficiencies, and that this is a more natural or authentic manner of reading (White, 1985). Although holistic scales may be more practical for certain situations are more common in large-scale assessment, analytic scales are generally considered to be more appropriate for evaluating L2 writing, since L2 writers often show uneven profiles across different aspects of writing (Kroll, 1998). As Weigle describes it, "a script may be quite well developed but have numerous grammatical errors, or a script may demonstrate an admirable control of syntax but have little or no content" (p.120).

One of the most widely used and well-known analytic scales is that by Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981), called The ESL Composition Profile, which is often referred to simply as The Jacobs Scale (Polio, 2001) but referred to by its creators as the PROFILE, and described as "a criterion-referenced guide to the elements and processes that are believed to be fundamental to an ESL student's success in composing written discourse" (Jacobs et al., 1981, p. v). Despite being more than 3 decades old, this scale remains in use largely because it was extensively piloted and revised over a period of more than 10 years to ensure that it could be applied reliably and that its criteria were justifiable from the perspective of construct validity.

The PROFILE asks raters to evaluate learners' scripts based on 5 different components associated with text quality—Content, Organization, Vocabulary, Language Use, and Mechanics—which the authors argue capture the three essential aspects of the composing process: "what to say, how to organize it, and how to say it effectively" (p. 34). Raters are first asked to rate the success of each aspect of the text as being "Excellent to Very Good", "Good to Average", "Fair to Poor", or "Very Poor", and then they must assign specific point values within these categories, based on a provided set of descriptors or rubrics, as seen in Figure 3.4). For the areas of Content and Organization the four levels of competence are evenly distributed but for the areas focused on usage—Vocabulary, Language Use, and Mechanics—there is a clear division between the second and third categories based on the extent to which errors interfere with the message. Jacobs et al. (1981) indicate that this division is intentional and is such that for these three components, "the PROFILE's minimum levels of mastery…are referenced against an ESL population, while the upper levels…are referenced against native speaker standards for effective written communication" (p. 36). The PROFILE has particular validity from the standpoint of theories that emphasize the communicative purpose of writing. Thus, for example, when evaluating linguistic accuracy, with regards to both vocabulary and language use more generally, the scale encourages raters to consider the gravity of errors, as opposed to the number of errors, and to focus on the success of communication when classifying an essay as “Good to Average” or “Fair to Poor”.

The 5 components on the Jacobs scale are differentially weighted based on the perceived relevance of each aspect to the overall “communicative effect”, which is central to the construct of writing ability as defined by its authors. They justify their weighting system through reference to a study by early TOEFL director David P. Harris, who found that teachers ranked content and organization among the most important aspects of writing, with usage and mechanics given lesser importance, yet their own evaluation practices revealed biases that ran counter to these claims, often heavily penalizing surface errors (Harris, 1969). The PROFILE was designed to correct for such biases and to encourage raters to focus on meaning at each stage while minimizing penalties for surface errors that do not interfere with meaning.

While researchers may opt to design their own rating scales based on the specific purposes of assessment, when aimed to obtain a general method of writing quality, it is often preferable to rely on an existing tool (such as the PROFILE) that has been properly validated and field-tested (Polio, 2001). As Polio points out, however, when selecting a established rating scale the researcher or test administrator must consider the population of test-takers for which the scale was created. That is, using a scale like the TWE or the Jacobs scale, which were both developed with advanced, university-level students in mind, would be inappropriate for a population of younger or beginning-level learners, and overextending them to these populations might negatively impact the reliability and validity of scores (Weigle, 2002).

3.1.3.2 Rater variables and rater training

Regardless of the scale selected, an important source of variability in scoring of writing quality stems from the raters themselves. Just as test-takers bring their different experiences, opinions, and expectations to the test, so may raters, and two important steps in writing assessment are the selection of qualified, competent raters, and the training of raters to ensure that they learn to evaluate, as much as possible, only those elements in the writing that pertain to the construct of interest. Raters may vary in a number of ways, for example: in the leniency and consistency of scoring, in biases against various characteristics of writing (or against perceived traits in test-takers), or in their tendency to grade towards the middle (or at the extremes) of the rating scale (Knoch, 2009). Research in this domain has focused both the ways in which raters vary, and on the ways in which rater training influences raters. Variability between raters was among the primary reasons that indirect tests of writing were promoted for several decades and among the reasons for the increasing interest in automated grading systems, which are perceived as more objective than human graders (Barkaoui, 2007). Researchers have found raters' scores may be influenced by a wide variety of factors such as personality, gender, age, ethnicity, personal and professional experience, and previous rating experience, to name just a few (Hamp-Lyons, 1990; Weigle, 2002).

The rater variables that have received the most systematic attention in research are language background (native vs. non-native speakers), professional background (ESL teachers vs. teachers of other disciplines, such as L1 composition teachers), and rating experience ('expert' vs. novice raters). In Barkoui's 2007 review article, he reports on a handful of studies that explored the role of language background and the degree to which native English speaking (NES) raters were similar to non-NES raters in terms of both leniency and their attention to different characteristics of EFL writing. He reports that Connor-Linton (1995) and Shi (2001) found no significant differences in the scores given by NES and non-NES raters but found that NES raters reported different qualitative reasons for their scores and appeared to have focused on different characteristics than did the non-NES raters. Kobayashi (1992) compared a large sample of NES and Japanese professors (N = 269), who were asked to evaluate the essays of Japanese EFL learners for features of grammaticality, clarity of meaning, naturalness, and organization. Kobayashi found that the NES graders were stricter than the Japanese graders on questions of grammaticality, but were more lenient on questions of clarity of meaning and organization. In a later study, Kobayashi and Rinnert (1996) found that Japanese graders had more positive reactions than NES graders to EFL essays that used Japanese rhetorical patterns, indicating that raters' familiarity with other languages might influence their reactions towards particular groups of test-takers, and ought to be taken into account when selecting and training raters. A similar result was also reported in Hamp-Lyons (1989), who found that NES raters' reactions to EFL essays by learners with different native languages was influenced by the raters' own familiarity with those languages. The different reactions of native and non-native raters, and the inconclusive evidence as to the direction of the effects associated with language background, suggest that the language background of raters (in terms of both native language and familiarity with other languages) should be controlled whenever possible. The effect of both teaching and rating experience has also been explored. Cumming (1990) used think-aloud protocols to compare the decision-making processes of expert and novice raters and found that expert raters paid attention to “higher order features” of content and organization while novice raters focused more on lower order aspects and surface errors. Both Delaruelle (1997) and Weigle (1999) found that experienced raters had a wider range of responses to draw upon than did novice raters, although Weigle (1999) found that the differences between novice and experienced raters in her study varied in relation to the type of task. In general, a number of early studies cited in both Weigle (2002) and Barkoui (2007) collectively suggest that experienced raters are more generous than novice raters in arriving at holistic scores, and appear to be less influenced by surface errors and to reward ‘higher order’ features, in line with the results of Cumming (1990).

Finally, several studies have compared the ways in which ESL teachers and teachers from other academic disciplines react to ESL essays, and the extent to which they value the same characteristics in writing. Barkaoui (2007) synthesizes the results of several different studies that compared the reactions of faculty members from different departments and found that academic discipline significantly predicted raters’ responses to linguistic accuracy; specifically, professors in the humanities and social sciences were found to be more tolerant of errors than were professors in the physical sciences (p. 8). Both Song & Caruso (1996) and Cumming, Kantor, and Powers (2002) compared rating behavior of English faculty and ESL teachers and both found that although the two groups showed evidence of similar decision-making processes (on think-aloud protocols) and arrived at similar holistic scores, the English faculty placed greater emphasis on the content and rhetorical features of essays while the ESL teachers placed greater emphasis on language use. As with language background, given the evidence that ESL teachers may value different features than L1 writing teachers, and given the mixed nature of research findings, professional background (and specifically ESL teaching experience) should be controlled for when selecting raters.

In addition to rater variation, another important focus of assessment research has been on the effects of rater training. Rater training is common practice in large-scale writing assessment, such as on standardized tests, and may help to reduce certain causes of variability. For example, providing raters with benchmark scripts for each point on a holistic scale may help them to use the full range of scores and reduce the tendency of some raters to rely on scores near the middle of the scale. One of the most important goals of rater training is to eliminate biases based on prior experiences, personality differences, or reactions to differences between test-takers. That is, although careful task design may attempt to guide learners towards more abstract and impersonal content and argumentation, moral and affective variables will invariably have an influence on test performance and make their way into test-takers scripts (Hayes, 1996). Test-takers' moral compass, religious beliefs, cultural orientation, and baseline personality traits may be clearly apparent in their writing, and raters may instinctively react to these characteristics despite the fact that they are clearly irrelevant to the construct of writing ability as defined by the rating scale. Hamp-Lyons (1990) describes two of her own earlier studies on this issue (Cooper & Hamp-Lyons, 1988; Hamp-Lyons, 1989) in which she found that readers "make judgments about affective and moral facets of the writer"…that is, that "they 'read the writer' as they read the text, unless carefully trained not to do so." (p. 78). An important aspect of rater training is thus to coach raters to explicitly ignore the personal opinions of value judgments expressed in the text and counteract these biases, which might otherwise introduce unwanted variance in test-scores. In general, studies of rater training have found that when done rigorously it significantly increases the reliability of scores, but does not completely eliminate variability between raters (e.g., Weigle, 1998). For this reason, raw scores from single raters are rarely considered reliable measures of writing ability and it has become the norm in large-scale assessment and any high-stakes testing contexts to rely on scores from multiple raters (at least two), which are then averaged (McNamara, 1996; Polio & Williams, 2009). Additionally, statistical evaluations of reliability, both within and between raters, must be conducted in any writing test to ensure that raters have adequately internalized the results of training and are consistently applying the criteria as defined by the scale.

3.2 Analysis of textual characteristics

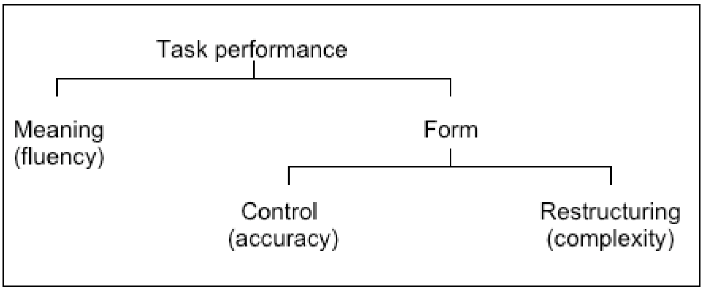
In addition to evaluations of overall quality, SLA researchers often wish to measure written progress in more focused ways, or to evaluate the development of L2 proficiency in writing, for which the most common approach is to rely on objective, quantifiable measures associated with a particular construct (Polio, 2001). Quantitative analysis of writing must engage with the same questions of reliability and validity as qualitative assessment, but when done properly is particularly illuminative because it allows researchers to hone in on highly specific features of language use. The objective measures that have most commonly been used to evaluate L2 writing are those associated with the constructs of complexity, accuracy, fluency, lexical diversity and sophistication, and coherence/cohesion. In each of these domains, empirical research has engaged with the best ways to operationalize these constructs so that analysis may be carried out with high degrees of reliability and validity; for some measures, focused studies have also aimed to determine concurrent validity with extrinsic measures of either writing quality or L2 proficiency (Wolfe-Quintero et al., 1998), while other measures remain "unvalidated" in this regard and awaiting further research. In the following subsections we present the theoretical and methodological issues surrounding objective analysis in each domain.

3.2.1 CAF

The concepts of complexity, accuracy, and fluency have been fundamental to the field of SLA, reflecting a theoretical argument that together they "adequately and comprehensively" capture the principal dimensions of L2 proficiency, as reflected in performance (Housen & Kuiken, 2009, p. 461). The theoretical basis for this argument may be traced to one of two psycholinguistic theories: that learners may choose to prioritize one aspect of their L2 over another based on individual differences; or that learners may choose to prioritize one aspect of their L2 because these aspects compete for limited cognitive resources, particularly in complex tasks (Ellis & Barkhuizen, 2005). Ellis and Barkhuizen trace the historical development of CAF theory back to an early model of language acquisition developed by Meisel, Clahsen, and Pienemann (1981), which posited that learner language progresses along two axes: the first axis references developmental sequences of language, presumably acquired in a logical and hierarchical order; the second axis references "non-developmental" features, or those which are not governed by developmental principles and thus might be acquired at any time. Meisel et al. held that while progress along the 'developmental' axis was prescribed, progress along the 'variable axis' was determined by each individual learner's 'socio-psychological orientation' (cited in Ellis & Barkhuizen, p. 140). More specifically, some learners are presumed to have a 'segregative orientation' and to prioritize communication over considerations of grammar, which others are presumed to have an 'integrative orientation' and to focus more intently upon combining grammar in sophisticated ways, to approximate native-speaker usage. Although these orientations are presumably part of a continuum and not strictly dichotomous, the notion that all learners must choose which aspects to attend to when formulating their messages, was supported both by experimental research and by influential theories of information processing and working memory, which argued that in complex tasks, multiple processes compete for limited cognitive resources (e.g., Baddeley, 1986). One of the first to apply these theories to a model of L2 production was Skehan (1998), who developed a model of task performance in which processing resources are fundamentally divided between attention to *meaning* and attention to *form*. Skehan argued that learners attend to meaning and form to greater or lesser degrees depending on the processing demands of a given task and depending upon individual differences, such as those proposed by Meisel et al. Skehan further suggested that form should be analyzed along two additional dimensions reflecting the processing demands of *control* vs. *restructuring*. In his model, he maps these internal representations onto their linguistic manifestations, such that meaning is related to linguistic *fluency*, control is related to linguistic *accuracy*, and restructuring is related to linguistic *complexity*. Ellis and Barkhuizen provide a helpful illustration of Skehan's mapping in their own review of CAF, reproduced in Figure 3.5.

While Skehan's is not the only account of the interplay between CAF in L2 production, it provides us with an adequate background to understand how these indices may be representative of language development, and an exploration of competing theories, such as Robinson's (2001) Cognition Hypothesis, is beyond the scope of this study. The primary claim of Skehan's model, called the Limited Attentional Capacity Model, is that different aspects of performance compete with one another for limited resources, and that for particularly demanding tasks learners must prioritize one of the three components (complexity, accuracy, and fluency) over the other.

Figure 3.5. Skehan's model of task performance (as visualized by Ellis & Barkhuizen 2005, p. 143)



Thus there is an initial competition between meaning and form and, within form, additional competition between control and restructuring (accuracy and complexity). Given that processing demands stem both from the nature of the task and also from the degree of automaticity or proficiency with the L2, Skehan proposed that learner production could be examined along these three dimensions, and thus that complexity, accuracy, and fluency are effective indices for measuring performance on a given task. Skehan and colleagues primarily investigated these dimensions of proficiency by manipulating task demands, showing that, for example, performance improves if learners are given more time to plan prior to task performance, reducing the demands on working memory (Foster &

Skehan, 1996). According to this line of thinking, then, if a (speaking or writing) task is held constant—and administered over a period of continued L2 learning, whether through focused instruction or immersion—changes in CAF, whether across all domains or only in specific domains, should reflect progress in L2 proficiency (Ellis & Barkhuizen, 2005).

While the theoretical justification for measuring CAF in learner production is robust, the process of actually doing so poses a real challenge, which has been explored in numerous review articles (Housen & Kuiken, 2009; Larsen-Freeman, 2009; Ortega, 2003; Norris & Ortega, 2009; Polio, 2001). As Norris and Ortega (2009) point out: "complexity, accuracy, and fluency are each quite complex subsystems with multiple parts, and trying to get a good look at all the elements that constitute any one of these constructs is a major measurement endeavor" (p. 556). In the following subsections we will analyze the specific challenges associated with the measurement of each construct, and discuss the best practices that have been gleaned from the research. Before delving into measurement of individual constructs, we will consider a general issue that affects measurement in all domains: the selection of units of analysis.

3.2.1.1 Units of analysis

As Ellis and Barkhuizen (2005) point out, the analysis of learner language in terms of CAF "requires a principled way of segmenting a text into units" (p. 147). That is, many of the specific measures discussed in the following sections attempt to quantify characteristics such as the relative frequency of errors, as an index of accuracy, or dependent clauses, as an index of syntactic complexity. The research reviewed in the following sections reveals that the most prominent unit of analysis in CAF research to date is the T-unit (see reviews of text-based research in Ortega, 2003; Polio, 2001); however we argue that this usage constitutes a methodological shortcoming and concur with Ellis and Barkhuizen (2005), Bardovi-Harlig and Bofman (1988, 1989) and others that the only valid unit of analysis for studies of L2 writing, except perhaps when studying young children, is the sentence, as defined by the presence of sentence-final punctuation (a full stop, question mark, etc). Below we will support this view with a discussion of the T-unit and theoretical arguments against its usage.

The T-unit stands for 'minimal terminal unit' and represents "the shortest units into which a pieces of discourse can be cut without leaving any sentence fragments as residue" (Hunt, 1970, p. 189). It specifically consists of an independent clause plus its dependent clauses, and was first introduced by Hunt (1965) to study the development of sentences in the L1 writing of grade school children, who had not yet mastered the conventions of punctuation. Hunt was interested in studying the length of units as a developmental index, and found that the surface characteristics of children's texts complicated this process. Thus the original purpose of the T-unit was to eliminate cases of excessive coordination (common in children's writing), and to allow researchers to study changes in unit-length in texts characterized by a lack of punctuation, and thus a preponderance of run-on sentences (Hunt, 1970). The T-unit was later adopted by SLA researchers interested in oral language for the same reason: that is, the T-unit provided them with a method of segmenting utterances in speech, where sentence boundaries are often difficult to define due to the absence of punctuation (although more recently the conceptually similar AS-unit has become the more popular method of segmentation for unit-analysis, e.g., Mora & Valls-Ferrer, 2012). Although the T-unit has been widely adopted in SLA research, and is frequently used to evaluate written as well as oral production, there are a number of convincing arguments against the use of this unit for written analysis, particularly when studying the production of advanced learners. Bardovi-Harlig and Bofman laid out many of these arguments in their 1988 and 1989 articles (see also Bardovi-Harlig, 1992), making a strong case that the use of T-units for the analysis of adult writing is an inappropriate over-extension of this measure.

Bardovi-Harlig and Bofman (1988) argue that the T-unit is particularly ill suited for analysis of writing produced by advanced second language learners, and that the sentence is the only valid unit of measurement for this group. Their primary argument is that the T-unit does adequately capture learner knowledge, and that it violates the "psychological reality" of the learner. That is, on the one hand "a T-unit analysis divides sentences which were intended to be units by the language learner" and prevents researchers from observing the learner's understanding of sentence-structure in English; on the other hand, T-unit analysis "divides learner-produced text into artificially homogenous units" and "treats all conjoined and non-conjoined sentences equally, as if they were non-conjoined sentences" (Bardovi-Harlig & Bofman, 1988, p. 5). That is, while the T-unit does serve to divide cumbersome run-on sentences, it also breaks up legitimately coordinated sentences which are both grammatical and reflect a certain rhetorical sophistication that moves beyond the presentation of two independent clauses, as in the examples given below in (1), adapted from Bardovi-Harlig and Bofman (1988, p. 4).

(1) Hundreds of schools were built / and tens of institutions
are starting to join in providing technical education to the
public. (2 T-units/ 1 sentence)

> Hundreds of schools were built. / Tens of institutions are starting to join in providing technical education to the public (2 T-units/ 2 sentences)

In additional to the validity concerns associated with T-unit analysis in adult writing, the use of the T-unit creates additional problems for reliability. That is, although T-units are relatively easy to define, inter-rater reliability when calculating T-units still rarely reaches 100% and thus may introduce an additional source of potential variability into analysis of CAF measures (Polio, 1997). In Polio's (1997) study, in which T-units were identified in order to calculate a common index of accuracy (error-free T-units, or EFTs), she reports that although reliability reached .99 in the identification of T-units after careful rater training, there were sources of disagreement, indicating that T-unit analysis was more time-consuming (requiring training), required greater scrutiny (to resolve disagreements) and still did not reach the level of 100% reliability that can be expected when researchers allow learners' punctuation to mark sentence boundaries.

Given the arguments in favor of the sentence as the most valid unit of analysis in adult writing, it is somewhat surprising that so much L2 research has continued to rely on the T-unit. In Ortega's (2003) synthesis of complexity measures in studies of college-level EFL writing, she reviews 25 studies published between 1976 and 1996 and shows that only 4 studies used any sentence-based measures, and that these studies all considered T-unit based measures as well. A look at the studies reviewed in Norris and Ortega (2009) reveals that this preference has hardly changed over the years, and that T-units still remain the most common units of analysis in SLA studies of CAF. Although many of the measures reviewed below thus reference the T-unit, we argue that each of these measures might be improved still further by adapting them to sentence-based analysis.

3.2.1.2 Measuring complexity

The construct of complexity is widely recognized as the most elusive and difficult to define of the three domains in CAF, which is perhaps why it has received the most research attention over the years. As Bulté and Housen (2012) point out, studies that have used complexity as a dependent variable to describe L2 proficiency have often reported mixed results, and this is partly due to the fact that complexity has been defined and operationalized in so many different ways across studies. To begin with, the construct of complexity may refer to any number of sub-types, such as interactional complexity, propositional/ideational complexity, grammatical complexity or lexical complexity (see Ellis & Barkhuizen, 2005, p. 153). In the present section, we focus exclusively on one type of complexity that is particularly relevant for objective analysis of writing: syntactic complexity.

In the large body of SLA research focused on syntactic complexity in writing, this construct has most commonly been operationalized as a function of unit length (e.g., length of clauses, sentences, t-units, etc), the amount of subordination, coordination, the variety of grammatical forms, or simply the presence of forms considered to be sophisticated (e.g., conditions, passives, comparatives) (Norris & Ortega, 2009; Ortega, 2003). The precise measures selected must reflect considerations of the test-taker, the task, and a theoretical understanding of the construct as multi-componential in nature, which Norris and Ortega (2009) argue is not consistently the case in SLA research. In their influential article, they argue that syntactic complexity consists of three measurable sub-constructs: (i) complexity via subordination, (ii) overall or general complexity, and (iii) subclausal complexity via phrasal elaboration (p. 561). They further argue that good practices in SLA research will entail avoiding redundant measures, or those that tap the same construct, but will capture each of these three aspects of complexity, since capturing only one aspect will provide an incomplete picture of development and may lead to a misinterpretation of results. This reflects the theoretical understanding of syntactic complexity as an index of development, developed within the framework of systemic functional linguistics, which holds that there is a developmental sequence in which the expression of ideas proceeds from parataxis, to hypotaxis, to grammatical metaphor (Halliday and Mathiessan, 1999, cited in Norris & Ortega, p. 562). Parataxis refers to the sequencing of independent ideas, expressed as words sentences, and clauses, primarily through the use of *coordination*; hypotaxis refers to the expressing logical relationships between ideas by linking them grammatically, through *subordination*; finally, grammatical metaphor refers to the enrichment and complexification of individual ideas through the use of *nominalization* and *phrasal elaboration*. Individual learners are predicted to proceed along this path as they learn to use their L2 productively and progress towards a mastery of written and formal registers (Norris & Ortega, 2009). The validity of this theoretical orientation is supported by text-based analysis of L1 writing, such as that conducted by Biber (2006), which shows that nominalization and phrasal elaboration are the most common modes of expressing complex ideas and relationships in academic discourse. Given this developmental sequence, Norris and Ortega argue that the progression from coordination to subordination is primarily relevant to describe differences in beginning to intermediate learners, while the progression from subordination to phrasal elaboration is primarily relevant to describe the progression between more advanced levels; thus measuring only one of these dimensions of complexity would be inappropriate in any study in which proficiency is of interest and might lead to misinterpretation. For example, if only subordination measures are used, the researcher "may completely misinterpret whether an increase or decrease is indicative of a positive or negative change in performance, because a decrease in subordination at the highest levels of proficiency may be related to an increase in the overall complexity of the language performance" (Norris & Ortega, 2009, p. 566).

For each sub-type of complexity, there are a number of available measures, some of which have been validated in relation to external measures of proficiency, the selection of which will depend largely upon practical considerations. Global complexity is typically calculated as a function of sentence or T-unit length, and aside from the debate over which of these two units is more appropriate (cf. section 3.2.1.1), these measures are widely accepted and uncontroversial, except for arguments that they are insufficiently descriptive, which may be put to rest by supplementing them with measures of the other sub-types. Complexity via subordination is typically calculated as the number of dependent clauses per sentence, clause, or T-unit, although researchers interested in studying development at lower levels often prefer to complement this with direct analysis of coordination, using measures such as the Coordination Index (Bardovi-Harlig, 1992). Phrasal complexity may be calculated in several ways. The most common measure is mean length of clause, although other measures include length of noun phrases or number of modifiers per noun phrase (e.g., Crossley & McNamara, 2012). Clause length is routinely perceived to be one of the best measures for capturing differences in the writing of advanced learners (Wolfe-Quintero, 1998), although Polio (2001) warns that calculations of clause length may pose problems for reliability and comparisons across studies, since clauses may be defined in different ways. She recommends explicitly defining clause length and training raters to ensure that they are consistently applying this definition across texts.

While most of the common syntactic complexity measures are associated with characteristics of language proficiency and L2 development, and have been validated both by theoretical arguments and empirical studies examining variation between learners of different levels, there is little evidence that they are associated with characteristics of writing quality, and thus their use for analysis of written production must be questioned so that results are interpreted reliably. That is, while most rating scales used to evaluate academic writing make reference to syntactic complexity, it is usually expressed as a dimension of the broader category of syntactic variety, as seen in the rating scales referenced in section 3.1. For example, the TOEFL rubric (reproduced in Figure 3.3) states that a top-rated essay (5) "displays consistent facility in the use of language, demonstrating syntactic variety", while an average essay (3) "may display accurate but limited range of syntactic structures". Thus the assumption is that a competent writer will rely on the full range of syntactic structures in their repertoire, including coordinate and subordinate causes. A highly proficient L2 writer may opt to alternate between simple and complex sentences or to express certain relationships via coordination for stylistic purposes, and this type of variety is clearly associated with greater writing ability. Thus while evidence of more complex structures (like complex noun phrases) is clearly evidence of greater proficiency, it is unlikely that the usage of these structures will increase linearly with proficiency, since their frequency will plateau based on the greater demand for syntactic variety. One way to better understand complexity measures is to compliment them with measures of syntactic variety. This is rarely done for practical considerations, as it is time-consuming and labor-intensive and there are no established measures for manual analysis that have been validated; however new computational tools now offer some promising proxy measures that may help researchers gauge this construct (e.g., Coh-Metrix, 2.0, developed by Graesser, McNamara, Louwerse, & Cai, 2004, which will be discussed in Chapter 5). Another way to ensure reliable interpretation of complexity measures is to compare L2 writing with writing produced by native speakers. That is, native speakers may be presumed to have the highest levels of linguistic proficiency and thus their usage of different syntactic structures (whether in terms of frequency in variety) will theoretically reflect differences in writing ability and establish norms against which L2 usage may be evaluated.

3.2.1.3 Measuring accuracy

In comparison with complexity, the definition of accuracy is relatively transparent; that is, most agree that accuracy describes "how well the target language is produced in relation to the rule system of the target language" (Skehan, 1996, cited in Ellis & Barkhuizen, 2005, p. 139). While Housen and Kuiken (2009) point out that accuracy is "probably the oldest and most consistent construct of the [CAF] triad", and the easiest to justify from the perspective of construct validity, there are still a number of theoretical and practical concerns that govern the analysis of accuracy in L2 writing, most of which relate to reliability. As Polio (2001) points out, while linguistic accuracy generally refers to the absence of errors, "the scope of the term varies from study to study and may or may not include word choice, spelling or punctuation errors" (p. 94). That is, there is a great deal of variation across studies in terms of how an error is defined, which represents an important first step in accuracy analysis.

As Ellis and Barkhuizen (2005) note, one of the most pressing concerns in error definition is whether one will be guided by considerations of grammaticality (defining errors as violations of prescriptive grammar), or considerations of acceptability (defining errors as violations of native-speaker usage). One argument for the latter approach stems from theories of writing as a communicative skill, and the notion that errors which are technically ungrammatical but are acceptable in native-speaker usage will not impede understanding and thus do not reflect negatively on the writer's communicative competence. As we saw in section 3.1, on the analytic scale developed by Jacobs et al. (1981), accuracy is addressed in three of the five components (language use, vocabulary, and mechanics); however at each point the rater is instructed to consider the relative gravity of the errors made, in addition to their frequency, and to focus on the extent to which errors impede communication (that is, whether "*meaning is confused or obscured*"). Focusing on errors that are truly ungrammatical, and thus may impede communication, is thus more coherent with most understandings of writing ability. Another issue to be considered is the treatment of spelling errors. While spelling errors are frequently disregarded in L2 writing research, there are convincing arguments that should be considered alongside other dimensions of accuracy (although perhaps separately) in that they have been found to influence holistic ratings (Bestgen & Granger, 2011) and are an additional reflection of the linguistic knowledge or control of the writer (Mollet, Wray, Fitzpatrick, Wray, & Wright, 2010). Finally, one of the most pressing concerns in accuracy analysis is the issue of reliability (Polio, 1997). That is, because errors may be difficult to interpret or classify, getting researchers to agree on the nature (or even number) of errors may be difficult, but is important to establish so as not to invalidate any findings.

The most common measures of accuracy are the number of errors per unit (word, sentence, clause, or t-unit), or the number of error-free units, and these are relatively uncontroversial. Within studies that have calculated error frequency, there are varying degrees of classification (e.g., grammatical errors, lexical errors, overall errors), and the decision to attend to these will depend upon specific research interests, although there is evidence that reliability is increased by the use of a hierarchical system that allows raters to work systematically and establishes clear rules about the scope of an error (James, 1998; Polio, 1997).

Polio identified some of the pitfalls involved in accuracy analysis in her 1997 study, in which she, and an additional rater, evaluated 38 EFL essays using rating schemes and measures common in the field and sources of disagreement and confusion were documented and quantified. Her article provides a valuable set of guidelines and recommendations for researchers interested in examining accuracy and has undoubtedly helped improve the treatment of accuracy in CAF analysis. She specifically focused on holistic evaluation (not considered here), the identification of EFTs and the identification of individual errors, using a system modified from Kroll (1990) that classified errors based on 33 different error types, such as "subject-verb agreement" or "incorrect tense". For the identification of error-free units, she found that intra-rater reliability was relatively high (above .90) and that inter-rater reliability was acceptably high as well (generally above .80). She identified 5 possible reasons for disagreement: legibility, prescriptive rule, questionable native-like usage, intended meaning not clear, and mistake on the part of the rater. Of these, the greatest source of disagreement was the "nativeness" of a given sentence. That is, sentences where raters felt the sentence would not have been written by a native-speaker, but may not have been ungrammatical. For error classification, she found that reliability for the overall number of errors was quite high (.89-.94), but that reliability on specific error classification was lower, as raters applied different rules to arrive at their classifications. She found that additional guidelines were required to ensure that errors were not double-marked, and that if one error caused an additional error but a single change could fix them both, only the first error should be marked. This confirms the advantages of using a highly-specified hierarchical system of error classification, such as that laid out by James (1998), in his exhaustive volume *Errors in Language Learning and Use*.

3.2.1.4 Measuring fluency

Although fluency in common parlance typically refers to a much broader notion of language proficiency (Housen & Kuiken, 2009; Segalowitz, 2011), the analysis of fluency in SLA research generally refers to far more specific and focused notions of processing speed or its physical manifestations. In oral language, fluency may be measured by the amount of speech produced in a given amount of time, through subjective ratings (usually by native-speaking judges) or through objective indices like speech rate, rhythm, the number and duration of pauses, interjections, or dysfluency phenomena (Lennon, 1990). In written language, fluency may be measured during the writing process as the amount of text produced in continuous bursts (e.g., Chenoweth & Hayes, 2001); however in studies of written products, fluency is most commonly measured as the total amount of text produced in a given amount of time, calculated either as the total number of words or the number of other syntactic units (e.g., clauses, sentences). This measure is a good baseline indicator of both proficiency and writing ability, as a number of studies have shown that text length, or the amount of text produced increases with proficiency and tends to correlate highly with subjective evaluations in empirical research (Weigle, 2002; Wolfe-Quintero, et al., 1998). In contrast to measures of accuracy and complexity, calculating fluency is thus quite simple, and can be conducted easily with a range of widely available software, including most word-processors (e.g., MSWord).

The one theoretical issue related to the operationalization of fluency was initiated by Wolfe-Quintero, Inagaki, and Kim (1998) in their book reviewing the use of CAF measures to analyze writing. These authors argued that all length-based measures, such as mean length of sentence, T-unit, or clause, are actually fluency measures (p. 14); however, as Ortega (2003) points out, they provided little theoretical justification for this somewhat controversial position, other than to argue that they are not sufficiently informative with regards to complexity. Ortega, in contrast, provides ample evidence to refute Wolfe-Quintero et al.'s interpretation, and to show that length-based measures like mean length of sentence or mean length of clause reflect important dimensions of syntactic complexity. Despite the greater theoretical and empirical support for the use of length-based measures as indices of complexity, Wolfe-Quintero's book is widely cited and is one of the standard references in CAF research, such that it has generated a certain amount of confusion surrounding fluency measurement in writing.

3.2.2 Lexical diversity and sophistication

Another fruitful domain for objective analysis of L2 writing is the analysis of lexical characteristics, broadly categorized as pertaining to the constructs of either lexical diversity or lexical sophistication. Although these are occasionally combined and considered with CAF analysis as a feature of complexity (e.g., Bulté & Housen, 2012; Ellis & Barkhuizen, 2005), the methods and theoretical considerations surrounding their operationalization are distinct, such that they may be more coherently examined as a separate domain of development. In particular, in contrast to measures of syntactic complexity, they appear to represent a far more linear course of development. Lexical diversity (which refers to the number of distinct word types in a text) and lexical sophistication (which refers to the relative frequency of those word types) are associated with L2 proficiency, following the assumption that they reveal information about the underlying vocabulary knowledge of the writer. Vocabulary size has been directly linked to proficiency in a number of empirical studies (Engber, 1995; Laufer & Nation, 1995). These constructs are also linked to L2 writing ability, as they reflect on the quality of written texts, given that all of the widely-used rating scales specify that good writing involves the use of varied and sophisticated lexis (Yu, 2009). Measures of lexical diversity are considered *intrinsic* measures, in that they can be measured through analysis of the text alone and do not rely on external sources to evaluate the use of lexis (Meara & Bell 2001). Measures of lexical sophistication, on the other hand, are considered *extrinsic* measures, because they rely on external corpora to obtain frequency counts or other information regarding the characteristics of the lexicon used. The operationalization of each is dependent, as with all assessment methods, on the characteristics of the test-takers and larger context.

Lexical diversity is most commonly analyzed through calculation of the type-token ratio of a text (TTR), which divides the number of different words in the text (types) by the total number of words (tokens). There is extensive research, however, showing that TTR correlates highly with text length, and that the longer a text is, the more likely that words will be repeated (Malvern & Richards, 1997). It has thus become a more common practice to make use of modified TTR calculations, such as Guiraud's Index (types/√tokens), which attempt to compensate for variation stemming from differences in text length. Another popular measure of lexical diversity, which attempts to more systematically counteract the effects of text length, is the D-measure, developed by Malvern and Richards (1997). D uses curve-fitting to represent how the TTR changes over a range of token sizes in a given piece of writing or transcribed speech: to arrive at D, a set of mean segmental TTRs is calculated for different sized samples of a text; these values are then matched to a series of curves, and D is considered the value that produces the best-fitting curve when plugged into the formula: (TTR= D/N * ((1 + 2*N/D)_ -1).˙ D is most commonly used to evaluate very short texts (under 50 words), which are commonly produced by beginning or younger learners and create particular problems for TTR analysis (Malvern, Richards, Chipere, & Duran, 2004). While the mathematical bases of D are quite complex, it can be computed automatically using dedicated software (e.g D-tools, developed by Meara & Miralpeix, 2004).

Lexical sophistication is, as mentioned, calculated using external corpora which can evaluate the relative sophistication of the words used based on their overall frequency in the English language. One of the earliest and most well known indices of lexical sophistication was the Lexical Frequency Profile (LFP) developed by Laufer and Nation (1995), which they argued was a reliable index of the productive vocabulary size of a writer. The LFP is calculated by deconstructing the lexicon of a text in terms of frequency bands, using three predetermined word lists developed by Nation (1995) for use with his *VocabProfile* program. The first list consists of the one thousand most frequently used words in the English language (1K list), the second thousand most frequently used words (2K list), and the university word list (UWL) which consists of 836 words common in academic texts though not in day-to-day usage (see Nation 1990). Each text can be assigned a profile, based on the percentage of words from each of these lists, as well as off-list, less frequent words. Laufer and Nation argue that the LFP gives a snapshot of the way a learners' vocabulary is distributed at that particular stage of development, by analyzing the proportion of high frequency, low frequency, and academic words produced in a piece of writing. The original study was able to show that a lexical sophistication measure obtained using the LFP correlated well with external measures of vocabulary size, discriminated between learners of different proficiency levels, and remained stable across writing samples produced by individual learners,

While Laufer and Nation's LFP remains influential, there are a number of computational tools that allow researchers to calculate absolute frequencies (see Mollet et al., 2010), as well as proxy measures, such as Advanced Guiraud 1000 (Daller, Van Hout, Treffers-Daller, 2003), which are significantly less labor-intensive and have been shown to provide similar information about vocabulary size. Advanced Guiraud 1000 removes the most frequent words and, rather than checking the frequency of the remainder, calculates proficiency based on the type count alone. Mollet et al. (2010) who conducted an exhaustive examination of tools for textual analysis found that this measure was a reliable proxy for full frequency counts (correlating at over .84) and correlated well with essay quality and with external measures of linguistic proficiency.

3.2.3 Cohesion

The final domain of objective analysis considered here relates to the construct of cohesion. This construct falls under the broader category of discourse organization, which is an important characteristic of essay quality and featured widely on holistic and analytic scales used to evaluate academic writing. Cohesion refers to the textual features that facilitate coherence, or a logical mental representation of the text in the mind of the reader (Graesser et al., 2004): that is, cohesion is "an objective property of the explicit language and text" that facilitates coherent interpretations (p. 193). The concept of cohesion has been particularly relevant in studies of discourse comprehension. When there is a lack of cohesion, an idea or relationship must be inferred by the reader, and the success of their inferences will depend upon their prior knowledge and degree of reading expertise (McNamara, 2001; McNamara & Kintsch, 1996). Thus research on cohesion has shown that while this construct is related to essay quality, this is dependent upon the characteristics of both the reader and the complexity of the topic (McNamara, 2001; McNamara, Crossley, & McCarthy, 1996).

In an influential 2001 experiment, McNamara showed that 'low-knowledge' readers are helped by greater amounts of cohesion, while 'high-knowledge' readers benefit from "cohesion gaps", which allow them to draw on their own knowledge and make inferences, which in turn leads to a greater number of connections between ideas and a more well-formed mental representation. In her empirical study, she had 80 undergraduate psychology students read different versions of a text about cell mitosis: the original text, which was classified as low cohesion because many relationships were expressed implicitly, and a manipulated text, which increased cohesion through 7 specific actions: replacing pronouns with noun phrases whenever the referent was potentially ambiguous; adding descriptive elaborations using familiar concepts; adding connectives to specify relationships between sentences or ideas; replacing or inserting words to increase referential overlap; adding topic headers; and adding thematic sentences linking each paragraph to the next, and to the overall topic. After reading the text, all students answered comprehension questions targeting information presented in the text. Participants were also classified as having high or low knowledge based on their performance on a 14-question test that assessed knowledge about the components involved in the text (e.g., the cell; the process of mitosis), but did not query information that was actually presented in the texts. Students who answered fewer than 6 questions correctly were classified as low knowledge, and students who answered 6 or more questions correctly were classified as high knowledge. Half of the students read the same version of the text twice, which the other half read both versions, in opposite orders, and McNamara considered the extent to which increased cohesion was beneficial to text comprehension, and how this interacted with prior knowledge. She found that high knowledge readers were more likely to benefit from the low cohesion text, because they were more likely to "generate knowledge based inferences while reading the text" (p. 56); in contrast low knowledge readers showed greater comprehension when reading the high cohesion texts.

In non-domain specific essays on general topics, such as those generally involved in writing assessment for proficiency purposes, one can assume that sufficient prior knowledge is available to readers, such that higher quality texts might be those classified as low coherence—that is, those that use fewer explicit markers of cohesion—allowing readers to make inferences about the relationships between ideas and arrive at their own, more profound interpretations of the text. Alternately, the overuse of cohesive devices might be perceived as redundant and cumbersome, explicitly marking relationships that are obvious to the reader. While this has not yet been investigated extensively in L2 writing, it bears further research as it provides a potentially important index of reader awareness, which is highly linked to writing expertise (e.g., Bereiter & Scardamalia, 1987).

Cohesive devices are defined as specific features, words, phrases, or sentences that guide the reader and allow them to interpret the ideas expressed, and to identify relationships with other ideas expressed in the text and with larger topics or themes expressed therein. Cohesion is thus typically operationalized through measures that quantify the use of specific cohesive devices, such as connectors or logical operators, or through the analysis of structural or referential cohesion, which gauges the extent to which words, concepts, or forms are repeated across sentences, paragraphs, or texts. One of the easiest classes of cohesive devices are "connectives". Connectives are words or phrases used to link ideas and facilitate interpretations, and they may have varying functions and be categorized based on these functions, such as: clarifying connectives ("in other words", "that is"); additive connectives ("also"; "moreover"); temporal connectives ("after"; "before"; "when"), or causal connectives ("because"; "consequently"). Another commonly analyzed facet of textual cohesion refers to the property co-reference. Co-reference occurs when one noun, pronoun, or noun phrase refers (back or forward) to another constituent in the same text (Graesser et al., 2004). Two sentences are considered to be linked by co-reference if they share a single referent, and the degree of linkage between sentences is considered to be an index of cohesion. As with connectives and other cohesive devices, a high degree of co-reference would be thought to benefit low knowledge readers but to be perceived as redundant or obtrusive to high knowledge readers, who are able to make their own inferences about the relationships between ideas.

3.3 Summary: evaluating progress in L2 writing

In this chapter we have considered different methods for evaluating written progress, both qualitatively and quantitatively. Our review of assessment research has highlighted findings that have influenced the methods of evaluation used in the empirical study, such as the benefits of adopting a well-validated rating scale and the importance of selecting raters from similar backgrounds and with similar levels of experience. We also considered methods for objective analysis of writing looking at textual features in domains associated with linguistic proficiency and/or writing proficiency, such as CAF and cohesion. We reviewed the most commonly used measures in the domains of interest, considered the extent to which they have been validated in previous studies, and some of the practical and theoretical questions associated with different measures. While some of these measures are well-validated and clearly linked to the constructs of interest in previous research, other measures have been less well-explored and merit further evaluation to determine how well they capture aspects of linguistic proficiency and writing quality, leading us to a secondary aim of the empirical study, which is to gain a better understanding of quantitative measures of assessment. Armed with the knowledge gleaned from this chapter we now turn to Part II and to our presentation of the empirical study in the following chapter.

PART II

Chapter 4

Objectives and Research Questions

4.1 Introduction to the Study

In Chapter 1 of this dissertation we reviewed the body of research on language development in SA contexts, and highlighted the importance of this area of research, particularly in the European context where the ERASMUS exchange program is the beneficiary of considerable public and institutional funds. We found that there are many open-ended questions about how SA exchanges benefit learners' linguistic development, and also found that certain skills and modalities are relatively understudied in comparison to others.

The majority of SA research has focused on oral production data (i.e., speaking) and has shown that learners reliably improve in fluency, producing longer runs, fewer hesitations and dysfluencies, and more native-like rhythm, among other things (Juan-Garau & Pérez-Vidal, 2007; Segalowitz & Freed, 2004; Valls-Ferrer, 2011); however there is less evidence that learners' speech improves in terms of accuracy or complexity, two other important components of overall language proficiency, as we saw in Chapter 3. Although some studies of oral production reported significant improvement in either general or more specific aspects of grammatical competence (e.g., Isabelli & Nishida, 2005; Mora & Valls-Ferrer, 2012) others reported that no progress was made, or that AH contexts were actually more beneficial than SA contexts (Collentine, 2004; Juan-Garau & Perez-Vidal, 2007).

The lack of consistent findings in the domains of accuracy and complexity may be because the relatively informal, uninstructed SA learning context is primarily, or even exclusively, beneficial to fluency; on the other hand, it may be that more robust evidence would have been found if writing had been studied as extensively as speech. As discussed in Chapter 2, writing and speech vary in many ways, and complexity and accuracy may be prioritized and operationalized differently in each production mode. Accuracy, for example, is highly valued in writing but less important in speech (Sperling, 1996); learners in immersion contexts, who must use their L2 to convey meaning to native speaking interlocutors and to solve real-life problems, may find that excessive monitoring for grammatical accuracy is detrimental to fluency and their ability to communicate. In Chapter 1 we found that, in comparison to oral production, relatively few studies of SA contexts have looked at language development in writing, or considered how writing skills change and improve after time abroad. Furthermore, because of the particularities of these studies, we still do not know whether and how writing might be expected to improve in a typical SA context, where learners are immersed in their L2 but not expressly focused on writing development in that language.

Sasaki's four studies (2004, 2007, 2009, 2011) were methodologically rigorous and exhaustive in many respects; however her studies all address participants who received extensive process-writing instruction while studying abroad, which is most often not the case for ERASMUS learners, or North-American exchange students in Europe. Thus although Sasaki's studies collectively indicate that SA has a significant positive effect on writing, is relatively more beneficial than EFL classes at home, and may increase learners' motivation to write in their L2, it is unclear whether these same benefits and effects would have been found if her students did not have so much writing practice and focused instruction while abroad. Freed, So, and Lazar's (2001) study suggested that SA periods were not beneficial to L2 writing; however the methods used and the heterogeneity of participants make it difficult to evaluate their claims and extend them to other populations. Perez-Vidal and Juan-Garau's (2009) study considered ERASMUS students who spent 3-months abroad and reported gains in fluency alone, finding no changes in complexity and accuracy and presenting an overall less positive picture of writing development than that in Sasaki's studies; however Perez-Vidal and Juan-Garau examined only a small set of measures, and selected these from previous research that had been conducted using oral production data. The research reviewed in Chapters 2 and 3 makes it clear that studies of language proficiency in writing must take into account the differences between writing and speech (for example, in terms of CAF), the nature of writing ability, and the fact that composing competence or communicative competence may develop alongside, but independently, from linguistic competence. That is, when studying language development in writing one must not focus exclusively on CAF but must also consider changes in the structure and organization of content, in cohesion, syntactic variety, and other aspects that indicate progress and may compete for a writer's attention and cognitive resources.

The present study uses data from the same population studied in Perez-Vidal and Juan-Garau (2009), but aims to reconsider and expand upon the findings of this earlier study by using a substantially larger array of quantitative measures and by complimenting quantitative analysis with a consideration of qualitative improvement, in the eyes of trained raters. The selection of measures and methods of assessment take into account the literature reviewed in Chapters 2 and 3. We aim to consider quantitative characteristics that are commonly associated with either linguistic competence and/or writing competence, and to use our own data to reevaluate these presumed relationships. The following sections give more background on the institutional and cultural context in which our research was conducted and then outline the specific objectives and research questions formulated for this study.

4.1.2 The SALA Project

The present study makes use of a corpus of longitudinal writing data collected through "SALA" ('Stay Abroad and Language Acquisition'), a large-scale state-funded research project based at the University Pompeu Fabra (UPF) in Barcelona, Spain[17]. The SALA Project was developed in 2004 by researchers at the UPF in collaboration with researchers at the University of the Balearic Islands (UIB) in Mallorca, and the University of Barcelona (UB). The primary objective of SALA researchers was to evaluate EFL acquisition during study abroad in comparison to acquisition in (AH) classroom contexts, looking at the full range of linguistic skills. SALA was the beneficiary of a 3-year grant from the Spanish Ministry of Education from 2004 to 2007 and the project's funding has since been renewed 3 times. SALA is ongoing and researchers have been collecting data and expanding the scope of the project since its conception; with each renewal of funding, additional participant groups and research questions have been added. The present study, however, looks exclusively at data collected during the project's first cycle and from the first two cohorts of participants, such that all descriptions of SALA's methods and materials

[17] Project led by Dr. Carmen Pérez-Vidal, Universitat Pompeu Fabra, under the umbrella of the consolidated research group "ALLENCAM, Grup d'Adquisició de Llengües a la Catalunya Multilingüe"

(here and elsewhere) pertain to that initial time period. A more complete description of SALA during this time period may be found in Pérez-Vidal, Trenchs, Juan-Garau, and Mora (2006), who present the project's full scope, design, and initial research goals.

The SALA project was designed to capitalize on the specific conditions of the undergraduate translation degree offered by the school of Translation and Interpretation at the UPF, where all students specializing in English as their primary foreign language followed the same course of study during the first two years of their degree. Beginning in 2005, researchers affiliated with SALA collected longitudinal data from students enrolled in this degree program as they participated in two consecutive learning contexts: a 6-month period of classroom instruction at home (AH), followed by a 3-month period of SA. Both the AH and SA periods were obligatory for all students; data was collected from the entire academic cohort before and after each context using the same tasks, materials, and procedures so that progress in different areas of EFL could be evaluated and compared. Baseline data was also collected from a comparable group of native speakers: undergraduate students studying abroad at the UIB, similar to the UPF students in age and educational background. Among the skills assessed during SALA data collection were speaking ability, lexico-grammatical competence, listening comprehension, aural perception, and formal writing ability. The latter, which was assessed via a timed argumentative essay, is the principal focus of this dissertation.

The following chapter provides more detail about SALA in relation to the experimental design and the materials, tools, and procedures used to collect the data that were considered in this thesis. Further information on SALA can also be found in work published by researchers affiliated with the project, including a handful of studies focused on oral production, which have collectively demonstrated gains in oral fluency after the SA and shown a comparative advantage of SA over AH study on native-like speech rate and fluency, though not on purely phonological measures (Juan-Garau and Pérez-Vidal, 2007; Mora, 2008; Mora & Valls-Ferrer, 2012; Trenchs-Parera, 2009; Valls-Ferrer, 2010, 2011).

4.1.3 Multilingual Catalonia

As referenced in the previous section, SALA is part of the larger research group ALLENCAM, which situates all of its research on language acquisition within the context of "*la Catalunya Multilingüe*" (Multilingual Catalonia). Catalonia is a region of Spain with two official languages (Spanish and Catalan), and where the majority of the autochthonous population may be considered bilingual, albeit to different degrees. Both languages are used in public education from the beginning of primary school, and in larger cities such as Barcelona, Spanish and Catalan are both highly visible in daily life: seen on television, in advertisements, in newspapers, on heard on the streets among the city's inhabitants.

Although the multilingualism of participants is not a focus of the present study, it is important to recognize that the majority of SALA participants were born and raised in a bilingual environment and thus that English, the foreign language of interest in our study, might reasonably be described as an "L3" as opposed to an "L2", in line with the body of research on multilingual acquisition referenced in Chapter 1 (i.e., Cenoz & Jessner, 2000; Rivers & Golonka, 2009). In the present study, we opt to use the label L1 to describe any language acquired during childhood, in the home or local community—regardless of whether this refers to one, two or more languages—and use the label L2 to refer to any language not spoken in the home or local community and acquired later in life. Notwithstanding these labels, we recognize that our participants' acquisition of English may be described as a case of L3 acquisition. More details on language backgrounds of participants are provided in the following chapter, in our description of the specific participants who took part in this study.

4.2 Objectives

The primary aim of this study is to explore the benefits of SA on writing ability, in general and in relation to lexico-grammatical proficiency (LGP). This study focuses on a sample of academic writing (argumentative essays) collected longitudinally from SALA participants before and after both AH and SA learning contexts. Participants' writing was evaluated in terms of perceived quality and in terms of objective indices of fluency, lexical diversity and sophistication, accuracy, complexity, and cohesion (the full set of quantitative measures is henceforth referred to as FLACC). We also considered writing samples collected from native speakers (NS), of comparable educational backgrounds (university students studying abroad in Spain) to determine whether the L2 learners converged with native speakers in any domains after spending time abroad, and to improve our understanding of quantitative measures and our interpretation of changes in any domains. This latter point relates to a secondary goal of the study, which is to explore a wide range of objective measures, in the domains of FLACC, with predicted relationships to writing quality and lexico-grammatical proficiency, and determine whether these relationships may be observed in our data. We accomplished this secondary objective through comparisons with native speakers and also by considering how well FLACC measures correlated with qualitative writing scores and with external indices of LGP, obtained using a grammar and cloze test that was part of the SALA battery. More specifically, we aim:

1. To compare the impact of SA and AH study on writing ability.
2. To explore the effect of initial level on improvement after SA or AH learning contexts.
3. To determine how L1 and L2 writing differs, in terms of perceived quality and FLACC, and how SA and AH learning contexts impact these differences.
4. To identify relationships between quantitative measures (in the domains of FLACC) and perceived writing quality
5. To identify relationships between quantitative measures (in the domains of FLACC) and lexico-grammatical proficiency.

In order to accomplish these global objectives, we analyzed essays written by a robust sample of SALA participants over a period of approximately1.5 years, before and after periods of English study at home and abroad. All essays in the corpus were evaluated by two trained raters using an analytic scale, and were also analyzed in terms of FLACC measures, using a range of computational tools.

4.3 Research Questions

The objectives above led us to formulate 4 principle research questions, which guided the analysis and discussion presented in Chapters 6 and 7. Each research question focuses on a single dependent variable (or combination of dependent variables) and is reformulated into one or more subquestions that evaluate different independent variables and/or different pieces of statistical analysis. These four questions are outlined below.

Research Question 1 (RQ1).

Does learners' writing improve over time, and after AH and SA, in terms of perceived quality and in terms of FLACC? Is one context relatively more beneficial than the other?

RQ1a Do qualitative writing scores improve significantly after either the AH or SA learning contexts?

RQ1b Are there significant changes in FLACC measures after either the AH or SA learning contexts?

Research Question 2 (RQ2).

Is writing improvement different for learners with different initial levels of proficiency?

RQ2a Are changes in perceived quality and FLACC different for participants with higher and lower initial writing proficiency (IWP)?

RQ2b Are changes in perceived quality and FLACC different for participants with higher and lower initial lexico-grammatical proficiency (IGP)?

Research Question 3 (RQ3).

How do learners' essays compare to those of native speakers, in terms of the perceived quality of their essays and in terms of FLACC?

RQ3a How do learners' essays compare to those of native speakers in terms of perceived quality?

RQ3b How do learners' essays compare to those of native speakers in terms of FLACC?

Research Question 4 (RQ4).

Do FLACC measures have the predicted relationships with a) writing quality and b) lexico-grammatical proficiency?

RQ4a Which FLACC measures are significantly correlated with qualitative writing scores and which discriminate between high and low scoring learners?

RQ4b Which FLACC measures are significantly correlated with grammar and cloze scores?

RQ4c Which FLACC measures discriminate between learners and native speakers with similar qualitative scores?

Chapter 5

Methods

This chapter presents the methods used to carry out the empirical study, and is divided into multiple sections. The first three sections present the design of the study (5.1), the participants (5.2), and the SA and AH learning contexts (5.3). Section 5.4 addresses the process of data collection and presents the two different tasks that were used to assess participants' writing skills and lexico-grammatical proficiency. Section 5.5 describes the process of transcribing the written corpus and the different procedural decisions that were made at this stage, such as the treatment of spelling and punctuation errors. Finally, the last two sections present the measures used to evaluate participants' writing. Section 5.6 describes the process of qualitative evaluation. Information is given on the rating scale, rater training, and provides on information on intra- and inter-rater reliability. The last section, 5.7, describes the process of quantitative analysis. This process was carried out using a variety of computational tools, which are presented first, and then the different FLACC characteristics selected for analysis are described along with the methods and tools used to measure them in our corpus.

5.1 Design

As mentioned in the previous chapter, the SALA project was designed to capitalize on the characteristics of the Translation and Interpretation degree at the UPF, in which all English majors were required to participate in two distinct learning contexts (AH and SA), in the same order, during their first two years at the university.[18] In order to study the progress made during this obligatory program of study, a repeated measures design was adopted.

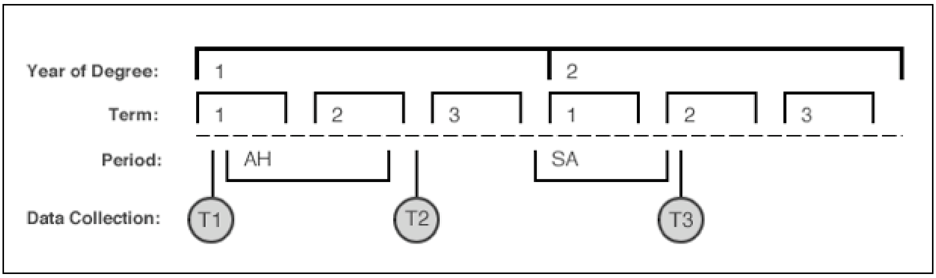
The repeated measures design, in which all participants receive all treatments (in this case, the AH and SA learning contexts may be considered "treatments"), is commonly used in longitudinal studies and in educational contexts, where it is important to reduce variability between subjects (Shuttleworth, 2009). While the absence of a control group and the possibility of cumulative effects or instrument decay make repeated measures designs vulnerable to criticism, especially from scientists accustomed to laboratory conditions and controlled randomization, it has several important advantages in studies of this nature. Firstly, using the same subjects throughout the study reduces unsystematic variability, which increases the power of statistical tests and reduces concerns about variability between subjects due to individual differences that are outside of the study's scope (Field, 2005). Secondly, repeated measures designs are more economical and require fewer participants for statistical analysis, which is a considerable advantage in longitudinal studies that span multiple years, like this one, in which participant fatigue often causes group sizes to shrink from the beginning to the end of the data collection period (Minke, 1997). Finally, the repeated-measures design has validity in the present context, since SALA assesses intact cohorts (academic classes) in a real educational context where treatments are received in a fixed order.

The UPF follows a trimester system, with each trimester including 10-weeks of instruction followed by a 2-week revision and exam period. During the time of SALA, Translation students completed two consecutive terms of English classes in the first and second terms of Year 1, and an obligatory SA in the first term of Year 2. SALA data collection was organized around this schedule: data was first collected from participants at the beginning of their first term at the UPF, before they had received any formal language instruction at the university level. Approximately 6-months later, after the two terms of English instruction, data was collected again. Following a 3rd term, with no formal language study, and a brief summer holiday, participants embarked upon an obligatory 3-month stay abroad (the SA treatment). After participants had completed the SA and returned to the UPF, data was collected a third time. These three data collection times are henceforth referred to as T1,

[18] The degree program has since undergone a number of changes and the description here pertains to the period of time during which data was collected.

T2, and T3, while the two learning contexts are referred to as SA and FI, respectively. The structure of data collection is illustrated in Figure 5.1.

Figure 5.1. Structure of SALA data collection



5.2 Participants

Writing from 30 L2 learners and 28 native speakers was considered in the present study, resulting in a total of 118 writing samples from 58 L1 and L2 participants. (The 30 learners produced 3 samples each, based on the design described above).

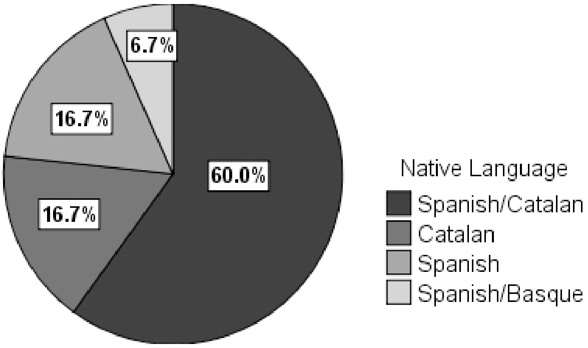
The 28 native speakers comprised the full sample of native speakers that participated in SALA, while the 30 L2 learners were randomly selected from the larger group of SALA participants (N = 81) who wrote essays at all 3 data collection times (T1, T2, and T3). A sample size of 30 was selected as this is frequently considered the minimum sample size for robust statistical analysis and allowed for equal group sizes when comparing learners to native speakers. The L2 learners were part of the first 2 cohorts of UPF students who participated in SALA, from whom data was collected between 2005 and 2008: 21 came from the 1st cohort, students who enrolled in the UPF in 2005, and the remaining 9 came from the 2nd cohort, students who enrolled in 2006. Both the L1 and L2 participants were predominantly female, as seen in Table 5.1, in line with the overall demographics of the translation degree at the UPF.

Table 5.1 Participants

Participants	N	% Female
L2 Learners	30	80% (*n* = 24)
Native speakers	28	79% (*n* = 22)

The native speakers were British and American undergraduate exchange students studying abroad in Spain. 19 of the native speakers were attending the University of the Balearic Islands while 9 were studying abroad at the University Pompeu Fabra. They were all monolingual English speakers, with 12 coming from the UK and 16 from the US, and ranged from 20-22 years in age. The L2 participants were between 17 and 25 years old at the beginning of data collection: the majority (93%) were either 17 or 18 years old, and the mean age for the group was 18.1 years old. Some information about their language backgrounds was reported on a socio-linguistic questionnaire administered through SALA at T1. On this questionnaire, learners were asked to identify their native languages, the results of which are reported in Figure 5.2. As illustrated, the majority of the L2 learners were Spanish/Catalan bilinguals, exposed to both languages from early childhood, while some grew up speaking only Spanish or Catalan at home and then learned the other language later in life, and a small minority (2 participants) came from outside of Catalonia and identified themselves as Spanish/Basque bilinguals.

Figure 5.2 Self-reported native languages of L2 participants



Like all L2 participants in SALA, the learners were working towards a degree in Translation and Interpretation at the UPF and specializing in English as their primary foreign language. Before beginning at the UPF, all learners had considerable exposure to English and were considered by SALA researchers to have “advanced” levels of proficiency (see, for example, Pérez-Vidal & Juan-Garau, 2009). They had received an average of 10 years of classroom instruction in English throughout their primary and secondary education and passed the ‘PAU’ (*Prova d’accès a la Universitat*)—the placement test required for entrance to Catalan universities, which includes sections testing both first and foreign language competence.[19] For acceptance into the English track of the Translation degree, students needed a 6.73/10 (in 2005) or 7.04/10 (in

[19] http://www.gencat.cat/economia/ur/ambits/universitats/acces/vies/pau/info/normativa/index.html

2006)[20]—scores felt to demonstrate academic competence and readiness for higher education. They also passed a general English competence test administered by the university, and while individual scores were not recorded the global description of the Translation degree indicates that before enrolling in the required language courses, all students with English as their primary foreign language should have a level of competence equivalent to at least a B2 on the CEFR[21]. Participation was on a voluntary basis and in line with the ethical code established by the university. Each participant was assigned an alpha-numeric code in the first data collection session, and these codes were used to maintain the anonymity of individual students when analyzing and reporting the data described in the following sections.

5.3 Learning contexts

All L2 participants took part in two sequential treatments, first at home (AH), where they received two semesters of formal English instruction (FI) and then study abroad (SA). These two contexts are described below, focusing on the amount of exposure to English and the amount of writing practiced in each context.

5.3.1 AH context

The FI treatment consisted of the first two English courses offered by the Faculty of Translation and Interpretation, "*Llengua BI*" and "*Llengua BII*", which were required for all English majors in their first and second terms at the UPF. The 'B' in the course title refers to the fact that this was their primary foreign language (their secondary foreign language, usually French or German, was their 'C' language). These courses were oriented around the B2 level, were taught entirely in English and made use of primary, unabridged materials. Although they were divided into two separate courses, and into two trimesters, this was primarily for administrative reasons and they could be considered two halves of a single course; that is, although a different professor taught and evaluated each half of the course, the format remained the same and the contents were progressive. Each FI term included 10-weeks of lectures and seminar sessions, for a total of 40 hours in the classroom, and ended with 2 weeks for revision and a final exam. Thus by the end of the FI treatment, students had received 80 hours of in-class instruction, where English was the medium of instruction and the language used for all written assignments.

[20] See http://www.upf.edu/universitat/upf_xifres/estudis/tra.html

[21] See http://www.upf.edu/factii/factii_grau/presentacio/index.html

Students spent 3 hours each week in large-group lectures (50-60 students) where the syllabus focused on formal linguistic features of the language, such as parts of speech or tense and aspect, and they were taught an analytic approach to constructing and deconstructing English clausal structures and contrasting them with their native languages (see SALA researchers Juan, Prieto, and Salazar (2007) for further explanation of the FI goals). An additional hour weekly was spent in seminar sessions with smaller group sizes (generally less than 20 students), where the focus was on giving students opportunities to practice reading and listening comprehension and to work on formal writing. Writing was the most explicitly practiced skill in the seminar sessions, with graded writing assignments comprising roughly 50% of students' grades. The seminar syllabi included approximately 1-2 class sessions dedicated to essay planning and structure (the use of topic sentences, paragraphing etc). While the exact methods and assignments varied from teacher to teacher, all students completed 1-2 formal writing assignments each term, either argumentative or personal essays, and were usually given the opportunity to revise and resubmit their work. In the lecture portion of the course there was minimal writing practice, and no formal argumentative writing was expected of students, although they did write paragraph-length responses to grammar-related questions during the term. The final exam for these courses tested only the grammar and structural knowledge given in the lectures and did not assess writing or language skills practiced in the seminar sessions.

5.3.2 SA Context

All participants embarked upon their SA in the 1st term of their 2nd year at the UPF, after completing the FI treatment with passing grades on the final exams. On a background questionnaire distributed by SALA researchers, all 30 participants indicated that the SA was their first substantial trip abroad (their first trip lasting more than a month). 13 participants indicated that they had never traveled to an English-speaking country except for a few days as tourists. 17 of them indicated that they had participated in some kind of extended vacation or language exchange in an English-speaking country prior to enrollment at the UPF, the majority having traveled to the UK for a period of 1-2 weeks. Participants completed their SA at a variety of institutions: 27 of the 30 participants attended a university in the UK, while 2 went to Canada and 1 went to Australia. The two students who went to Canada were both at the University of Ottawa, but in different departments (the School of Arts and the Department of Translation, respectively). The 27 students studying in the UK were at a range of universities and were all placed with other UPF students (a minimum of 2 and a maximum of 7). Overall, 76% of participants lived in university residence halls, while 16% lived in shared apartments off-campus and 6% lived with host-families.

The majority of participants (80%) were enrolled in language departments at their host institutions, either Modern Languages or Spanish/Hispanic studies, while the remaining 20% were in Humanities or Translation departments. All participants were expected to register for a minimum of 16 credits, although these credits were transferred as pass/fail and many students did not complete the final exams associated with their SA courses because the exam period occurred in January, after they had returned to the UPF. Most students took Translation and Philology courses focused on their native and secondary foreign languages (Spanish, French or German), but did not take ESL or English grammar courses, and thus would not have had explicit writing instruction in English, although they may have had instruction in other languages. All of the students maintained a personal diary during their time abroad—at the request of the program coordinator and in collaboration with the SALA project[22]—in which they were asked to write weekly entries describing their academic and social experiences. These diaries gave students the chance to practice informal writing at least once a week, and also provided information about the different courses they took and the amount of formal writing required of them during their stay. In the sample of diaries reviewed, some of the students mention completing formal writing in their foreign languages, in exam contexts: *"In this week, I had two exams: French and Applied translation exams. The French exam consisted of two parts: the first part was a formal letter (a job application); the second part was a reading comprehension"*, and many mention making handouts, giving presentations, and completing class assignments that may have involved some amount of writing in English, but most do not mention writing the longer essays or exam papers that are typical in humanities degrees, and which might have been expected of them during the final exam period had they completed it.

5.4 Data collection

At each of the 3 data collection times outlined in Figure 5.1 participants completed a battery of tests assessing their overall English language competence, each focused on a different language skill. Data was collected in exam-like conditions: participants sat in a large lecture hall

[22] The coordinator of the UPF exchanges, Dr. John Beattie, a professor in the English department of the DTCL, is one of the senior researchers affiliated with the SALA project.

and completed a series of timed tests over the course of a 2-hour period. (Baseline data from the native speakers was collected using the same materials and procedures but in smaller groups at the UIB and UPF, respectively). Only 3 of the SALA tests administered during these data collection sessions are considered in this study: the writing test, used to compile the corpus, and the grammar and cloze tests, which provided information on participants' lexico-grammatical competence and were used to investigate one dimension of the independent variable 'initial level'. Descriptions of these 3 tests are provided below; however further details and information about the entire SALA test battery can be found in Pérez-Vidal, Trenchs, Juan-Garau, and Mora (2007).

5.4.1 Writing task

The writing task was presented to all participants with the same instructions and under the same conditions: they were given a ruled, double-sided, exam sheet and were told they had 30 minutes to write a response to the following prompt:

> *"Someone who moves to a foreign country should always adopt the customs and way of life of his/her new country."*

Participants wrote their essays by hand in the allotted time period, without the use of a dictionary or any additional resources. They were told to budget their own time to allow for planning and revision. (A relatively small number of participants sketched outlines or drafts on their exam papers; participants were allowed to request additional sheets of paper if necessary, but none of the final drafts analyzed in this study exceeded a single double-sided sheet.)

The decision to assess writing skills through an argumentative essay was made in light of the context of our study: as discussed in Chapter 3, argumentative writing has high validity for assessment in university contexts, where this type of writing is frequently expected in classroom and exam settings (Weigle, 2002). This discourse mode is also sufficiently complex to motivate more skilled writers to engage in at least some degree of knowledge-transforming (Bereiter & Scardamalia, 1987), and to outperform their peers who adopt the more straightforward knowledge-telling approach. As discussed in Chapter 2, less complex modes, such as narrative writing, may be completed adequately with either approach once writers are familiar with the genre and would have been less suitable for the purpose of differentiating between writing expertise at higher levels of education and proficiency. The specific topic was selected based on the assumption that participants would be motivated and have at least some personal knowledge, due to their experience living in Barcelona, a cosmopolitan city where tourism is among the primary industries and cultural and linguistic diversity are omnipresent.

5.4.2 Grammar and cloze tests

Two additional tasks given during the 2-hour group session were designed to measure participants' lexico-grammatical competence: the first of these was a 20-item cloze test and the second was a 20-item sentence-rephrasing task in which participants were asked to construct grammatically correct alternatives based on example sentences. Participants had 15 minutes to complete each section. While the rephrasing task was considered a grammar test (indeed, it is listed as the "grammar task" in the SALA battery), the cloze test was considered by SALA researchers to be a "global test" requiring the mastery of "vocabulary, grammar, discourse and even reading skills" (Juan et al., 2007, p. 3). This understanding is supported by early SLA research in which cloze tests were considered potentially useful proxies for "lower-order" ESL proficiency (e.g., Alderson, 1979) and found to correlate well with global proficiency measures obtained from writing tests (Fotos, 1991). Later research has refined the dimensions of linguistic competence measured by cloze tests (e.g., Purpura, 1999) but they are still widely used to assess "grammatical knowledge" and knowledge of "vocabulary in context", and included on large-scale exams like the Examination for the Certificate of Proficiency in English (ECPE) (see Saito, 2003).

The SALA cloze task consisted of a gapped text with 20-items entitled "*The Lady Who Liked Adventure*". Participants were given 15 minutes to read the 250-word text and fill in each gap with a single acceptable word. Some of the gaps had more than one acceptable answer, while others were more restrictive. Although each gap had a single correct answer based on the original text, full or partial credit was given for any response that was grammatically correct and semantically appropriate given the context. For example, in gap (1) below, full credit was given for the expected response "*when*", while partial credit was given for the response "*and*".

Examples from cloze test:

> Mary Bruce was in London looking for a nice dress .... (1) she noticed a showroom with a light aircraft for .... (2) at a terribly reasonable price.

The grammar task asked participants to rephrase 20 sentences (or sentence pairs), given a new initial structure, while keeping the meaning as close as possible to the original. Again, participants were given full or partial credit for all responses that were grammatically correct and approximated the meaning of the original sentence. For example, in item 1., full credit was given for both "*Would you mind not using the shower after midnight*" and "*Would you mind not showering after midnight*".

Examples from grammar test:

1. Please don't use the shower after midnight.

Would you mind...……

2. The weather was fine at the seaside last Saturday.

We had...…..

Given that many items on the grammar and cloze tests had multiple potential answers, all tests were graded by a single researcher, to eliminate problems of inter-rater reliability. The researcher, a native-speaker of American English, kept a detailed log of all acceptable answers and minor variations so that judgments could be applied consistently, and randomly selected scripts were regraded to verify intra-rater reliability. Participants' combined scores on these two tests (a percentage correct out of 40) were taken as a global measure of lexico-grammatical competence.

5.5 Transcription

The writing corpus selected for this study—consisting of 118 essays, 90 by L2 learners and 28 by native speakers—was transcribed and digitally formatted by the researcher prior to analysis, working directly with the handwritten original versions. The transcription process required making a number of procedural decisions with implications for analysis, including how to treat spelling and punctuation errors. These issues, though often glossed over in studies of L2 writing, are far from trivial. Indeed, as pointed out by Mollet et al. (2010), "misspellings can be precisely what separates out one writer from another, but they will be unhelpful in many analyses" (p. 434). Decisions as to how to deal with misspelled words were informed by both the constructs analyzed and the requirements of the computational tools used, which culminated in the creation of multiple data sets.

In the first stage of transcription, all essays were typed in MSWord, with auto-correction disabled, and saved as plain text files (.txt). The transcriptions were kept as close as possible to the original handwritten texts: each writer's spelling and punctuation choices were transcribed verbatim; spacing and paragraphing was replicated as faithfully as the word processor allowed. In cases of poor handwriting or ambiguous spelling, colleagues were recruited to give a second opinion, and illegible words were replaced with 'xxx'. There were very few instances of ambiguous or illegible words, however; since the essays were written in exam conditions, participants made an effort to write legibly. This set of "raw" transcriptions was saved as Version 1, which was used for the qualitative evaluations described in section 5.3. Multiple copies were then created and formatted for use with different software. Eventually 4 distinct sets of data were created, with different treatments of errors and formatting based on concerns discussed below (see Appendix 1).

5.5.1 Spelling errors and non-words

As Mollet et al. (2010) and others have noted, spelling errors may be influential in analyses of writing quality and communicative effectiveness, but they tend to obstruct analysis of specific linguistic characteristics and may negatively impact the performance of computational tools. In the present study, the decision was made to create multiple versions and to correct spelling errors in the versions used for quantitative analysis, but to log and classify all spelling errors so that they could be analyzed as a dimension of accuracy.

To begin spelling correction, the spell-check feature of Microsoft Word was enabled and used to highlight words not found in the dictionary. Participants were evenly distributed in their preferences for British or American spellings, and the appropriate dictionary was selected based on this preference. Proper names and foreign words with multiple or unclear standards (e.g., *burka*, *burkha*, *burqa*) were all changed to a single arbitrarily selected standard to maintain consistency; irregular hyphenization and spacing in high frequency compound words was corrected or standardized as well (e.g., *every-day* was corrected to *everyday*; *openminded*, to *open-minded*; *time table* to *timetable*). Following these steps, errors flagged by the spell-checker were scrutinized and classified as either Type 1 spelling errors or lexical errors (non-words) based on their phonetic or orthographic similarity to a context-appropriate target form. If a highlighted word varied only marginally from the target (e.g., *complet*, *allways*; *definetely*, *incompetible*) and the meaning of the word was clear and appropriate in the context, it was recorded as a spelling error and was corrected. In cases where writers used Spanish and Catalan spellings of closely related words, the spell-checker was used as a metric of whether a monolingual English speaker would be likely to recognize and comprehend the word produced. Thus, *exemple* and *advantatge*, were considered spelling errors, corrected to *example* and *advantage*, whereas *humil* (meaning *humble*) and *musulm* (meaning *Muslim*), were considered "non-words".

Non-words were classified as either lexical or morphological errors and were treated differently depending on the focus of analysis. Most non-words could be attributed to L1 transfer and were classified as lexical errors. These included cases of direct borrowing and cases of modifying an L1 root with English morphology (e.g., producing *reflexed*, modified from the Spanish/Catalan verb *reflexionar*, meaning *to reflect*). A few non-words, however, resulted from creative or erroneous morphology (e.g., *fastly*; *overwent*) and these were classified as morphological errors. For analysis of lexical diversity and sophistication where the goal was to accurately estimate learners' lexical knowledge, non-words with clearly identifiable English roots were modified so that learners would get credit for knowing the root (e.g., *fastly* was replaced with *fast*) but all other non-words were eliminated entirely so that the degree of lexical sophistication would not be overestimated. For analysis of other linguistic characteristics, where lexical knowledge was not of primary concern, non-words based on the L1 were corrected and replaced with the most likely English translation (e.g., *relationed*, derived from the L1 *relacionado* (or *relacionat*, in Catalan), was replaced with *related* in the phrase: "*she was afraid of the customs relationed to this country*"). This was done so that the syntactic parsers and lexicons used to extract structural and semantic information would perform optimally. Overall, non-words were very infrequent in the corpus, with only 12 instances of non-words counted in the 90 L2 essays, which consisted of 21,327 words.

After correcting errors identified by the spell-checker, the researcher read each text carefully to find any spelling errors that were camouflaged because the form produced coincided with a different English word. These consisted primarily of what are described by James (1998) as "slips", or "lapses of the tongue or pen" (p. 83) and were logged as 'Type 2' spelling errors. These were corrected when the target word was easily identifiable in the context, high frequency, and used correctly elsewhere in the text or corpus (e.g., writing *sing* instead of *sign* in "*bowing is a sign of respect*"). Errors resulting from phonetic spelling or confusion were also classified as Type 2, and were corrected as long as the target form was context-appropriate and high frequency in the corpus. Thus *tent* was corrected to *tend* in "*we tent to hold on to our own culture*"; *leaved* was corrected to *lived* in "*when I have leaved in Australia or Canada for one or two years*"; and *sump* was corrected to *sum* in the formulaic phrase "*to sump up*". The last example demonstrates the importance of correcting Type 2 errors before carrying out analysis of lexical diversity and sophistication in texts. That is, a learner who erroneously produces "*sump*" a relatively rare word should clearly not be given credit for having a more sophisticated vocabulary than a peer who is able to correctly produce "*sum*", nor should the learner who writes *expend* when meaning *expand*, or *provably* when meaning *probably*. Mistakes with homonyms such as *their*/*there*, *your*/*you're*, or *its*/*it's* were also treated as Type 2 spelling errors. Since both forms of each pair are high frequency they would have had no bearing on lexical frequency, yet leaving them intact would have disrupted computational analysis of syntactic complexity and cohesion, which relied on part-of-speech (POS) taggers. Similarly, minor errors that affected word class (e.g., *live* instead of *life* in a noun phrase, or *the* instead of *them* as an indirect object) were corrected as long as both forms were of equal frequency and there was evidence of correct usage elsewhere in the text. If the writer produced the erroneous form repeatedly (i.e., more than once), it was left intact and coded as a grammatical error in analysis of accuracy. As with non-words, these slips were relatively rare, but it was deemed important to correct them to improve the accuracy and reliability of computational analysis. All spelling errors and non-words were considered in analysis of accuracy and are reported in Appendix 2.

5.5.2. Punctuation errors

Unlike spelling and lexical errors, which were relatively infrequent, punctuation errors were ubiquitous in this corpus. In both the L1 and L2 essays there were numerous instances of run-on-sentences, sentence fragments, improper usage of commas, colons, semi-colons, and other punctuation markers. The decision as to how to deal with punctuation errors was given lengthy consideration, made in tandem with the decision to adopt the sentence as the primary unit of analysis for syntactic complexity, as opposed to the T-unit (Hunt, 1965), which organizes units at the clausal level. As we discussed in Chapter 3, although the T-unit has been used in many studies of CAF and may be useful for oral production and child language, there are convincing arguments that the sentence is preferable for analysis for writing produced by adult learners (Bardovi-Harlig & Bofman, 1988; Bardovi-Harlig, 1992). This is true in part because of the difficulty of "correcting" a writer's punctuation without making assumptions about their intentions and/or ignoring their stylistic choices. In the words of Mollet et al. (2010), correcting punctuation "is a significant violation of the text and implies a shared view about what is "correct"." (p. 435). For these reasons, it was deemed preferable to respect the sentence boundaries established by learners and leave punctuation errors intact whenever possible.

Although sentence boundaries were thus left intact, other punctuation errors and idiosyncrasies were corrected because they obstructed computational analysis of syntactic complexity and cohesion. The software used to analyze these two constructs (L2SCA and Coh-Metrix, described in section 5.7) counted sentences by means of sentence-final punctuation marks (full stops, question marks, or exclamation points). Thus ellipses (…) were problematic, and these were either eliminated (when they were used in the middle of a sentence) or changed to full stops. Similarly, double question marks and double exclamation points were changed to single instances, and question marks or exclamation marks in the middle of sentences (e.g., "*this attitude can help you meet new people and, why not?, to have a very good friendship with them*") were eliminated. All paragraph boundaries established by the learners were respected, but were marked by a hard return and a blank line for computational analysis. Version 4 of the corpus required further modification to format texts according to CHAT transcription conventions for use with the CLAN program (MacWhinney, 2000).

5.6 Qualitative evaluation methods

To meet the principle objectives of this thesis, all essays in the corpus were analyzed both qualitatively and quantitatively. Qualitative evaluations were carried out by two trained raters, both experienced ESL teachers, using the analytic scale developed by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981), described in Chapter 3 (Figure 3.4). In the following subsections we will describe the rating scale, rater selection and training and report intra- and inter-rater reliability data.

5.6.1 Rating Scale

As discussed in Chapter 3, The "ESL Composition PROFILE" was developed by Jacobs et al. (1981) at Texas A&M University, within a comprehensive program for teaching and testing ESL writing skills: the 'ESL Composition Program' (ECP), and remains widely used in SLA research despite having been created more than 30 years ago in part because it was extensively validated, pilot-tested and revised before publication and is "well supported by content and construct validity" (Grabe & Kaplan, 1996, p. 409). It was deemed preferable to adopt a widely-used scale, to enhance the validity and comparability of our findings. The Jacobs scale was appropriate for the present context as it was designed with a similar population in mind (undergraduate EFL students) and was practical because the scale was published within a detailed usage manual with instructions for training raters, maximizing reliability, and interpreting scores.

Jacobs et al. (1981) report baseline data from their own study, carried out with 599 international students at Texas A&M University, in order to link PROFILE scores to concrete EFL skill-levels and establish the PROFILE as a test instrument with generalizability beyond the limits of individual, independent studies. They correlated participants' Total scores on the PROFILE with their performance on two large-scale standardized tests (the TOEFL and the Michigan Battery) and used percentiles to associate scores with EFL skills levels (Table 5.2). The ability to evaluate learners' writing skills independently from their overall ESL skills helped identify differences in writing and overall proficiency within our relatively homogenous corpus.

Table 5.2 "Interpretive Guide Referenced to ESL Writing Skill Levels" (adapted from Jacobs et al., 1981, p.66)

Interpretive Guide Reference to ESL Writing Skill Levels		
PROFILE Score Range	ESL Writing Skill Level	
100-92	High Advanced	Advanced
91-83	Advanced	
82-74	Low Advanced	
64-56	High Intermediate	Intermediate
55-47	Low Intermediate	
46-38	High Beginning	Beginning
37-below	Beginning	

5.6.2 Rater selection and training

All essays in the corpus were evaluated by two raters, working independently. Following guidelines and recommendations outlined in the PROFILE's usage manual, the two raters were selected based on their competence, experience and similarity of background (Jacobs et al., 1981, p. 46). Both raters were experienced EFL teachers who were familiar with standardized proficiency tests and using scoring rubrics to evaluate speaking and writing performance; both raters were female, in their early thirties, were native-speakers of American English, had been living in Barcelona for 3-4 years, were proficient in Spanish and had passive comprehension of Catalan.

The PROFILE usage manual specifies that raters be told to read each essay twice, as quickly as possible, to evaluate the Content and Organization components after the first reading, and to evaluate Language Use, Vocabulary, and Mechanics after the second reading. The stated goal is for raters to reach their decisions quickly and "instinctively", based on their first impressions, as opposed to dissecting the precise features of syntax, lexis, or accuracy, that influence their opinions. The expectation is that the 5 component scores will be highly inter-correlated as they all come from an impression of the same essay, with the same overall communicative effect, but reflect slightly different perspectives. Jacobs et al. (1981) explain that the interrelationship of scores can be used as a measure of internal consistency, or evidence of proper training (this is discussed further in section 5.3.4 along with other aspects of reliability).

The two raters participated in a single training session led by the researcher that lasted approximately 45 minutes. During training, raters were informed of the general goal of the study (to investigate writing development in university EFL students) and the goal of the session: to familiarize them with the PROFILE and practice applying it consistently so that all essays would be evaluated in the same way by both of them, and across multiple readings. Both raters had experience using scoring rubrics and they quickly understood the requirements of the task and the goal to be consistent and reliable. Because the essay topic prompted a handful of participants to reveal information about their cultural backgrounds, the raters were informed that some of the essays had been written by native English speakers; however, they were explicitly asked to ignore any clues in the text about the writer's background and make a conscious effort to grade all essays similarly, using the descriptors in the PROFILE rubrics. After this introduction, the two raters practiced evaluating sample essays from the larger SALA corpus, along with the researcher. Each sample essay was evaluated independently and then scores were compared and discussed. For the first 3 training essays discrepancies of more than 3 points for any component score were examined and agreement was negotiated by rereading the essay in question and discussing the features that influenced scores, in relation to PROFILE criteria. By the 4th and 5th training essays, no major score discrepancies occurred and the training session was concluded.

5.6.3 Evaluation procedures

Each rater was given a binder with multiple copies of the PROFILE and word-processed transcriptions of the 118 essays (Version 1). Although the essays had been word-processed, they were faithful to the original texts in terms of spelling and punctuation errors and formatting, so raters could evaluate all aspects of mechanics except for handwriting. All essays were printed in 12 pt Times New Roman Font and were labeled with a numeric code to maintain anonymity. Two essays were printed on each page and the pages were shuffled in each pack so that the L1 and L2 essays were interspersed and so each rater would encounter the essays in a different order. The raters were given separate sheets to write their scores for each component (see Appendix 3); they were not asked to total their scores, although Rater 2 chose to do so.

Raters were instructed to dedicate no more than 2 to 3 minutes to reading and evaluating each essay, and to evaluate the essays in small batches to avoid fatigue. They were also asked to complete their evaluations in similar conditions (e.g., at the same time of day, in the same location). Both raters completed their evaluations within a week and reported evaluating between 16-24 essays per day. One week after the raters had submitted their initial evaluations, they were given a similar pack with 20 randomly selected essays to be reevaluated, so that intra-rater reliability could be checked. These essays were assigned a new numeric code and raters were asked to replicate the conditions of the first round of evaluations. Both raters completed this follow-up task within two days.

5.6.4 Reliability

After all scores had been obtained from raters and recorded in an Excel spreadsheet, 3 types of reliability were calculated: inter- and intra-rater reliability and internal consistency reliability, in order to evaluate raters' performance and in order to evaluate the PROFILE as a test instrument. All three types of reliability were calculated in SPSS version 15.0.

Rater reliability was calculated using the intra-class correlation (ICC) method (Shrout and Fleiss, 1979), with a two-way fixed-effects model and the confidence interval set at .95. Since the final PROFILE score for each participant was obtained from the average of the two raters' scores, the correlation coefficients reported here represent average measure reliability. Reliability coefficients are supplemented with values for the standard error of measurement (SEM), which account for sampling error and help one estimate the degree to which the reported scores resemble participants' "true" scores (Brown 1999). This measure was calculated in Excel using the formula: SEM = SD * √(1 - reliability). These methods used allow for direct comparison with the normative data reported by Jacobs et al. (1981). These authors aim for reliability coefficients of .85 or higher, especially in high stakes conditions, and set the cutoff point for adequately high reliability at .80, following previous researchers (p. 39).

Intra-rater reliability was assessed first, using the sub-samples of 20 essays that were evaluated twice by each rater. Both raters had high reliability, well above the .85 value recommended by Jacobs et al. (1981), which indicated that they had used the rubric consistently and produced their evaluations competently and conscientiously. SEM values were low, giving confidence that reasonably small differences in scores were indicative of differences in essay quality. Next inter-rater reliability was calculated by comparing the total scores reported for the complete set of 118 evaluated essays. (The average scores reported by each rater were taken for the 20 essays that were graded twice.) Inter-rater reliability was above the established cutoff point and indicated that training was successful, that the rubric had been applied similarly by both raters, and that reported scores were a reflection of performance based on the PROFILE criteria described above.

Table 5.3. Intra- and inter-rater reliability

Intra-rater reliability		
	Rater 1	Rater 2
Reliability	.929	.934
SEM	4.61	3.99

Inter-rater reliability	
Reliability	.85
SEM	6.44

Examination of the scores reported by each rater revealed that Rater 2 used a slightly smaller range and gave slightly lower scores on average but that overall scores were highly comparable. When scores were converted to ranks, eliminating the differences in range, an independent means t-test revealed that the two raters' scores were not statistically different.

Table 5.4. Descriptive statistics: raw scores reported by each rater

Total PROFILE Scores	Rater 1	Rater 2
Mean	83.3	82.9
SD	9.2	7.7
Median	81	83
Min	65	67
Max	100	98

The final measure of reliability considered was internal consistency reliability, measuring the extent to which the component scores were correlated with each other and with Total scores. As Jacobs et al. (1981) explain: "Since the five component scales of the PROFILE are designed to measure the same thing…we would expect to observe a substantial correlation among all the components" (p. 71), and strong correlations between scores are thus an additional method of checking the reliability of evaluations and of the PROFILE as an instrument for this particular corpus. Internal consistency reliability was analyzed by examining the intercorrelation matrix and by calculating Cronbach's alpha, following recommendations by Jacobs et al. The alpha value indicated adequate reliability, α = .82, and there were substantial positive correlations between scores (see Table 5.5), with the lowest correlation between scores at r = .74, and all Components correlating with Total scores at $r \geq .85$. This suggested that all scores were sufficiently interrelated and that internal consistency reliability was acceptably high. (Cronbach's alpha a conservative estimate, reflecting the "lowerbound" of internal consistency reliability, and most researchers feel that .80 is more than adequate to claim "good" reliability (see Garson, 2010).

Table 5.5. Correlations between Total scores and component scores in full corpus (N=118)

Score	Con	Org	Voc	Lang	Mech	Tot
Content	1	.90	.74	.76	.75	.92
Organization	.90	1	.79	.79	.76	.93
Vocabulary	.74	.79	1	.88	.79	.92
Language Use	.76	.79	.88	1	.79	.93
Mechanics	.75	.76	.79	.79	1	.85
Total	.92	.93	.92	.93	.85	1

In sum, analysis of reliability—both between and within raters and as a function of internal consistency—indicated that raters were able to use the PROFILE consistently and effectively to evaluate the quality of essays in this corpus, and that the Total scores and component scores were reflecting similar subjective impressions of communicative effectiveness, though with the expected subtle differences based on the unevenness often observed in L2 writing. After establishing acceptable reliability and consistency, the two raters' scores (both Total and component) were averaged and these mean scores became dependent variables, used to measure qualitative improvement in scores and differences based on language background.

5.7 Quantitative evaluation methods

Following qualitative evaluations, the writing corpus was analyzed quantitatively, measuring characteristics in the domains of fluency, lexical diversity and sophistication, accuracy, syntactic complexity and cohesion

(FLACC). Analysis was conducted using a combination of computational tools and manual analysis. Before discussing the specific measures used, the four different computational tools (AntWord Profiler, the L2 Syntactic Complexity Analyzer, Coh-Metrix, and CLAN) are described in detail, as are the general procedures of their use.

5.7.1 Computational tools

5.7.1.1 Coh-Metrix 2.0

One of the primary tools used for computational analysis of the corpus was the web-based text analysis tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), version 2.0[23]. Coh-Metrix (C-M) takes plain text files as input and computes 60 indices of textual features, many obtained through 3rd party parsers, lexicons, and databases from around the web. We used a small handful of Coh-Metrix indices to consider features of lexical sophistication, syntactic complexity (at the phrasal level) and cohesion. Cohesion, and the extent to which it relates to text difficulty and reading comprehension, is one of the primary interests of the Coh-Metrix's creators (see McNamara 2010) and the reason why this tool was initially developed.

The general procedures for running Coh-Metrix were quite simple: essays were copied and pasted into the web-based tool, one by one, until the full corpus had been analyzed. Output was downloaded and imported into Excel, and then checked against output obtained from other programs (word and sentence counts, which were provided by multiple tools, were used to cross-validate results). Any discrepancies were resolved by scrutinizing essays and by calculating manual counts of select indices. Errors, generally caused by irregular punctuation or paragraphing, were corrected when necessary and essays were re-analyzed until Coh-Metrix output coincided with that of other programs and/or manual analysis. Comparing the sentence counts produced by L2SCA (see below) and Coh-Metrix revealed a minor glitch in the latter program: sentence-final quotation marks threw off sentence frequency counts, which affected many other measures. The decision was made to eliminate quotation marks from all texts analyzed with Coh-Metrix as they were not relevant to any measures selected for analysis. After formatting changes, the sentence-counts across programs were interchangeable. After the accuracy of Coh-Metrix output was confirmed, measures of interest were imported into Excel, and then SPSS for further analysis.

[23] Accessed via http://cohmetrix.memphis.edu/cohmetrixpr/index.html

5.7.1.2 AntWordProfiler 1.200

The freeware concordance program AntWordProfiler 1.200 (AWP)—written in Perl 5.10 by Laurence Anthony and available for download from the authors' website[24]—was used to analyze lexical diversity and sophistication (the latter complemented by two Coh-Metrix indices). AWP was designed based on the theory of the Lexical Frequency Profile (Laufer & Nation, 1995): when given a plain text file as input, the program outputs a "vocabulary profile" for the text using corpus frequency lists specified by the user. The output produced includes type/token counts and percentages for the text as a whole, and then for vocabulary on each level list specified, also producing a glossary of all words appearing in the input, organized by level list, along with the number of times they appear.

In the present study, AWP was first used to scrutinize all essays for non-words, spelling or lexical errors that might have been overlooked during manual analysis. The entire corpus was run through AWP as a single text, using all 14 of the BNC frequency lists created by Paul Nation, downloaded within the BNC version of the Range program (Heatley, Nation, & Coxhead, 2002).[25] All words classified as "off-list" were scrutinized to ensure that they were not due to errors made during transcription. After output was checked and errors were corrected, essays were re-run individually, using only the first 3 BNC lists. The overall number of types and tokens was recorded for each text, as were the numbers of types and tokens on the 1K and 2K BNC lists respectively. These data were all entered into an Excel spreadsheet and were used to calculate measures of lexical diversity and sophistication discussed below.

5.7.1.3 L2SCA 2.3.1

Syntactic complexity was analyzed primarily using a recently developed computational tool called the L2 Syntactic Complexity Analyzer, or L2SCA (Lu, 2010), which is implemented in python and designed for UNIX-systems. L2SCA is open-source and may be freely downloaded

[24] http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html. © Lawrence Anthony 2008

[25] These organize word-families into thousand-word frequency bands based on their occurrence in the 100,000,000 British National Corpus, which is largely comprised of formal, written language. The lists are bundled in BNC version of Range, downloadable from Paul Nation's website, and are described in the documentation for the software. (http://www.vuw.ac.nz/lals/staff/Paul_Nation.)

from the author's website.[26] The system is operated from the command line and relies on two bundled tools, the Stanford parser and Tregex (Levy and Andrew, 2006), to parse and query plain text files given as input: the Stanford parser, which has built-in sentence segmentation, tokenization, and POS tagging functionalities, is used to process text and generate parse trees; while Tregex is used to query the parse trees, using patterns specified by Lu (2010) and written in Tregex syntax.

L2SCA takes plain text files as input and computes 9 frequency counts, and 14 measures of syntactic complexity, which were selected by Lu (2010) based on CAF research evaluated in Wolfe-Quintero et al. (1998) and in subsequent studies of syntactic complexity in L2 written production (e.g., Ortega, 2003). The 9 frequency counts given are: (W), sentences (S), verb phrases (VP), clauses (C), T-Units (T), dependent clauses (DC), complex T-Units (CT), coordinate phrases (CP), and complex nominals (CN). The 14 syntactic complexity indices computed are: mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), clauses per sentence (C/S), verb phrases per T-unit (VP/T), clauses per T-unit (C/T), dependent clauses per clause (DC/C), dependent clauses per T-unit (DC/T), T-units per sentence (T/S), complex T-unit ratio (CT/T), coordinate phrases per T-unit (CP/T), coordinate phrases per clause (CP/C), complex nominals per T-unit (CN/T), and complex nominals per clause (CP/C).

The procedures for using L2SCA were straightforward. The full corpus was first run through L2SCA and the output was scrutinized and compared to manual analysis of several randomly selected texts. Despite findings reported in a validation by Lu (2011), who found that his tool was able to reliably replicate manual analysis once raters had been properly trained, it was found that certain learner errors threw off clause counts (discussed in more detail below in the sections on accuracy and complexity). Once these errors had been identified and corrected, the corpus was reanalyzed and results were used to cross-validate output produced by Coh-Metrix and AWP. Following this, selected measures were imported into Excel and used to calculate the set of syntactic complexity variables described in 5.7.4.

5.7.1.4 CLAN

Analysis of accuracy, unlike quantitative measures in other domains, was carried out manually by the researcher: errors in essays were identified through careful reading and classified using a hierarchical system,

[26] Version 2.3.1. http://www.personal.psu.edu/xxl13/download.html

following procedures described below. Errors were tagged and counted, however, using the CLAN (Computerized Language Analysis) software package (MacWhinney 2000). In order to use CLAN, transcribed essays were converted into .cha files and formatted following CHAT transcription conventions. A number of essays were converted using a macro created in MSWord, and formatting was updated manually by using the ‘check’ command. After texts were in CHAT format, they were read and coded, one by one, using the set of accuracy codes discussed below and included in Appendix 4. After the full corpus was coded, the FREQ command was used to count errors in each category and these were recorded in Excel. Selected categories were then imported into SPSS for analysis.

5.7.2 Fluency measures

Fluency in writing was measured as the number of *words* and the number of *sentences* produced in the 30-minute time frame. Because all participants had the same amount of time to write their essays, the raw counts are used (as opposed to calculations of words/minute or sentences/minute, which are useful in studies where the time allotted for writing is variable—e.g., Sasaki, 2004; Freed, So & Lazar, 2003).

The number of words (#W) was taken from the output given by AWP. Word counts were obtained after eliminating non-words and unintelligible words from texts. Contractions and hyphenated words were counted as 2 words each. The number of sentences (#S) was taken from the L2SCA output. In L2SCA, a sentence is defined by the presence of sentence-final punctuation marks: periods, question marks, quotation marks or ellipses.

Table 5.6 Fluency measures

Domain	Measure	Abbrev.	Method
Fluency	Number of Words	#W	AWP
	Number of Sentences	#S	L2SCA

5.7.3 Lexical diversity and sophistication measures

Lexical diversity was measured as a function of Guiraud’s index (GI), calculated using the formula: $GI = types/\sqrt{tokens}$. As mentioned in Chapter 3, GI is preferred to traditional type/token ratios when comparing texts of varying length, since the likelihood of introducing new word types tends to decrease as text length increases. Type and token counts were taken from the AWP output, and GI was calculated in Excel.

The AWP output was also used to calculate our primary measure of lexical sophistication: Advanced Guiraud 1000 (AG1k). As discussed in Chapter 3, AG1k measures the proportion of "advanced" or "sophisticated" words in a text, looking at the overall number of types and tokens and also at the number of types on the first BNC level list (the first 1000 words). AG1k recalculates Guiraud's type/token relationship after eliminating the most frequent words in English, using the formula: *AG1k = (types-1K types)/ √tokens*. It was proposed as a good proxy measure of lexical richness by Daller, Van Hout, and Treffers-Daller (2003) and was validated by Mollet et al. (2010), who found that it strongly correlated with measures obtained by more exhaustive and time-consuming methods, such as obtaining corpus frequency for all words in a given text and then calculating mean or log frequency. Although more exhaustive measures of lexical sophistication were available using Coh-Metrix, AG1k was selected as this was calculated in AWP where non-words and other obstructive lexis could be easily identified and eliminated from analysis. The primary sophistication measure calculated by Coh-Metrix was used to confirm Mollet et al.'s claims and cross-validate results. This measure calculates the mean frequency of words in a text using the CELEX database (Baayen, Piepenbrok & Gulikers, 1995), which in turn relies on the COBUILD corpus (which includes 17.9 million words, of which about 16.9 million are from written language corpora). When AG1k was compared to the mean frequency measure given by Coh-Metrix, a significant negative correlation was indeed found between the two (r = -.77, p < .001), indicating that the AG1k measure showed a reduction in high frequency words in the COBUILD corpus.

In addition to looking at the use of infrequent words, we considered two potentially complimentary measures of lexical sophistication taken from the Coh-Metrix output. These measures focused on the extent to which writers used vocabulary that was highly specific, as opposed to abstract, as measured via noun and verb hyponymy. Hyponymy is a measure of the *specificity* of words and is calculated by Coh-Metrix using the WordNet database. Wordnet's authors describe hyponymy as a measure of the super-subordinate relations between words when they are classified based on their semantic content. Thus words that are very specific will tend to have many hyponym levels (that is, they fall under many categories). Wordnet provides the following example for noun hyponomy: "the category *furniture* includes *bed*, which in turn includes *bunkbed*; conversely, concepts like *bed* and *bunkbed* make up the category *furniture* and all nouns are classified under the maximally abstract category *entity*." Verbs are similarly categorized, with more specific verbs, like *to whisper* falling under the broader category of *to talk* and then *to communicate*.[27] Therefore very specific or concrete words, like *bunkbed* or *whisper*, have high hyponym values in comparison with words like *furniture* or *communicate*. Coh-Metrix reports mean hyponym values for all nouns and verbs used in the input text and then calculates an average hyponymy score for each category. These average values capture the degree of specificity of a given text, which may be related to both vocabulary size and to the use of concrete examples, both potentially relevant to writing ability and L2 proficiency. The influence of hyponomy on a text may be appreciated by considering excerpts from texts in our corpus with very high (Example 1) and very low (Example 2) noun hyponomy values.[28]

Example 1: High HyN	Example 2: Low HyN
...On the other hand, this importation of foreigners to the island is extremely profitable for a number of different job sectors. Those who choose to purchase homes and cars need English or German speaking lawyers, "funcionarios", and car dealers/real estate companies to serve them. In this way I would say that adaptation is not necessary because it increases the demand of cars/transportation services, housing, and other jobs such as translators. (Participant #57, NS)	*...One must be tolerant of other people's actions because people may behave differently. What may be considered rude in Spain might be seen as appropriate in the United States. An individual may feel nostalgia for their place of birth or where they grew up, but the best part of living in another country is the enriching experience of living a life that is probably unlike your own. Nostalgia is natural and one should retain his or her native customs to keep them connected to their roots.* (Participant #54, NS)

Table 5.7 Lexical diversity and sophistication measures

Domain	Measure	Abbrev.	Method
Lexical Diversity	Guiraud's Index	GI	AWP
Lexical Sophistication	Advanced Guiraud 1000	AG1k	AWP
	Noun Hyponymy	HyN	C-M
	Verb Hyponymy	HyV	C-M

[27] Examples from "About WordNet." Word Net. Princeton University. 2010. http://wordnet.princeton.edu

[28] Examples in this chapter are taken indiscriminately from learner and NS texts. In order to further protect their anonymity, all participants were assigned a number from #1-#58, which replaced the original alpha-numberic codes assigned during the SALA project.

5.7.4 Accuracy measures

To evaluate accuracy, errors in each text were identified manually by the researcher and coded in CLAN. Since all error detection was carried out by a single researcher, this effectively eliminated the problem of inter-rater reliability. Double-coding of 20 randomly selected essays was used to calculate intra-rater reliability, as described below. An attempt was made to maximize intra-rater rater reliability by classifying errors according to a detailed hierarchical system, following suggestions and observations made in Polio (1997). We will first describe the procedures used to calculate accuracy, and then give a list of the measures finally selected.

To code errors, the researcher read through each text, sentence by sentence, in CLAN. If no errors were detected, the sentence was coded as an "Error-free Sentence" (EFS). If an error was identified, it was classified as grammatical, lexical, or pragmatic, in preferential order. That is, if an error could be classified as grammatical, that was the preferred option. Once an umbrella category was established, the error was classified as a sub-type, using the list provided in Appendix 4, and tagged on the coding line. Sub-types were used to help the researcher make consistent decisions but were not considered in analysis. Error identification relied on the researcher's intuitions as a native speaker of English more than on considerations of prescriptive grammar rules; thus usage errors common in L1 speech and informal discourse were ignored if they did not interfere with meaning: For example the use of *was* instead of *were* was not tagged in the phrase "*If it was me who moved abroad, I would...*" as this error is common in L1 speech.

Grammatical errors were primarily coded based on the POS of the affected constituent (e.g., noun error, verb error). Additional categories were available for errors that affected multiple constituents, like subject/verb agreement or word order, and for common errors, like subject omission. Grammatical errors were tagged as 'transfer' when L1 phrases or structures were directly translated and led to ungrammatical English phrasing, which captured cases where both word order and individual constituents were ungrammatical. Preposition errors were tagged as grammar errors except when the preposition affected the meaning of a phrasal verb, in which case they were tagged as lexical errors. Errors with function words included determiners, articles, and pronouns, when they were not counted as subject omission. Lexical errors were considerably rarer than grammatical errors, but included 'non-words' and word choice that was clearly inappropriate given the context. The majority of word choice errors were due to L1 transfer (e.g., producing *sensible* to mean *sensitive* or confusing English verb distinctions like *make* vs. *do* or *hold* vs. *keep*). Learners were given the benefit of the doubt with regards to the lexis used: while many word choices were mildly inappropriate or idiosyncratic, they were only coded as errors if there was a clear misunderstanding of the dictionary definition, or if meaning would have been obscured without reader knowledge of Spanish or Catalan, as in the example "*For instance, in september I am going outside to study English*", where *outside* is used to mean *abroad*, following Spanish/Catalan usage. Pragmatic errors included referential errors that created intra- or inter-sentential ambiguity, as in the example: "*Unless we do it, people around us in that new country may not accept us*" where the pronoun *it* was not clearly linked to anything in the preceding discourse. Pragmatic errors were also used to mark problems with formulaic language and discourse connectors, when learners produced forms that were not idiomatic and distracted from meaning (e.g., "*we must live and let it live*" or "*on another hand*").

After all texts were coded in CLAN, the FREQ command was used to tally errors. These counts were entered into an Excel spreadsheet to compute total grammatical, lexical, and pragmatic errors, and an overall error count. Spelling errors, which were logged during the transcription process described above, were added to the spreadsheet and considered independently. Finally, all accuracy measures were converted to ratios controlling for text length, so that they could be more appropriately considered as measures of L2 proficiency and/or text quality.

As mentioned above, after all essays had been coded once for accuracy, 20 essays were randomly selected for recoding after a period of several weeks. Intra-rater reliability was then calculated, using the ICC method with a two-way fixed-effects model and the confidence interval set at .95. The results revealed that intra-rater reliability was quite high, at .953, and that the researcher was consistently following the established guidelines and observing the error definitions laid out in Appendix 4.

Table 5.8 Accuracy measures

Measure	Abbrev.
Error-free sentences per sentence	%EFS
Errors per word	%TotE
Grammar errors per word	%Gre
Lexical errors per word	%Lex
Pragmatic errors per word	%Prag
Type 1 Spelling errors per word	%Sp1
Type 2 Spelling errors per word	%Sp2

5.7.5 Syntactic Complexity measures

Analysis of syntactic complexity was carried out primarily using L2SCA (Lu, 2010), using a selection of the frequency counts and indices of syntactic complexity produced automatically by this tool; however an additional measure of phrasal complexity was adopted from the Coh-Metrix output, as described below. Table 5.9 (at the end of this section) presents the 4 measures that were considered in our analysis, which included one measure of global complexity, one measure of subordination, and 2 measures of clausal complexity, or phrasal elaboration, a construct deemed to be particularly relevant for describing participants with advanced levels of proficiency (Norris & Ortega, 2009).

The global complexity measure selected was mean length of sentence (MLS). This was taken directly from the L2SCA output, although the MLS index from Coh-Metrix was used to cross-validate results. The measure of subordination was the number of dependent clauses per sentence (DC/S), which was taken by dividing the number of dependent clauses by the number of sentences, using Excel. Dependent clauses are defined as "finite adjective, adverbial, and nominal clauses that are immediately dominated by an independent clause" (Lu, 2010, p. 483) following the definition of a clause given below.

The global measure of clausal complexity selected was mean length of clause (MLC), and was adopted directly from the L2SCA output. As Polio (2001) points out, it is important to explicitly state one's definition of "clause", as various researchers have adopted different measures over the year. In L2SCA, a clause is defined as "a structure with a subject and a finite verb…and includes independent clauses, adjective clauses, adverbial clauses, and nominal clauses" (Lu, 2010, p. 481). Non-finite verb phrases are *not* counted as clauses, but punctuated sentence fragments *are* counted as clauses even when there is no overt verb, following methods developed by Bardovi-Harlig & Bofman, (1989).

Manual identification of these measures in sample texts revealed that, for the L1 essays, the frequency counts given by L2SCA correlated nearly perfectly with those obtained by hand. For the L2 essays, on the other hand, there were certain discrepancies between automatic and manual counts due to learners' errors. L2SCA was designed for analysis of L2 texts and in the original study, which was carried out using a corpus of 3,554 essays written by university-level English majors in China (part of the Written English Corpus of Chinese Learners), errors were not found to create significant problems. Lu (2010) concluded that: "error analysis indicates that learner errors found in the corpus do not constitute a major cause for errors in parsing or in identifying the production units and syntactic structures in question" (p. 488). However, he noted that "most of the learner errors that do exist in the corpus (e.g., errors with determiners or agreement) are of the types that do not lead to structural misanalysis by the parser or misrecognition of the production units and syntactic structures in question by the system" (p. 488). In the present study, scrutiny of L2SCA output in comparison to manual counts revealed that, while the majority of errors did not disrupt L2SCA, there was one error type—obligatory subject omission—that led to consistent discrepancies: since Tregex counts clauses by identifying finite verbs linked to subjects, obligatory subject omission leads to errors in clause frequency counts. Subject omission was a relatively common error in the L2 essays, and clause frequency counts were important to many measures considered in the domain of syntactic complexity, thus this presented a problem. Ultimately, the decision was made to correct errors of subject omission before analysis of complexity (correcting this errors was facilitated by accuracy coding (see Appendix 4), during which subject omission was an independent category of error).

Finally, one additional measure of phrasal complexity was adopted from the Coh-Metrix output. That measure was noun phrase modification (SYNNP), which counts the mean number of modifiers (all adjectives, adverbs, or determiners that modify the head noun) per noun phrase (NP). Syntactic features are analyzed in Coh-Metrix using the Charniak parser for part-of-speech tagging and an internal method of querying (again, the corrected texts were used and Coh-Metrix's performance was confirmed by manual analysis of several randomly selected texts). SYNNP captures the length and complexity of noun phrases, which is an important characteristic of proficient academic writing. Consider the difference between the first two sentence of the text with the highest SYNNP count (Example 3) and the text with the lowest SYNNP count (Example 4):

Example 3: High SYNNP	Example 4: Low SYNNP
In this day and age, the clash of cultures is ever present in the news and our daily lives, as western culture struggles to find a way to relate to a seemingly disparate Middle Eastern set of beliefs. (Participant #49, NS)	*Like it is said in topic, when you move to a foreign country you should adopt their customs and way of life, maybe because if you do that you will stay better and people who live there will respect you as well you respect them.* (Participant #30, T1)

Very high SYNNP counts are derived from very complex NPs, such as "a seemingly disparate Middle Eastern set of beliefs", while very low

SYNNP counts are derived from simple NPs, including those with obligatory article omission, like “topic”, in Example 4.

Table 5.9 Syntactic complexity measures

	Variable	Abbrev.	Method
Sentence Level	Mean length of sentence	MLS	L2SCA
	Dependent clauses per sentence	DC/S	L2SCA
Clause Level	Mean length of clause	MLC	L2SCA
	Modifiers per noun phrase	SYNNP	C-M

5.7.5.1 Syntactic variety measures

In addition to our measures of syntactic complexity, associated with L2 proficiency in the CAF theories discuss in Chapter 3, we also considered two Coh-Metrix indices associated with syntactic variety. We were interested in the relationship between syntactic variety and overall measures of quality, based on explicit mentions of syntactic variety on most holistic and analytic scales (see Chapter 3), and we were also interested in exploring potential interactions between syntactic variety and complexity in development. The first measure of syntactic variety was an index of structural similarity (StrutA): structural similarity is calculated in Coh-Metrix using an algorithm that builds intersections between syntactic trees, and indices of similarity are given for adjacent sentences. High values of StrutA are associated with frequent of repetition of syntactic structures between adjacent sentences, while low values are associated with infrequent repetition of syntactic structures. Consider the examples below. Example 5 is the essay that obtained the highest value for StrutA, indicating very little syntactic variety, while Example 6 is the essay that received the lowest value for StrutA:

Example 5: High StrutA	Example 6: Low StrutA
What if you go to South Africa and you end up eating Valencian Paella? Or hear a conversation in Basque while having a walk in Central Park? What? Strange? Perfectly possible. *Immigration is a common phenomenon. People who change their original country for another one and establish themselves there. As we know, no country is the same*	*When moving to a new country it is important to adjust to the customs and way of life of that country. Not only will this make the transition a little bit easier, but it will also give you a greater appreciation for the culture, it is likely to be less expensive to live "like a native", and you are more likely to make new friends.* *It may be tempting to maintain*

as its neighbour, let alone to the one in the opposite corner of the Earth. That is why immigrants tend to hold on to at least some of their customs wherever they move to. *But to what extent should we maintain our habits instead of taking the ones from the country we are moving to? Speaking with our friends and family in the language we were brought up is ok, as long as we learn properly the one spoken where we are living. Keeping our religious tradition is ok too, as long as it doesn't clash with the religious customs there.* *To sum up, maintaining our culture alive is very important and enriches ourselves, but as long as we adapt to living there and avoid living in cultural ghettoes.* (Participant #9, T2)	*the customs and ways of life of your native country, to spend time exclusively within the ex-patriot community, and not adjust to the local language, food, and other differences. However, what is the point of going through the time and effort to moving to a new country if you aren't going to open yourself up for the full experience?* *Learn the local language, try new foods, change your normal eating and sleeping patterns to adjust to local norms and you just might start to call that new country "home".* (Participant #50, NS)

The other measure of syntactic variety evaluated a slightly different aspect of repetition between adjacent sentences: tense and aspect repetition (Temp). This measure reflects the extent to which tense and aspect are repeated between adjacent sentences throughout the entire text. In Coh-Metrix, all verb phrases are assigned values based on these features and the index of temporal cohesion is calculated by determining repetition scores for verb tense, repetition scores for aspect, and then averaging these two scores. High Temp values are associated with frequent repetition and lower variety (Example 7) while low Temp values are associated with infrequent repetition and greater variety (Example 8).

Example 7: High Temp	Example 8: Low Temp
One of the main issues to take into account when moving to a foreign country is that of adaptability. This does mean that one has to forgo the customs that the person has been brought up with, but it does mean that one has to keep an open mind about what they will find and be prepared to give the new country's customs their due respect. (Participant #44, NS)	*Moving to a foreign country can be very hard in many ways. Firstly, the language spoken there is more than likely going to be different to your own, which could cause problems. Furthermore, the way of life in the country you move to is bound to be different in some, if not all, ways to the life you are used to at home.* (Participant #48, NS)

Table 5.10 Syntactic variety measures

Variable	Abbrev.	Method
Structural similarity (adjacent sentences)	StrutA	C-M
Repetition of tense and aspect (adjacent sentences)	Temp	C-M

5.7.6 Cohesion

Finally, the last domain explored in quantitative analysis was that of cohesion. As discussed in Chapter 3, cohesion is associated with writing quality, to the extent that it facilitates discourse comprehension, but is largely relative to the context (the complexity of the topic) and the knowledge-level of readers (McNamara, 2001). We considered two aspects of cohesion: referential overlap, and the use of connectives (e.g., because, although, etc).

Referential overlap concerns the degree to which words and concepts are repeated in adjacent sentences and paragraphs throughout the text. High amounts of referential overlap indicate highly cohesive texts, while low amounts of referential overlap indicate low cohesion texts. Coh-Metrix provides three indices of referential overlap: anaphor overlap (RefP), or the extent to which anaphoric references were linked to referents in adjacent sentences; argument overlap (RefA), or the extent to which adjacent sentences share a common noun, pronoun, or noun-phrase; and content word overlap (RefC), or the extent to which sentences share common content words of any syntactic category. Our hypothesis was that texts with high amounts of referential overlap, of any type, would be perceived as redundant given the nature of our topic (drawing on common knowledge accessible to all) and the expertise of our trained raters. In the below examples, Example 9 is taken from a text with high values for content word overlap, reflecting the repeated use of the same words across adjacent sentences. Example 10 is taken from a text with low values for content word overlap, reflecting the lower amount of referential overlap.

Example 9: High RefC	Example 10: Low RefC
When moving abroad there are a lot of important things to consider. I think that it is important both to take on board the new culture but for some people it is also important to maintain a bit of the culture of your home country. *There are many reasons for emigrating to a new country.*	*One of the main issues to take into account when moving to a foreign country is that of adaptability. This does mean that one has to forgo the customs that the person has been brought up with, but it does mean that one has to keep an open mind about what they will find and be prepared to give the new country's customs their due respect.*

Many people crave better weather, or a better way of living, and for some people the attraction is change, and a new culture and the opportunity to learn about a different way of life, different customs and even a new language. In my opinion it is rude to move to a new country and not make the effort to adapt to a new culture. Some people emigrate, not for a new culture or language, but for a new climate, and they expect to live as they did in the country they are from, but abroad. It is unfair to expect the natives of their new country to adapt to them, as they are the ones who chose to move abroad and therefore they should attempt at least to speak the language and try out the new customs and way of life. (Participant #37, NS)	*It stands to reason that if through personal choice you decide to move to another country then it would be wise to obtain information on customs, language and if necessary, religion. These tend to be the three aspects that could differ from one's native country. Part of the adventure should be to compare these differences and see them as enriching and positive. There would be no point in moving to a foreign country in which you had no desire to learn from the experience. After all, learning a new language can only be beneficial. It would be one of the best ways to integrate oneself with the community, learn about their way of life and ultimately feel comfortable in the new setting.* (Participant #44, NS)

The other aspect of cohesion considered was the use of connectives. As discussed in Chapter 3, connectives are words and phrases that signal relationships between elements in the text, and are considered "cohesive devices", which provide readers with cues for interpreting texts and facilitate reading comprehension. As with other types of cohesion, the assumption is that a heavy use of connectives may benefit novice readers, particularly for complex topics, but be unnecessary and perceived as redundant by expert readers. Coh-Metrix counts the overall frequency of connectives in a number of different subcategories: additive connectives refer to words and phrases used to extend relationships (e.g., also, moreover); causal connectives refer to words and phrases used to indicate causal relationships (e.g., because, consequently); logical connectives refer to words and phrases used to express logical relationships (e.g., if, actually); and temporal connectives refer to words and phrases used to express temporal sequences (e.g., before, after). We considered the use of connectives in each of these categories and also considered the overall incidence of connectives in relation to the total number of words in the text. We primarily focused on the total incidence of connectives, but expanded this category to look at specific subtypes in several stages of analysis. The full list of elements tagged as connectives pertaining to each category is available on the Coh-Metrix website, and is reproduced in full in Appendix 5.

Table 5.11 Measures of cohesion

	Variable	Abbreviation	Method
Referential cohesion	Anaphor overlap	CrefP	C-M
	Argument overlap	CrefA	C-M
	Content word overlap	CrefC	C-M
Connectives	Additive connectives/word	%AdCon	C-M
	Causal connectives/word	%CausCon	C-M
	Logical connectives/word	%LogCon	C-M
	Temporal connectives/word	%TempCon	C-M
	Connectives/word	%Con	C-M

5.7.6.1 Pronoun density

The final quantitative measure considered was pronoun density (DenPr). This measure was taken from the Coh-Metrix output, and was calculated as the total number of personal pronouns divided by the total number of words. Although pronoun usage did not fit neatly into any of the larger categories explored, previous research has shown that non-native speakers overuse personal pronouns in comparison to native speakers (Shaw & Liu, 1998) and research on register variation had indicated that academic writing is characterized by a lower density of personal pronouns, and by greater impersonality and generality (Biber, 1988). That is, texts with high density of personal pronouns give the impression of being more personal and informal while texts with low pronoun density give the impression of being more formal and academic. Consider examples 13 and 14.

Example 13: High DenPr	Example 14: Low DenPr
One thing if you go to another country only for tourism, then you can adopt their customs if you want to feel more comfortable, but if you do your daily routine and only visit places you can do it, but know that timetables are different, so maybe you can't do something that you would like to. On the other hand, if you are going to live there for more than a month maybe the better you can do is to adapt at their customs, but never leaving your culture. You must remember forever who you are and what you want to be. (Participant #30, T1)	*When in Rome do as the Romans. For any person who has ever traveled or spent some time abroad this code of living is a crucial aspect to the foreign experience. When first arriving to a foreign country, one is immediately bombarded with a plethora of new customs, people, and, at times, a new language. For many, this sudden and drastic change bursts their bubble of comfort, causing a feeling of awkwardness and possibly fear.* (Participant #36, NS)

Chapter 6

Results

This chapter presents the results of data analysis and is divided into four main sections corresponding to the four research questions presented in Chapter 4. Section 6.1 addresses longitudinal improvement in participants' writing, in order to answer RQ1.

RQ1. Does learners' writing improve over time, and after AH and SA, in terms of perceived quality and in terms of FLACC? Is one context relatively more beneficial than the other?

In this section we analyze improvement in the L2 participants' essays, first in terms of qualitative scores (in section 6.1.1) and then in terms of quantitative measures in the domains of FLACC (in section 6.1.2). For each set of dependent variables (qualitative scores and FLACC measures), we first provide descriptive statistics for the corpus as a whole, which serve as a reference point throughout the remainder of the chapter, and examine the distribution of variables within groups. We then conduct statistical analysis to determine whether the learners' writing improves significantly over the course of the study and whether improvement is significant after either of the two learning contexts (AH and SA).

Section 6.2 reexamines the patterns of longitudinal improvement observed at the group level for participants with different levels of proficiency, in order to address RQ2.

RQ2. Is writing improvement different for learners with different initial levels of proficiency?

We consider initial level in two ways: as a function of participants' writing scores at the beginning of the study (initial writing proficiency) and as a function of their performance on the grammar and cloze tests, an external measure of their lexico-grammatical proficiency. We first explore the relationship between these two indices of proficiency, and then divide participants into high and low proficiency groups based on their performance in each area at T1 and explore the interactions between progress over time and initial proficiency.

In section 6.3 we compare the learners' essays to those written by native speakers, to test the statistical significance of patterns observed in the descriptive analysis and respond to RQ3.

RQ3. How do learners' essays compare to those of native speakers, in terms of the perceived quality of their essays and in terms of FLACC?

We use between-groups comparisons to determine whether native speakers perform significantly better than learners in terms of perceived quality and in terms of FLACC and to determine which characteristics differentiate between the two groups.

Finally, in section 6.4 we explore the relationships between FLACC measures, writing quality and grammar and cloze scores, to determine if our predictions for each set of measures are met, in response to RQ4.

RQ4. Do FLACC measures have the predicted relationships with a) writing quality and b) lexico-grammatical proficiency?

All statistical analysis reported in this chapter was carried out using SPSS version 15, and the terminology and conventions used to describe and report statistical tests were selected following the recommendations of Field (2005). An alpha level of .05 was adopted for all statistical tests.

6.1 Longitudinal changes after AH and SA contexts

In this section we explore data in response to *RQ1* and consider longitudinal development in learners essays, over time and after the AH and SA contexts respectively. We consider improvement in terms of perceived quality and in terms of objective measures of FLACC, and explore each set of dependent variables separately, in two subsections. First, in Section 6.1.1, we evaluate improvement in the perceived quality of learners' texts by exploring longitudinal changes in qualitative writing scores awarded by trained raters using the ESL Composition PROFILE. We present descriptive statistics for the full corpus and consider the baseline levels of proficiency in relation to the normative data published in Jacobs et al. (1981) before conducting statistical analysis to determine whether the learners' scores change significantly over time. We primarily focus on Total scores, but also consider the additional information provided by the 5 component scores (Content, Organization, Vocabulary, Language Use and Mechanics) and use effect size to evaluate the extent to which all component scores change over time. Next, in section 6.1.2 we consider changes in the domains of fluency, lexical diversity and sophistication, accuracy, complexity and cohesion (FLACC). We consider the changes observed in each domain in turn, first presenting descriptive statistics and exploring variables and then conducting statistical analysis to evaluate longitudinal changes in the L2 corpus.

In both sections we begin by presenting descriptive statistics for the corpus as a whole, including both the native speakers and L2 participants; however statistical analysis in this section focuses exclusively on longitudinal changes, and statistical comparisons with native speakers are reserved for section 6.3.

6.1.1 Changes in perceived quality

This subsection reports the results of qualitative evaluations carried out by trained raters using the analytic scoring rubric developed by Jacobs et al. (1981) and addresses *RQ1a. Do qualitative writing scores improve significantly after either the AH or SA learning contexts?* We first conducted descriptive analyses and considered participants' levels of ability at the beginning of the study, and then we conducted statistical analysis to determine whether there is significant improvement over time, after either the AH or SA context, and whether one context appears more beneficial than the other at the group level.

Reported scores were first considered in relation to the interpretative guide provided by Jacobs et al. (1981), which equates PROFILE scores with ESL skill levels, as described in Chapter 5. This revealed that all learners scored within the "High Intermediate" to "High Advanced" range, with the majority scoring between 74-82 points, thus falling into the "Low Advanced" category. Roughly the same distribution was found for the 30 essays written at T1, although there were relatively fewer scores in the "Advanced" range, and no scores in the "High Advanced" range. In contrast, all 28 of the native speakers' essays fell into the "Advanced" or "High Advanced" range, with the majority scoring between 92-100 points. The distribution of participants by skill level is reported in Table 6.1.

Table 6.1. Number of essays in each skill level based on Total PROFILE score

Score Range	Level	# L1 Essays	# L2 Essays *(#T1)*
92-100	High Advanced	18	5
83-91	Advanced	10	23 *(6)*
74-82	Low Advanced		50 *(18)*
65-73	High Intermediate		12 *(6)*

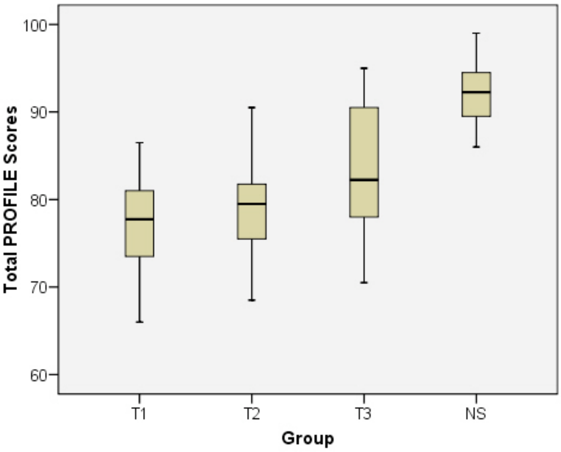
A look at the raw data reveals that the learners' mean and median scores increased at each data collection time. (Means are reported in Table 6.2; median scores increased from 77.75 at T1, to 79.5 at T2, to 82.25 at T3). There was also a relatively larger standard deviation at T3, indicating greater variability in learners' scores after the SA context, and a relatively larger difference between scores achieved by learners and native speakers, in comparison to differences between learners at any two data collection points. These group differences are illustrated in Figure 6.1.

Table 6.2. Descriptive Statistics: PROFILE scores by group (Mean/SD)

	Mean (SD)			
SCORE	T1	T2	T3	NS
Content	23.8 (2.3)	24.3 (1.9)	25.8 (2.3)	27.2 (1.5*)*
Organization	15.7 (1.3)	16.2 (1.4)	17.1 (1.5)	18.1 (.82)
Vocabulary	15.4 (1.4)	15.8 (1.4)	16.6 (1.9)	19.0 (.73)
Language Use	18.7 (1.2)	19.1 (1.4)	20.5 (1.7)	23.2 (1.1)
Mechanics	3.9 (.4)	4.1 (.3)	4.3 (.3)	4.7 (.36)
Total	**77.4 (5.4)**	**79.4 (5.9)**	**84.1 (7.1)**	**92.1 (3.5)**

The distribution of Total and component scores was then explored by group, using Q-Q plots and Kolmogorov-Smirnov tests to check the assumption of normality. Mechanics scores were not normally distributed for any of the 4 groups, since the range of scores obtained was very small and the vast majority of participants scored a 4 (61% overall, 76% in the L2 corpus), which led to high kurtosis values. Interestingly, all other component scores, as well as Total scores, were normally distributed at T1 and T2 but not at T3: graphical exploration of the T3 group revealed a bimodal distribution for all scores except for Language Use, with learners clustered at higher and lower score-ranges, suggesting that there might have been a split in the group means after SA, an issue that we flagged for further analysis (see section 6.1.1.1, below).

Figure 6.1. Box-plot: Mean Total scores by group



In order to facilitate between-groups comparisons and allow for the use of parametric tests, the decision was made to transform data and create a set of normally distributed variables. Scores were transformed in SPSS using Blom's formula, an operation that creates ranking variables based on proportion estimates, using the formula (r - 3/8) / (w + 1/4), where w is the sum of the case weights and r is the rank (Blom, 1958). Further K-S tests revealed that all of the transformed variables were normally distributed (p > .200) except for Mechanics, which remained problematic due to persistently high kurtosis. Mechanics scores were thus left out of between-groups comparisons with parametric tests and were evaluated independently.

After exploring the data descriptively, the observed differences in learners' scores were evaluated statistically. One-way repeated measures ANOVAs were carried out to determine whether there were significant changes in learners' Total and component scores over time, and standard repeated contrasts were used to determine whether significant changes occurred after AH (between T1 and T2) or SA (between T2 and T3), respectively. The set of normally distributed variables was used, and Mauchly's test confirmed that the assumption of sphericity was met for all variables, including Mechanics.

The results of the main ANOVAs (reported in Table 6.3) revealed a significant main effect of time on Total scores and all 5 component scores. The component score that showed the largest main effect of time, equivalent to the effect seen for Total scores, was Language Use, $\eta^2 = .17$, followed by Content, $\eta^2 = .16$; the component that showed the smallest effect of time was Vocabulary, indicating that, although Vocabulary scores did show significant improvement, they remained relatively more stable than the other component scores.

Table 6.3. RM-ANOVAs: Changes in PROFILE scores over time

	Mauchly's test			*Main ANOVA*				*p-values contrasts*[29]	
Score	χ^2	df	Sig.	N	η^2	*F(2,58)*	*p*	T1 - T2	T2 – T3
Cont.	.471	2	.790	30	.16	13.476	<.001	.283	< .001, *r* = .59
Org.	.254	2	.881	30	.15	11.893	<.001	.118	.005, *r* = .50
Voc.	.114	2	.944	30	.09	6.678	.002	.168	.030, *r* = .39
Lang.	.352	2	.839	30	.17	15.113	<.001	.253	.001, *r* = .62
Mech.	1.501	2	.472	30	.12	8.671	.001	.262	.003, *r* = .52
Total	**.281**	**2**	**.869**	**30**	**.17**	**15.054**	**<.001**	**.086**	**.001, *r* = .57**

Examination of focused contrasts revealed that the improvement registered after the AH context did not reach statistical significance (*p* = .086) but that improvement after the SA context did. Again, the component scores followed the same pattern observed for Total scores; however after SA the effect size seen for changes in Language Use and Content actually surpassed that seen for Total scores, indicating more dramatic changes in these two areas. As expected, Vocabulary scores continued to show the smallest effect size of all components, indicating that the improvement registered was less extreme than that observed for other components.

Together the results presented in this section confirm that, as a group, learners' writing showed qualitative improvement over the course of the study, and that they showed significant improvement after the SA context but not after the AH context. The larger range and bimodal distribution in raw scores at T3 suggested that there may have been considerable individual differences in score gains, however, and we elected to expand upon *RQ1* and explore these differences further, in subsection 6.1.1.1.

[29] Following Field (2005), main effects for RM-ANOVA are calculated as η^2, while effect sizes for focused comparisons are calculated as an *r*-coefficient, where $r = \sqrt{}((F(1, df_R)/F(1, df_R) + df_R))$.

6.1.1.1 Individual differences

The fact that there was a significant jump in scores after the SA but not after the AH context suggested that the SA context was more beneficial for the group as a whole; however, the bimodal distribution of scores after the SA indicated that not all individuals benefited equally from this context and that there may have been significant changes in both directions. The data were thus explored in more detail to determine to what extent the perceived advantage of the SA context was true for the entire group.

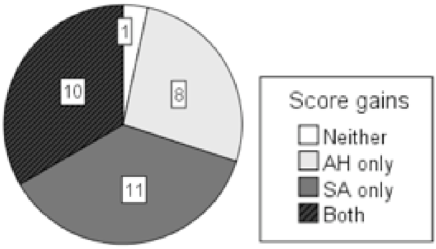
Three new variables were created to explore the distribution of gains among individual participants: *AH gains*, calculated as *T2 scores – T 1 scores*; *SA gains*, calculated as *T3 scores – T1 scores*; and *Overall gains,* calculated as *T3 scores – T1 scores*. These variables were then explored descriptively. As seen in Table 6.4, the majority of participants (26 out of 30) did indeed improve their scores over the course of the study, and the majority improved after each individual context; however a small number of participants saw a decline in scores over time, and 43% of participants did not improve their scores at all after the AH context.

Table 6.4. Description of analytic score gains by context

			Changes in Scores		
Context	Group	N	Mean (SD)	Median	Maximum
AH	gained	18	6.2 (4.0)	5.5	16.5
	declined	12	-4.3 (2.3)	-3.8	-8.8
SA	gained	21	8.1 (4.9)	9.0	15
	declined	9	-3.4 (3.6)	-2.0	-12.0
Overall	gained	26	8.3 (5.9)	6.6	25.8
	declined	4	-4.4 (4.0)	-3.8	-9.0

These data suggested that many participants improved in only one of the two contexts, which was confirmed by examining the Pearson's correlation between AH-gains and SA-gains (all variables measuring gains met the assumptions of normality, checked with Kolmogorov-Smirnov tests, and homogeneity of variance, checked with Levene's test). Analysis revealed that there was a significant negative correlation between the two variables, r = -.417, p = .022. That is, for many individuals, any improvement in one context corresponded to a lack of improvement in the other context. A graphical exploration of individual gains (classifying participants into 4 groups, based on whether they improved in the AH context only, in the SA context only, in both contexts, or neither context) revealed that, the group was relatively evenly split into participants in 3 groups, though a slightly greater number of participants improved only in SA (37%) versus only AH (27%) or in both contexts (33%), as seen in Figure 6.2.

Figure 6.2. Individual participants' Total score gains by context



The single participant whose scores did not improve at all in any context was eliminated from consideration, and repeated-measures ANOVAs were used to examine the longitudinal development of each of the three remaining groups. Again, ANOVA examined the main effect of time on participants' Total scores. Mauchly's test revealed that for the SA-only group, the assumption of sphericity was not met, and Greenhouse-Geisser estimates were used to adjust results, following Field (2005). Focused contrasts (Bonferroni-adjusted) were used to explore significant differences between all 3 data collection times and determine whether there was overall improvement in scores for all groups. The results of these ANOVAs are reported in Table 6.5 and the different patterns of development of each group are illustrated in Figure 6.3.

Figure 6.3 Changes in Total scores by groups with context-specific gains

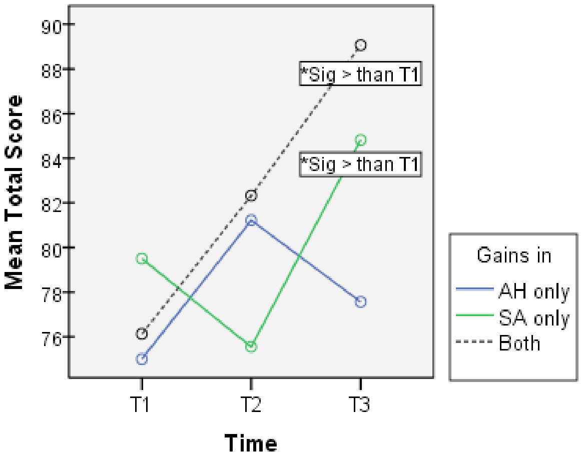


Table 6.5. Repeated-measures ANOVAs of groups with context-specific gains. Gains in AH group (n = 8), Gains in SA group (n = 11), Gains in both group (n = 10).

Gain	Mean Score (SD)			Main Effect of Time				*p*-values contrasts		
in	T1	T2	T3	*F*	Df_M, df_R	η^2	*Sig.*	AH	SA	T1vT3
AH	75.0	81.2	77.6	12.04	2, 14	.32	<.001	.001	.029	.112
SA	79.5	75.5	84.8	32.81	1.3, 12.9	.59	<.001	<.001	<.001	.004
Both	76.1	82.3	89.1	27.09	2, 18	.38	<.001	.003	.004	<.001

As we can see in Table 6.5, ANOVA revealed that there were significant changes over time for all groups; however examination of the contrasts revealed that this only translated into significant overall improvement for two of the three groups. That is, the group that improved only in the AH context, saw their score gains counteracted during the SA context, and although their scores appeared to be slightly higher at T3 than at T1, there was no statistically significant difference between these scores. In contrast, the group that improved only in SA, despite registering a decrease in mean scores after the AH context, saw enough improvement during the SA to leave them with significantly higher scores at T3 than at T1, just like the group that improved in both contexts. Overall, examination of individual differences confirm the impression given by the group data, that the SA is more beneficial to participants' writing than the AH context; however it also revealed that participants who benefited particularly from one context were less likely to benefit from the other, an issue that we will return to in section 6.2 in our examination of the effect of initial level of proficiency. The next section considers longitudinal changes in the domains of FLACC.

6.1.2 Longitudinal changes in FLACC

After examining improvement in qualitative scores, we next examined longitudinal improvement in learners' essays by looking at changes in the domains of FLACC, using the selection of measures described in Chapter 5. We explore these quantitative changes in 5 subsections, each devoted to measures in a given domain: fluency (section 6.1.2.1), lexical diversity and sophistication (section 6.1.2.2), accuracy (section 6.1.2.3), syntactic complexity and variety (section 6.1.2.4), and cohesion, along with the measure of pronoun density (section 6.1.2.5). As in the previous section, we first present descriptive statistics for the corpus as a whole, including the learners at each of the three times and the native speakers. We then conduct statistical analyses on the L2 data in order to answer *RQ1b. Are there significant changes in FLACC measures after either the AH or SA learning contexts?*

6.1.2.1 Fluency

The first domain examined was fluency, looking at the number of words and sentences produced in the allotted 30-minutes at each time. The learners' essays ranged from 111 to 454 words, with a mean of 237.8 words for the corpus as a whole. The number of sentences produced ranged from 3 to 21, with a mean of 10 sentences per essay. Examination of the mean number of words and sentences produced at each data collection time revealed a slight decrease from T1 to T2 and a larger increase from T2 to T3, and suggested an overall increase in fluency over time (see Table 6.6). Descriptive analysis revealed that #S was not normally distributed, and was positively skewed in at least one group. Skewness was corrected using a log-transformation (logarithm to the base of 10) and normally-distributed variables were used for all statistical analysis reported in this section.

Table 6.6. Descriptive statistics: fluency measures

	Mean (SD)			
	T1	T2	T3	*NS*
#W	233.4 (83.9)	212.4 (48.9)	266.8 (75.5)	*268.3 (96)*
#S	9.8 (3.9)	9.0 (2.5)	11.20 (3.8)	*11.4 (4.7)*

The significance of longitudinal changes for each of our fluency measures was tested statistically using one-way repeated measures ANOVAs with Time (T1, T2, T3) as the within-subjects factor. The main ANOVA was used to determine if changes were significant over time and standard repeated-measures contrasts were used to determine whether significant changes occurred after either the AH or SA learning contexts. Mauchly's test revealed that the assumption of sphericity was not met for either measure (for #W, $\chi^2 = .701$, $p = .007$, for #S, $\chi^2 = .777$, $p = .029$), and Greenhouse-Geisser estimates of sphericity were used to correct degrees of freedom, evaluate significance levels, and calculate effect sizes, following Field (2005). Results of the main ANOVA and contrasts (reported in Table 6.7) revealed that both fluency measures changed significantly over time and that the apparent decrease after the AH context was not significant but the increase after the SA context was significant.

Table 6.7. RM-ANOVAs: Longitudinal changes in fluency measures

ANOVA					*After AH*			*After SA*		
DV	*F*	Df$_M$, df$_R$	$\eta2$	*p*	*F*	*p*	*r*	*F*	*p*	*r*
#W	5.956	1.5, 44.7	.09	**.004**	1.750	.196	.24	22.253	**<.001**	.67
#S	3.597	1.6, 47.4	.07	**.044**	.324	.574	.11	9.959	**.004**	.51

Bonferroni-adjusted pairwise comparisons were then used to compare #W and #S at T1 and T3, which revealed that although participants appeared to improve in fluency from beginning to end of the study, this improvement did not reach statistical significance at the group level. When the data were analyzed qualitatively, it was revealed that 12 out of the 30 learners wrote longer essays at T2 than at T3, while 25 participants wrote longer essays at T3 than at T2, and 20 of the 30 participants wrote longer essays at T3 than at T1. These results suggested that for fluency, as for qualitative scores, there were considerable individual differences but that the SA context had a positive effect at the group level, while the AH context did not.

6.1.2.2 Lexical diversity and sophistication

Next we considered changes in lexical diversity and sophistication, via 4 different dependent variables: Guiraud's index (GI) of lexical diversity; Advanced Guiraud 1000 (AG1k), a general measure of lexical sophistication; noun hyponymy (HyN) and verb hyponymy (HyV), two more specific measures of lexical sophistication. Examination of the means for lexical variables (reported in Table 6.8) revealed that lexical diversity appeared to increase over time but in a non-linear U-shape similar to that seen for #W (this was unsurprising given that GI, like all measures of lexical diversity to at least some extent, is sensitive to text length). AG1 showed the same pattern, although the improvement from T1 to T3 was considerably less dramatic. HyV appeared to change very little in the L2 corpus, though the mean was slightly higher at T3 than at T1, while HyN showed a linear decrease over time, with learners moving progressively farther from native speaker levels.

Table 6.8. Descriptive statistics: lexical diversity and sophistication measures

	Mean (SD)			
Variable	T1	T2	T3	*NS*
GI	7.80 (.80)	7.37 (.70)	7.98 (.79)	*8.14 (.95)*
AG1k	1.17 (.40)	1.07 (.44)	1.18 (.33)	*1.63 (.54)*
HyN	4.36 (.42)	4.29 (.31)	4.27 (.33)	*4.45 (.34)*
HyV	1.33 (.15)	1.33 (.12)	1.34 (.13)	*1.37 (.14)*

One-way RM-ANOVAs were used to explore the statistical significance of these changes, with repeated measures contrasts in place to determine whether changes were significant after either context, the results of which are reported in Table 6.9. First, the distribution of variables was explored using Q-Q plots and Kolmogorov-Smirnov tests, which revealed that while GI and HyV met the assumption of normality, AG1k, and HyV presented either skewness or kurtosis problems in at least one group. These two variables were both transformed using Blom's formula, after which normality was met.

Table 6.9. RM-ANOVAs: Longitudinal changes in lexical diversity and sophistication measures

	ANOVA				*After AH*			*After SA*		
	F	Df_M, df_R	η^2	*p*	*F*	*p*	*r*	*F*	*p*	*r*
GI	8.087	2, 58	.11	**.001**	7.311	**.011**	.45	11.034	**<.001**	.52
AG1k	2.049	2, 58	.03	.138	--	--	--	--	--	--
HyN	.471	2, 58	.01	.627	--	--	--	--	--	--
HyV	.043	2, 58	.00	.958	--	--	--	--	--	--

Results revealed that there was a significant main effect of time on GI, with learners' texts showing greater lexical diversity over time, but no changes in lexical sophistication, as measured by these three variables. Contrasts revealed that, as with #W and #S, lexical diversity decreased significantly after the AH context but then increased significantly after the SA context. Bonferroni-adjusted pairwise comparison revealed that, as with #W and #S, the differences in lexical diversity between T1 and T3 were not significant, suggesting that the significant main effect of time was due to the U-shaped pattern of development in which after the SA participants recovered losses in lexical diversity after AH.

6.1.2.3 Accuracy

The next domain explored was that of accuracy. Examination of means (given in Table 6.10) suggested that accuracy increased from T1 to T3, but that learners remained far less accurate than native speakers: for example, the percentage of Error-free sentences steadily increased, yet at T3 only 34% of learners' sentences were error-free as opposed to 80% of native speakers' sentences. The total number of errors appeared to mirror the pattern seen for lexical diversity and fluency measures, in that the number of errors per word increased after the AH context but decreased again after the SA, with participants producing fewer errors per word at T3 than at T1. Type 1 spelling errors appeared to decline over time, while Type 2 spelling errors (those identified as "slips", as described in the previous chapter) remained constant.

Q-Q plots and Kolmogorov-Smirnov tests revealed that grammar, lexical and pragmatic errors (%Gre, %Lex, %Prag) had non-normal distribution in several groups, and the decision was made to evaluate these variables only descriptively while focusing statistical analysis on the umbrella measures: the total number of errors per word (%TotE) and the number of error-free sentences per sentence (%EFS). Both of these measures met the assumption of normality; however %TotE did not meet the assumption of homogeneity of variance. The variable was transformed with Blom's formula, which resolved this problem (Levene's test was non-significant at $F(3, 114) = 2.264$, $p > .05$). Both Type 1 and Type 2 spelling errors were positively skewed and normally distributed variables were created with log-transformations (base 10).

Table 6.10. Descriptive statistics: accuracy measures

	Mean (SD)			
Variable	T1	T2	T3	*NS*
%EFS	28.9 (19.8)	29.5 (20.7)	33.6 (17.7)	*80.1 (14.1)*
%TotE	5.9 (2.6)	6.2 (2.8)	5.5 (2.7)	*.85 (.6)*
%Gre	3.5 (2.2)	3.9 (2.3)	3.4 (1.9)	*0.57 (.5)*
%Lex	1.5 (1.2)	1.4 (.8)	1.3 (.9)	*0.14 (.2)*
%Prag	.92 (.8)	.87 (.7)	.87 (.6)	*0.14 (.2)*
%Sp1	.54 (.5)	.32 (.4)	.24 (.4)	*0*
%Sp2	.13 (.3)	.14 (.3)	.15 (.3)	*.08 (.2)*

The significance of the observed changes in means was tested statistically using one-way repeated-measures ANOVAs with time as the within-participants factor and each of our accuracy measures as the dependent variables. Mauchly's test confirmed that the assumption of sphericity was met in all cases. The main ANOVA, reported in Table 6.11, revealed that the only accuracy measure that showed a significant main effect was %Sp1. Type 1 spelling errors decreased significantly over time, and examination of focused contrasts revealed that the change in spelling errors after the AH context did not reach significance, $F(1, 29) = 3.883$, $p = .058$, but that there was a significant decrease in Type 1 spelling errors after the SA context, $F(1, 29) = 9.219$, $p = .005$, $r = .49$.

Table 6.11. RM-ANOVAs: Longitudinal changes in accuracy measures

	Mean			ANOVA			
Variable	T1	T2	T3	*F*	Df_M, df_R	η^2	*p*
%EFS	28.9	29.5	33.6	.640	2, 58	.01	.531
%TotE	5.9	6.2	5.5	1.037	2, 58	.02	.361
%Sp1	.54	.32	.24	10.048	2, 58	.13	**<.001**
%Sp2	.13	.14	.15	1.464	2, 58	.02	.240

Further descriptive analysis of changes in the overall percentage of errors by context revealed that only 40% of participants (12/30) improved their overall accuracy after the AH context, with 12 participants making fewer errors per word while 18 participants actually made more errors per word at T2 than at T1. In contrast, 70% of participants improved their accuracy after the SA context, with 21 participants making fewer errors per word at T3 than at T2. This corresponded to an overall increase in accuracy for 57% of participants, or 17 out of 30, as measured by the number of errors made per word (63% improved in accuracy when this was measured as the number of error-free sentences per sentence).

6.1.2.4 Syntactic complexity and variety

Next we explored changes in the domains of syntactic complexity and variety, beginning with the former. Our measures of syntactic complexity included 2 measures of sentence-level complexity (MLS and DC/S) and 2 measures of clausal complexity (MLC and SYNNP). As seen in Table 6.12, clausal complexity appeared to increase over time in the L2 corpus, with learners at T3 producing longer clauses, with more noun-phrase modification per clause. The length of sentences (MLS), as well as the number of dependent clauses per sentence (DC/S) decreased over time; while this suggested potentially less complex sentences, the decrease actually brought learners slightly closer to native speaker levels and thus suggested improvement. All 4 complexity measures violated the assumption of normality in at least one group when tested with Kolmogorov-Smirnov tests. Blom's formula was used to transform variables and create a set that was normally distributed; this worked for all measures except for SYNNP, which was transformed using a similar operation, but using Tukey's formula[30]. After transformation, all variables met the assumptions of both normality and homogeneity of variance, checked with Levene's test.

Table 6.12. Descriptive statistics and RM-ANOVAs: syntactic complexity measures

		Mean				ANOVA	
	Variable	NS	T1	T2	T3	*F*(2, 58)	*p*
Sentence Level	MLS	*24.9*	25.4	24.4	24.7	.024	.976
	DC/S	*.94*	1.5	1.4	1.3	.440	.646
Clause Level	MLC	*11.6*	9.2	8.8	9.6	3.385	**.041**
	SYNNP	*.89*	.72	.71	.76	1.980	.147

Changes in all measures over time were then tested statistically using a series of repeated measures ANOVAs, reported in Table 6.12 along with

[30] Ranks using the formula (r-1/3) / (w+1/3), (Tukey, 1962)

the descriptive statistics. Mauchly's test confirmed that the assumption of sphericity was met for all variables. As we can see, the only measure which showed a statistically significant main effect of time was mean length of clause (MLC), $p = .041$, $\eta^2 = .05$ and the observed changes in the other measures of complexity were not significant. Focused contrasts were examined to explore the changes in MLC after each learning context and these revealed that the decrease after the AH context was not statistically significant but that there was significant increase in clause length after the SA context, $F(1, 29) = 6.383$, $r = .42$.

Next we explored the two measures associated with syntactic variety: Temp, the measure of tense and aspect repetition, and StrutA, the measure of structural overlap between adjacent sentences. The means suggested that tense and aspect repetition decreased over time, indicating greater variety, while the amount of structural overlap increased after the AH and decreased again after the SA, but not to below initial levels. The distribution of both syntactic variety measures was explored using Q-Q plots and Kolmogorov-Smirnov tests. These revealed that Temp was normality distributed but that StrutA was positively skewed at T2. The latter measure was transformed using Blom's formula, after which the assumptions of both normality and homogeneity of variance (checked with Levene's test) were met. Repeat measures ANOVAs (reported in Table 6.13) revealed that a significant main effect of time was found for tense and aspect repetition (Temp), $p = .021$, $\eta^2 = .06$, but that the observed changes in StrutA were not significant. Exploration of contrasts for Temp revealed that, in contrast to a number of the other variables observed thus far, the changes after the AH context were significant, $F(1, 29) = 8.009$, $p = .008$, $r = .47$, but the changes after the SA context were not, suggesting that the improvement in syntactic variety occurred primarily during the AH context.

Table 6.13. Descriptive statistics and RM-ANOVAs: syntactic variety measures

	Mean (SD)				ANOVA	
Variable	*NS*	T1	T2	T3	*F*(2, 58)	*p*
Temp	*.76*	.84 (.11)	.77 (.12)	.78 (.10)	4.153	**.021**
StrutA	*.07*	.076 (.02)	.083 (.03)	.078 (.03)	.482	.620

6.1.2.6 Cohesion and pronoun density

Finally, we explored measures in the domain of cohesion as well as our measure of pronoun density, or the number of personal pronouns per word (DenPr). For cohesion we explored 3 measures of referential overlap, indicating more cohesive texts, as well as 5 measures of connectives. We explored the overall number of connectives per word (%Con) as well as the number of connectives associated with specific functions: additive (%AdCon), temporal (%TempCon), logical (%LogCon), and causal (%CausCon). Kolmogorov-Smirnov tests revealed that all 3 measures of referential overlap met the assumption of normality, as did DenPr, while the measures of connectives had positive skewness and kurtosis values in several groups. %Con, %CausCon, and %TempCon did not meet the assumption of normality at T3, while %LogCon did not meet the assumption of normality at T1, and %TempCon did not meet the assumption of normality at T2. We elected to transform these variables using Blom's formula so that parametric tests could be conducted and between-groups comparisons could be interpreted reliably.

All measures of cohesion appeared to decrease over time and move towards native speaker levels.

One-way ANOVAs with focused contrasts were used to test the significance of changes over time and identify any differences between the two learning contexts. These revealed that there was a significant main effect of time on the umbrella measure of connectives (%Con), p = .001, η^2 = .10, and on all types of connectives with the exception of temporal connectives (%TempCon); however the changes in referential overlap and pronoun density did not reach significance (see Table 6.14).

Table 6.14. RM-ANOVAs: Effect of Time on Cohesion

	Mean (SD)				ANOVA	
Variable	*NS*	T1	T2	T3	*F*(2, 58)	*p*
RefP	*.46*	.62 (.21)	.61 (.22)	.56 (.24)	.674	.514
RefA	*.63*	.74 (.21)	.70 (.21)	.70 (.18)	.524	.595
RefC	*.14*	.16 (.07)	.16 (.06)	.14 (.06)	1.181	.314
%Con	*.35*	.40 (.13)	.42 (.13)	.34 (.10)	7.567	**.001**
%AdCon	*.23*	.21 (.01)	.24 (.09)	.19 (.07)	5.030	**.010**
%CausCon[a]	*.09*	.15 (.08)	.15 (.07)	.10 (.05)	4.509	**.022**
%LogCon	*.15*	.25 (.12)	.23 (.10)	.18 (.07)	5.037	**.010**
%TempCon	*.04*	.05 (.04)	.03 (.03)	.04 (.04)	2.136	.127
DenPr	*75.4*	112.3 (31.7)	122.9 (42.3)	108.5 (30.0)	2.580	.084

[a] For %CausCon the assumption of sphericity was not met and degrees of freedom were corrected using Greenhouse Geisser estimates, *df* = (1.63, 47.34).

We explored focused contrasts for the umbrella measure %Con to determine whether significant changes had occurred after either learning context. These revealed that the changes after AH context were not significant, but that there was a significant decrease in the use of connectives after the SA context, $F(1, 29) = 19.451$, $p < .001$, $r = .63$, in the direction of native speaker norms.

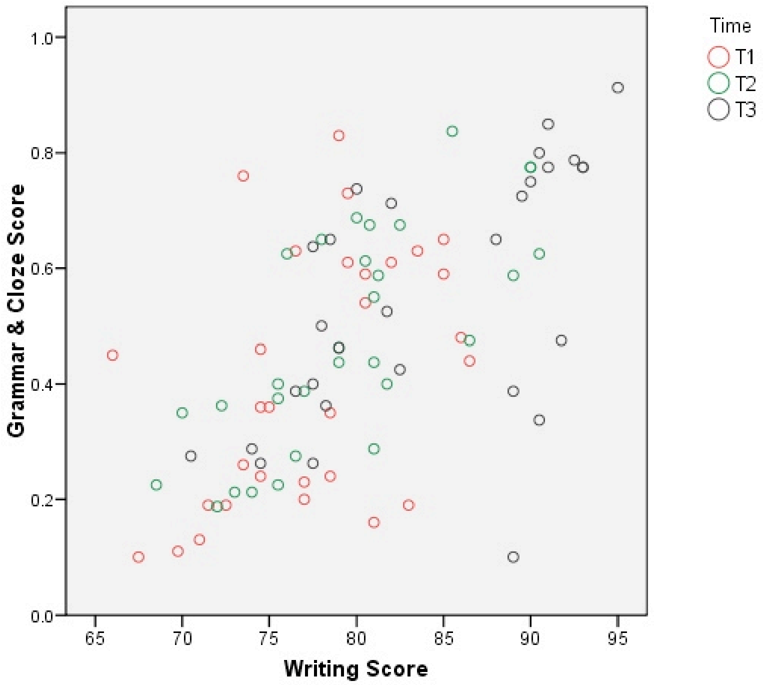
Together the results reported in this section indicate that participants' essays exhibited a number of significant changes in the domains of FLACC over the course of the study, and that the majority of these changes were context specific. After the AH context, learners essays significantly decreased in fluency, as measured by both the number of words and sentences, in lexical diversity, as measured by GI, all indicating a decrease in the quality of their texts; however they showed a significant increase in sentence variety (as measured by a decrease in tense and aspect repetition), which suggested that at least one aspect of their writing significantly improved. No significant changes were observed after the AH context in the domains of lexical sophistication, syntactic complexity, accuracy, or cohesion. After the SA context, learners essays significantly increased in fluency and in lexical diversity, in accuracy (as measured by spelling errors), in syntactic complexity (as measured by mean length of clause) and in cohesion, as measured by a decrease in the use of connectives suggesting that they were producing texts with fewer explicit markers of relationships between ideas and more cohesion gaps. Together the quantitative analysis indicates that, as suggested by the qualitative score changes, the SA context was relatively more beneficial to the group as a whole but that there were individual differences in performance which may have been masking significant differences at the group level in several domains. Based on the hypothesis that some of these individual differences may have been related to participants' proficiency levels, we opted to proceed directly to analysis of this factor, as reported in the next section.

6.2 Effect of initial level

In this section we reconsider the longitudinal changes in participants' essays in relation to their initial levels of proficiency, in response to RQ2. We considered initial level of proficiency in two different ways: initial level of writing proficiency (IWP), as measured by their qualitative scores at T1; and initial level of lexico-grammatical proficiency (IGP), as measured by scores on the grammar and cloze tests at T1.

Before grouping participants and conducting these analyses, we opted to examine the relationship between writing scores and grammar and cloze scores in the L2 corpus, in order to determine to extent to which these two measures were capturing different aspects of proficiency. The relationship between grammar and cloze test scores and qualitative writing scores was explored using Kendall's Tau coefficient, as there were a large number of tied ranks in the set of grammar and cloze scores and this method of correlation is considered preferable in such cases (see Field, 2005: p. 131). Analysis revealed that there were significant positive correlations between the grammar and cloze scores and writing scores at all three data collection times, and for the corpus as a whole, as illustrated in Figure 6.4.

Figure 6.4. Relationship between writing scores and LGP in the L2 corpus



For the full corpus of 90 L2 essays, we found that the correlation between grammar and cloze scores and Total writing scores was significant at τ = .478. Significant correlations were also found between grammar and cloze scores and all 5 component scores, with the strongest correlations seen for the 3 components that evaluated surface features related to linguistic control: Vocabulary, Mechanics, and Language Use (see Table 6.15). We also calculated the correlations between Total scores and grammar and cloze scores at each data collection time and found that it was stronger at T2 and T3, but was significant at all three times: at T1, τ = .328, p = .012; at T2, τ = .567, p <.001; and at T3, τ = .541, p <.001.

Table 6.15. Correlations between grammar and cloze scores and PROFILE component scores in the L2 corpus.

Score	Kendall's τ	N[31]
Content	.389(**)	89
Organization	.402(**)	89
Vocabulary	.527(**)	89
Language Use	.421 (**)	89
Mechanics	.437(**)	89
Total Score	.478(**)	89

** Correlation is significant at .01 (two-tailed)

The observation that the correlation between writing scores and grammar and cloze scores was significant, but only moderately so at T1 (τ = .328), confirmed our decision to explore these two aspects of proficiency separately in our examination of the effect of initial level.

6.2.1 Initial writing proficiency (IWP)

In this section we explored the effect of initial level of writing proficiency on longitudinal changes in response to *RQ2a. Are changes in perceived quality and FLACC different for participants with higher and lower initial writing proficiency (IWP)?* To explore the effect of initial level of writing ability (IWP), participants were classified based on the percentile rank of their Total score at T1. Three groups were created: low-IWP, whose mean scores fell into the "Upper Intermediate" category based on the normative guide provided in Table 6.1 (*M* = 71.7, *SD* = 2.9); mid-IWP, who fell into the "Low Advanced" range (*M* = 77.8, *SD* = 1.5); and high-IWP, whose scores were in the "Advanced" category (*M* = 83.3, *SD* = 2.3). We first examined whether IWP affected improvement in qualitative scores and then examined the effect on FLACC.

6.2.1.1 Effect on perceived quality

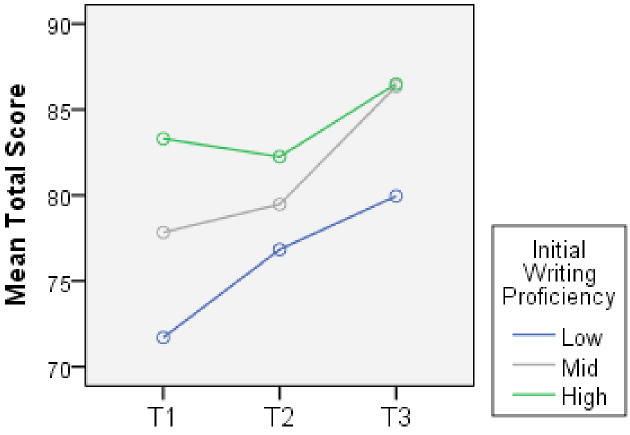
First, a mixed-design ANOVA with IWP as a between-participants factor was conducted to determine whether there was a significant interaction between Time and IWP on Total PROFILE scores, with repeated contrasts used to determine any context-specific interactions. This revealed that the interaction between Time and IWP was significant, at *F*(4, 54) = 2.58, *p* = .047; however contrasts revealed that the interaction was significant

[31] One participant did not take the cloze test at T3, and was eliminated from all analyses of lexico-grammatical proficiency

between T1 and T2, $F(2, 27) = 4.24$, $p = .025$ but not between T2 and T3, $F(2, 27) = .788$, $p > .05$, indicating that initial writing proficiency affected the amount of progress made after the AH context but not after the SA context.

Graphical and descriptive analysis suggested that while all three groups showed some improvement after the SA context only the two lower-level groups made progress after the AH context, with the higher-level group actually obtaining lower Total scores at T2 than at T1 (see Figure 6.5). This was not attributable to any kind of ceiling effect, as the mean score of this group (83.3) was still below the "High Advanced" range and below the scores obtained by the native speakers, as seen in Section 6.1.1. Descriptive statistics also suggested that the mid-level group made the greatest progress over time, reaping benefits in both contexts and appearing to converge with the higher-level group by T3.

Figure 6.5. Improvement in perceived quality in relation to IWP



These impressions were tested statistically by conducting separate one-way repeated measures ANOVAs for each of the three IWP groups. Again, the effect of Time on Total scores was evaluated in the main ANOVA and standard contrasts were used to determine if significant changes occurred after either learning context. Results revealed that while the low- and mid-level groups made statistically significant progress over time, consistent with the results observed for the whole group, similar improvement was not seen for the high level group (Table 6.16). The effect size was moderately larger for the mid-IWP group, confirming the impression that this group made the greatest gains over time. Examination of the contrasts revealed a qualitative difference between the lower two groups as well: while both groups appeared to make some progress after each learning context, improvement for the lower level group was only statistically significant after the AH context, while improvement for the mid-level group was only significant after the SA context.

Table 6.16. RM-ANOVAs: Changes in Total scores of learners with high (n=10), low (n=11), and mid (n=9) IWP

	Mauchly's test			*Main ANOVA*			*p-values contrasts*	
IWP	χ^2	*df*	Sig.	η^2	*F(2,58)*	*p*	T1 - T2	T2 – T3
Low	.149	2	.928	.24	9.225	.001	**.017, *r* = .61**	.173
Mid	.034	2	.983	.28	10.570	.001	.506	**.007, *r* = .79**
High	.720	2	.698	--	2.203	.139	.423	.099

Finally, independent-means t-tests were used to compare the scores of the three groups at each time and determine whether the mid-level group did, indeed, converge with the high-level group by T3. Since the grouping variable was created using T1 scores, scores were inherently significantly different at the beginning study. Results of t-tests at T2 and T3 (reported in Table 6.17) revealed that after the FI the low-IWP writers caught up with the mid-level group, and that the mid- and high-level groups converged at T3, with both groups continuing to score significantly higher than the low-level group.

Table 6.17. T-tests comparing Total scores of learners with high, low, and mid IWP

		T1			T2			T3	
	t	*df*	*p*	*t*	*df*	*p*	*t*	*df*	*p*
High vs. Low	9.286	19	**<.001**	2.134	19	**.046**	2.441	19	**.025**
High vs. Mid	6.788	17	**<.001**	1.127	17	.276	-.051	17	.960
Mid vs. Low	5.239	18	**<.001**	1.199	18	.246	2.360	18	**.030**

While context-specific differences could have been partially due to variability in individual performance, overall these results indicated that initial writing level had an effect on improvement over time and by context. Higher-level writers (those classified as “Advanced” at T1) seemed to progress less than their peers overall. Mid- and high level writers benefited more from the SA than the AH context, and only the lower-level writers (those classified as “Upper Intermediate”) saw significant progress after the AH context.

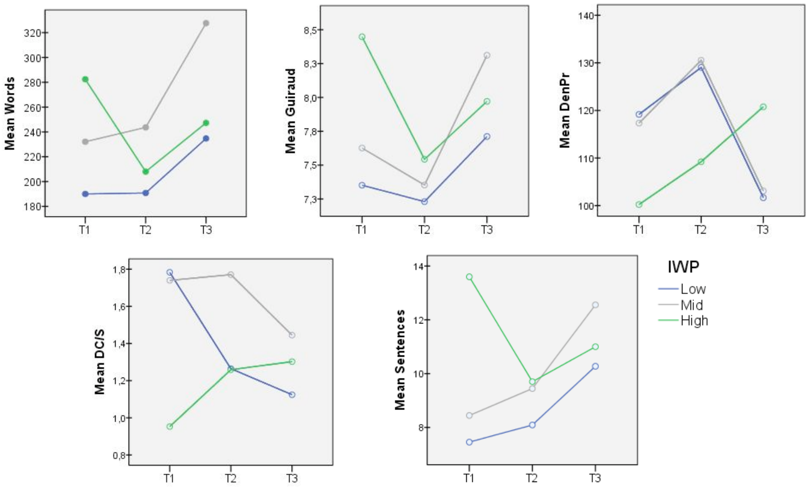
6.2.1.2 Effect of IWP on improvement in FLACC

We then explored the effect of IWP on improvement in the domains of FLACC. A mixed-design ANOVA was carried out with IWP as the between-participants factor in order to determine whether there were any significant interactions between Time and IWP on changes in FLACC measures over time. Significant interactions were found for both fluency measures and for lexical diversity as well as for DC/S and DenPr, as reported in Table 6.18 (sphericity assumed for all measures). These interactions are illustrated in the Figure 6.6 below.

Table 6.18. ANOVA: FLACC variables showing significant interaction between Time x IWP

	Interaction *Time* x *IWP*			*p*-values for contrasts	
Measure	*F*(4, 54)	*Sig.*	η^2	AH	SA
#W	3.565	**.012**	.05	.052	.247
#S	3.800	**.009**	.06	**.026**	.498
GI	3.228	**.019**	.05	.092	.205
DC/S	3.152	**.021**	.05	.055	.415
DenPr	2.762	**.037**	.04	.970	**.025**

Figure 6.6. Significant interactions between Time and IWP (Words, Guiraud, DenPr, DC/S, Sentences)



As we can see, in terms of fluency, at the beginning of the study the high level group wrote essays with more words and sentences than either of the other two groups. A one-way ANOVA revealed that at T1 there were significant differences in #W, F(2, 29) = 3.786, p = .036, and Bonferonni post-hoc tests revealed that the high level group produced significantly more words than the low level group, (p = .031) but that neither group differed significantly from the mid-level group. At T3, there were again significant differences between the groups, F(2, 29) = 5.598, p = .009; however now the mid-level group produced significantly more words than the high level group (p = .039) and the low level group (p = .012), and there were no significant differences between these two groups. The pattern for sentences and GI was similar, with the mid level group appearing to overtake the high level group, who sharply declined after the AH context, while the other two groups appeared to plateau or moderately improve.

For DC/C, both the mid- and low-level groups saw a decrease over time, producing fewer dependent clauses at the end of the study than at the beginning, moving in the direction of the native speakers (who produced a mean of .94 dependent clauses per sentence), while the high level group saw a moderate increase in the number of dependent clauses. For pronoun density, a similar pattern was observed, with the low and mid level groups showing an increase in the proportion of personal pronouns over the AH context but a decrease after the SA, moving in the direction of native speaker norms and indicating improvement, while the high level group used a greater proportion of pronouns at the end of the study than at the beginning. Overall the results in this section indicate that the high IWP group improved less than the other two groups, and that the mid IWP group showed the greatest improvement over time, ending the study with greater fluency, lexical diversity than the high level group, despite beginning at a relative disadvantage, and reducing the number of dependent clauses per sentence and the proportion of personal pronouns, moving in the direction of native speaker norms. These results confirm the patterns observed for qualitative scores and suggest that participants who began in the “low advanced” range at T1 were able to reap a greater benefit than those in either the “advanced” or “upper intermediate” ranges.

6.2.2. Initial Lexico-Grammatical Proficiency

In this section we reexamined the effect of initial level as a function of initial level of lexico-grammatical ability, which was found to correlate modestly with writing ability but presumed to represent a different dimension of proficiency, in order to answer *RQ2b Are changes in perceived quality and FLACC different for participants with higher and lower initial lexico-grammatical proficiency (IGP*)*?* In order to analyze the effect of IGP, learners were divided into groups based on the percentile ranks of their combined grammar and cloze scores at T1. Participants were initially divided into 3 groups, to match the analyses done for IWP; however due to a large number of tied ranks and an uneven, bimodal split in the middle group, the decision was made to eliminate mid-level scorers and compare only the high and low IGP groups. The high-IGP group scored a mean 66% on the grammar and cloze tests at T1, with a maximum score of 83%, while the low-IGP learners scored a mean of just 18% (see Table 6.19). To put these low scores in perspective, the 28 native speakers scored a mean of 83%, with a maximum of 95%.

Table 6.19. Descriptive statistics: T1 grammar & cloze scores of high- and low-IGP learners

IGP Group	#	Mean	SD	Min	Max
Low	11	18.0	5.0	10	24
High	10	66.3	8.2	59	83

6.2.2.1 Effect of IGP on improvement in qualitative scores

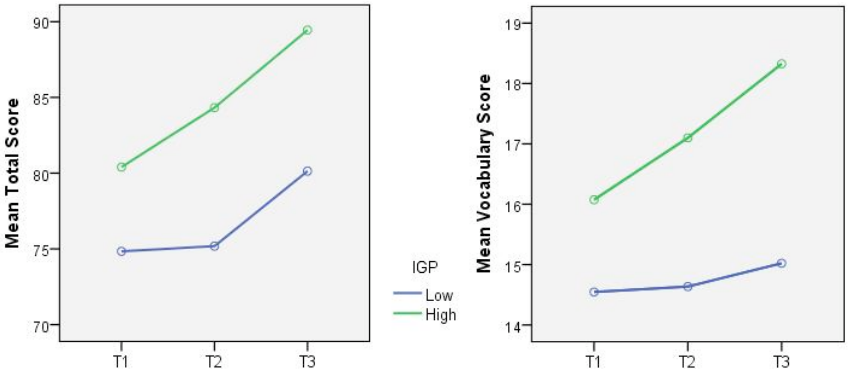
First, we explored the effect of IGP on changes in perceived writing quality, as measured by Total PROFILE scores. One-way repeated measures ANOVAs were carried out to examine the effect of time on the Total scores obtained by each group, again using focused contrasts to identify significant differences between data collection times. Mauchly’s test confirmed that the assumption of sphericity was met for Total scores for both groups. Results revealed a significant main effect of time for both groups (see Table 6.20), with a larger effect size found for the high-IGP group indicating that, learners with more advanced lexico-grammatical proficiency at T1 made more progress over time than did their lower-proficiency peers. The high-IGP group showed no significant jumps in Total scores after either the AH or SA context independently, despite showing significant overall improvement; in contrast, the low-IGP group showed a significant jump in scores after the SA context.

Table 6.20. RM-ANOVA: Longitudinal changes in high (n=11) and low (n=10) IGP learners.

IGP	Mean (SD)			Main Effect of Time				*p*-values contrasts	
	T1	T2	T3	*F*	*Df*$_M$, *df*$_R$	η^2	*Sig.*	AH	SA
Low	74.8	75.2	80.1	5.63	2, 20	.18	.011	.839	.043
High	80.4	84.3	89.5	8.32	2, 18	.24	.003	.072	.059

Despite these apparent differences, a mixed-design ANOVA with IGP as the between-participants factor revealed that there was no significant interaction between Time and IGP on learners' Total scores, *F*(2, 38) = 1.135, *p* = .332, indicating that the apparent variation in progress was probably not due to a systematic difference between groups. Mixed-design ANOVAs were also run on the component scores, however, and a significant interaction between Time and IGP was found for Vocabulary scores, *F*(2, 38) = 3.651, *p* = .035, η^2 = .08 (sphericity assumed). Examination of means revealed that high-IGP learners made vocabulary gains in both contexts, which mirrored the gains seen for Total scores, low-IGP learners showed virtually no gains on the vocabulary component.

Figure 6.7 Total and vocabulary scores of high-IGP and low-IGP groups



For the low IGP group, the component score that saw the greatest effect of time was Language Use, *F*(2, 20) = 11.775, *p* <.001, η^2 = .27, followed by Mechanics, *F*(2, 20) = 5.884, *p* <.010, η^2 = .19. There were no significant effects found for Content, Organization, or Vocabulary, although the effect of Time on Content was right at the cut-off point, *F*(2, 20) = 3.492, *p* = .05. In contrast, the high IGP group saw significant changes in all components over time and the relative effect sizes of the different components was reversed. That is, the largest effect was found for Vocabulary, *F*(2, 18) = 8.580, *p* = .002, η^2 = .24, followed by Organization, *F*(2, 18) = 7.129, *p* = .005, η^2 = .22, and then Content, Mechanics, and Language use, respectively.

The scores obtained by the two IGP groups were compared statistically using Mann-Whitney's non-parametric U-tests, which revealed that learners with high IGP obtained significantly higher writing scores at all 3 times (see Table 6.21). Non-parametric tests were used as there were several violations of homogeneity of variance, as measured by Levene's test, and given the small group sizes we deemed it best not to further transform the data.

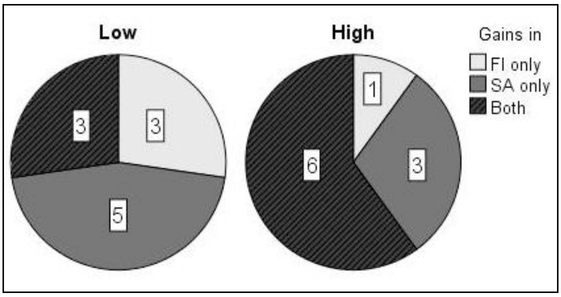
Table 6.21. Mann-Whitney test: Total and component score differences between high-IGP and low-IGP groups

		High-IGP Median	Low-IGP Median	*U*	*Z*	*Sig.*
T1	Con	25	22.5	34.5	-1.450	.147
	Org	16.38	15.25	34.0	-1.489	.136
	Voc	16	14.5	13.0	-2.975	**.003**
	Lang	19.25	18	22.0	-2.350	**.019**
	Mec	4	4	28.0	-2.385	**.017**
	Tot	80	74.5	20.0	-2.467	**.014**
T2	Con	26.25	23.5	16.5	-2.729	**.006**
	Org	17.38	15.5	17.5	-2.665	**.008**
	Voc	17.25	14.5	4.5	-3.574	**.000**
	Lang	19.5	18	6.0	-3.482	**.000**
	Mec	4	4	21.0	-2.747	**.006**
	Tot	83.25	75.5	11.5	-3.066	**.002**
T3	Con	27.5	24.5	19.5	-2.531	**.011**
	Org	18.25	16.5	21.5	-2.378	**.017**
	Voc	18.5	15	3.5	-3.638	**.000**
	Lang	21.5	19	35.5	-1.382	.167
	Mec	4.5	4	19.0	-2.740	**.006**
	Tot	91	77.5	11.0	-3.104	**.002**

Examination of the component scores revealed that at T1 there were no significant differences in Content and Organization and that score differences between high and low level groups was primarily due to higher scores for components assessing surface structures and linguistic control. Interestingly, at T3 this pattern was reversed: high IGP writers scored significantly higher on all components except for Language Use. This suggested that the lower IGP students had caught up in terms of linguistic control but were lagging behind in all other aspects of their writing.

As with the larger group, there were considerable individual differences within the high and low IGP groups, and individuals in each group appeared to benefit differently from different contexts. A slightly larger number of participants in each group improved in the SA than in the AH context, as seen in Figure 6.8. In the low-IGP group, 10 of the 11 participants made gains from the beginning to the end of the study: 6 improved in the AH context and 8 improved in the SA. In the high-IGP group, 9 out of 10 participants improved their scores over the course of the study: 7 improved in the AH context, and 9 improved in the SA.

Figure 6.8. Individual gains of participants based on IGP



6.2.2.2 Effect of IGP on improvement in FLACC

Next we explored the effect of IGP on improvement in the domains of FLACC. Again, the 2 groups were first compared at T1, looking at the full set of FLACC measures, this time using independent-means t-tests to compare learners with high and low IGP (Table 6.22).

Table 6.22. *T*-tests: FLACC differences between high-IGP and low-IGP groups

Fluency & Lexical characteristics				Accuracy			
	t	*p*	*r*		*t*	*p*	*r*
#S	.790	.439	.18	%EFS	3.231	**.004**	.60
#W	.863	.399	.19	%TotE	-6.343	**.000**	.82
GI	3.080	**.006**	.58	%Sp1	-2.314	**.032**	.47
AG1k	2.325	**.031**	.47	%Sp2	.212	.835	.05
HyN	-.264	.795	.06				
HyV	-.500	.623	.11				

Syntactic Complexity & Variety				Cohesion & Pronoun usage			
	t	*p*	*r*		*t*	*p*	*r*
MLS	-.378	.709	.09	RefP	-2.452	**.024**	.49
DC/S	-1.678	.110	.36	RefA	-1.983	.062	.41
MLC	4.016	**.001**	.68	RefC	-2.141	**.045**	.44
SYNNP	1.934	.068	.41	%Con	-1.711	.103	.37
StrutA	-1.152	.264	.26	DenPr	-2.851	**.010**	.55
Temp	-1.475	.157	.32				

There were between-groups differences in lexical diversity and sophistication, in all accuracy measures except for Type 2 spelling errors, in clause length, referential overlap, and pronoun density. After observing these differences, a mixed-design ANOVA was conducted, with time as the within-participants factor and IGP as the between-groups factor, testing changes in all FLACC measures; however no significant interactions were found, suggesting that there were no systematic differences in the developmental patterns of each group.

6.3 Comparisons with native speakers

In this section we conduct analyses comparing native speakers essays to those of the L2 learners, in order to test the statistical significance of differences observed in the descriptive analyses presented in Section 6.1 and obtain a response to RQ3. As in previous sections, we first conduct analyses looking at qualitative writing scores (in section 6.3.1), and then conduct analyses on quantitative measures, in the domains of FLACC (section 6.3.2).

6.3.1 Differences in qualitative scores

Consideration of the descriptive data indicated that, despite improving steadily over the course of the study, learners continued to score below native speakers, on average, and demonstrate greater variability in scores. This was tested statistically in response to *RQ3a. How do learners' essays compare to those of native speakers in terms of perceived quality?*

Because learners were found to have significantly improved their qualitative scores over time, comparisons in this section focused on comparing the 30 essays produced by the learners at T3 to the 28 essays produced by native speakers; a series of independent-means *t*-tests was carried out to compare the Total and component scores achieved by these two groups (descriptive statistics reported above in Table 6.2, section 6.1.1). The transformed, normally-distributed, variables were used and the assumption of homogeneity of variance was evaluated with Levene's test, which revealed that the Organization and Vocabulary component scores, did not meet the assumption of homogeneity of variance (p = .017 and p = .013, respectively). Following, Field (2005), the *t*-statistic and degrees of freedom were corrected to account for this, and the *p*- and *r*-values reported (see Table 6.23) reflect these changes. Mechanics scores were eliminated from this stage of analysis and were compared separately with Mann-Whitney's non-parametric *U*-test, given the problems with kurtosis described in Section 6.1.1.

Table 6.23. *T*-tests: Differences in PROFILE scores between NS and learners at T3

Component	*t(56)*	*df*	*p*	*r*
Content	2.585	56	.011	.33
Organization	2.839	48.21	.008	.38
Vocabulary	6.310	48.75	<.001	.67
Language Use	7.349	56	<.001	.70
Total	5.270	56	< .001	.58

The results of the *t*-tests confirmed that native speakers' qualitative scores were significantly higher than those of learners, even at T3. Predictably, the largest effect of L1 background was seen for the Language Use component score, followed by the Vocabulary scores, while the Organization and Content scores showed relatively smaller effects; the effect size found for Mechanics was squarely in the middle of these two groups: Mann-Whitney's *U*-test showed that Mechanics scores were significantly different at, $U = 183, p < .001, r = .52$.

Together results in this section reveal that learners did not converge with native speakers by the end of the study, and that their qualitative scores were significantly lower even at T3. Consideration of the relative effect sizes of component scores indicated that the persistent differences between learners and native speakers were primarily due to differences in Language use and Vocabulary, as well as surface features (Mechanics) and less attributable to differences in Content and Organization, although native speakers achieved significantly higher scores in all areas.

6.3.2 Differences in FLACC

Next the linguistic characteristics of learners' essays were compared statistically to those of native speakers, in order to answer *RQ3b. How do learners' essays compare to those of native speakers in terms of FLACC?* Given that a number of variables seemed to progress or change in a non-linear fashion, the learners at T2 and the learners at T3 were compared to native speakers separately using independent-means *t*-tests, reported in Table 6.24. (The descriptive statistics for these measures are given in Section 6.1.2: Fluency is reported in Table 6.6; Lexical diversity and sophistication in Table 6.8; Accuracy in Table 6.10; Syntactic complexity in Table 6.12; Syntactic variety in Table 6.13; and Cohesion in Table 6.14). The set of normally distributed variables was used and homogeneity of variance was checked with Levene's test. When comparing native speakers to T2 learners, Levene's test was significant for words and

%EFS (indicating that assumption of homogeneity of variance was not met) and statistics were adjusted to account for this violation. The same was true for %TotE when comparing native speakers to T3 learners.

Table 6.24. *T*-tests comparing NS and learners on FLACC measures

	NS vs. Learners at T2				NS vs. Learners at T3			
Variable	*t*	*df*	*p*	*r*	*t*	*df*	*p*	*r*
#W	2.765	39.5	**.009**	.40	-.136	56	.892	.02
#S	2.102	56	**.040**	.27	.067	56	.947	.01
GI	3.532	56	**.001**	.43	.713	56	.479	.09
AG1k	5.109	56	**<.001**	.56	4.404	56	**<.001**	.51
HyN	1.714	56	.092	.22	2.005	56	.050	.26
HyV	1.384	56	.172	.18	.784	56	.437	.10
%EFS	10.948	51.3	**<.001**	.84	11.040	56	**<.001**	.83
%TotE	-10.175	56	**<.001**	.81	-9.813	48.8	**<.001**	.81
%Sp1	-5.364	50.1	**<.001**	.60	-3.060	48.4	**.004**	.40
%Sp2	-2.378	56	**.021**	.30	-.629	56	.532	.08
MLS	.197	56	.845	.03	.051	56	.959	.01
DC/S	-3.456	56	**.001**	.42	-2.937	56	**.005**	.37
MLC	6.456	56	**<.001**	.65	4.026	56	**<.001**	.47
SYNNP	4.452	56	**<.001**	.51	2.785	56	**.007**	.35
StrutA	-1.685	56	.098	.22	-.912	56	.365	.12
Temp	-.346	56	.730	.05	-.931	56	.356	.12
RefP	-2.275	56	**.027**	.29	-1.458	56	.150	.19
RefA	-1.209	56	.232	.16	-1.238	56	.221	.16
RefC	-1.479	56	.145	.19	-.098	56	.922	.01
%Con	2.461	56	**.017**	.31	-.314	56	.755	.04
%AdCon	.778	56	.440	.10	-1.578	56	.120	.21
%CausCon	4.627	56	**<.001**	.53	1.596	56	.116	.21
%LogCon	3.634	56	**.001**	.44	1.080	56	.285	.14
%TempCon	-.668	56	.507	.09	.207	56	.837	.03
DenPr	-4.993	56	**<.001**	.56	-4.316	56	**<.001**	.50

As we can see in Table 6.24, NS produced significantly more words and sentences than the learners at T2 but not at T3. That is, after the SA learners converged with native speakers on fluency measures. The same was found for lexical diversity, as measured by Guiraud's index. The native speakers scored significantly higher than both T2 and T3 learners on AG1k, but not on other lexical sophistication measures. As expected, natives performed significantly better than learners at both T2 and T3 in terms of overall accuracy, although the T3 learners did converge with native speakers on Type 2 errors, indicating that there were fewer slips after the SA. As for syntactic complexity, the native speakers wrote significantly more complex clauses than learners, although effect sizes indicated this difference was less profound after the SA, which is in line with the observation that clausal complexity increased after this context. NS produced significantly fewer clauses and dependent clauses per sentence than learners at either time, and there were no differences in MLS or either measure of syntactic variety. (The native speakers did use significantly more syntactic variety than the learners at T1, $t(56) = 2.758$, $p = .008$, $r = .35$) Native speakers used less referential overlap than learners at T2 but not at T3, suggesting that after the SA learners were able to use pronominal reference in more appropriate, native-like ways. This was supported by the observation that, although the native speakers scored better than both groups of learners on the pronoun-density measures, there was a slightly larger effect size observed for the T2 groups. Finally, while native speakers produced significantly fewer connectives per word than learners after the FI, learners after the SA converged with native speakers on this measure.

6.4 Relationship between FLACC, writing quality, and LGP

This final section explores the relationship between FLACC measures, measures of writing quality, and grammar and cloze scores, in order to test our predictions for these measures and respond to *RQ4. Do FLACC measures have the predicted relationships with a) writing quality and b) lexico-grammatical proficiency?* Section 6.4.1 focuses on sub-question *RQ4a*, while section 6.4.2 focuses on sub-questions *RQ4b* and *RQ4c.*

6.4.1 Relationship between FLACC and qualitative scores

This section focuses on the relationship between FLACC measures and qualitative writing scores, in response to *RQ4a. Which FLACC measures are significantly correlated with qualitative writing scores and which discriminate between high and low scoring learners?*

First, correlations between Total writing scores and FLACC variables are carried out for each domain in turn, looking at relationships in the full corpus (N = 118) and in the L2 (n = 90) and L1 (n = 28) corpuses independently. Spearman's non-parametric rank correlation coefficient (r_s) was used to quantify these relationships, allowing for analysis of raw measures, some of which did not meet the assumption of normal distribution as described in section 6.1.2. Next, we considered which FLACC measures discriminated between high and low scoring learners. The full corpus of 90 L2 essays was divided into 4 groups based on the percentile rank of their qualitative scores, and then the bottom-scoring and top-scoring groups were compared. Although the repeated-measures design was ignored when creating these groups, and each essay was treated as a single case, the distribution of participants by data collection time is given in Table 6.25, along with a description of the scores achieved by each group. As we can see, the majority of essays in the low-scoring group were written by participants at T1, while the majority of essays in the high-scoring group were written by participants at T3. Based on the skill levels identified in the first section (see Table 6.1), we can see that the low-scoring group scored on average in the "upper intermediate" skill level, although several participants reached the "low advanced" range, scoring above 73/100. The high-scoring group, in contrast, fell at the high end of the "advanced" range, or in the "upper advanced" range.

Table 6.25. Distribution of learners with high and low PROFILE scores

	Distribution				Qualitative Scores			
Group	N	%T1	%T2	%T3	Mean	Median	Min	Max
Low	21	12	6	3	72	72.5	66	75
High	22	2	6	14	89.9	90	85.5	95

In the subsections below, we first consider correlations and then consider between-groups comparisons for each domain in turn. For practical reasons, given the structure of our databases, we elected to group fluency measures together with lexical diversity and sophistication measures in the remaining stages of analysis. All between-groups analysis in this section was conducted using Mann-Whitney's U-tests, given the relatively small sample sizes and the fact that several quantitative measures did not meet the assumptions of normality and homogeneity of variance required for comparison with parametric tests. For practical reasons given the structure of our databases, we grouped lexical diversity and sophistication measures along with fluency measures in the remaining subsections.

6.4.1.1 Fluency, lexical diversity and sophistication

We first explored variables in the domains of fluency, lexical diversity and sophistication. It was predicted that all six of the specific measures in these domains would all correlate positively with writing scores, based on previous research. We conducted Spearman's correlations for the full corpus, as well as for the L1 and L2 corpuses separately, in order to test these hypotheses, the results of which are reported in Table 6.26.

As predicted, both fluency measures were positively correlated with Total writing scores in the full corpus, and in the L2 corpus; the same was true for the measures of lexical diversity (GI) and the general measure of lexical sophistication (AG1k). The measures of hyponymy and concreteness, on the other hand, were not found to have any significant relationship with writing scores in the full corpus or in the L2 corpus, in contrast to expectations.

Table 6.26. Spearman correlations: Total writing scores, fluency and lexical measures

	Full corpus		L2 Corpus		L1 Corpus	
Measure	r_s	N	r_s	N	r_s	N
#W	.432**	118	.492**	90	.281	28
#S	.436**	118	.489**	90	.513**	28
GI	.481**	118	.544**	90	.066	28
AG1k	.554**	118	.401**	90	.160	28
HyN	.080	118	-.136	90	.493**	28
HyV	.094	118	-.050	90	.165	28

** $p < .01$; * $p < .05$f

In the L1 corpus, neither #W, GI or AG1k were seen to have a relationship with scores, suggesting that these measures were predominately relevant for discriminating between high and low-scoring L2 essays, which is explored further below. #S did have the expected relationship to scores, suggesting that even among L1 essays, the amount of text produced influenced the perceived quality of writing. Since the #W did not play a role, it may be surmised that L1 writers who wrote fewer, longer sentences scored relatively lower than L1 writers who wrote a greater number of shorter sentences. In the L1 corpus, a significant positive relationship was also observed for noun hyponomy, indicating that texts with more general, abstract nouns were rated higher; no such relationship was found in the L2 corpus, however, suggesting that this particular aspect of lexical sophistication was more relevant for describing differences in essays written by native speakers than by L2 learners.

In order to better understand the extent to which fluency and lexical variables were related to perceived quality in L2 texts, we next considered which measures discriminated between high and low scoring L2 essays, using a series of Mann-Whitney's U-tests, the results of which are reported in Table 6.27. These revealed that high-scoring L2 essays had significantly more words, sentences, greater lexical diversity and lexical sophistication but, as expected based on analysis of correlations, showed no differences in either noun or verb hyponymy. Effect sizes were large for fluency and lexical diversity, medium for AG1k.[32]

[32] Following Field (2005) effect sizes for Mann-Whitney U-tests are calculated as r = ABS(Z/√N)

Table 6.27. Mann-Whitney tests: Fluency and lexical differences between high and low scoring L2 essays

	Mean Rank Low (n = 21)	Mean Rank High (n = 22)	*U*	*Z*	*Sig.*	*r*
#W	14.38	29.27	71	-3.888	**<.001**	0.59
#S	14.69	28.98	77.5	-3.744	**<.001**	0.57
GI	12.00	31.55	21	-5.102	**<.001**	0.78
AG1k	15.67	28.05	98	-3.231	**.001**	0.49
HyN	24.48	19.64	179	-1.264	.206	0.19
HyV	21.57	22.41	222	-.219	.827	0.03

6.4.1.2 Accuracy

The next domain explored was that of accuracy. Again, correlations were explored first, examining the full corpus and the L1 and L2 corpuses independently. It was predicted that higher scoring essays would be more accurate, and that significant positive correlations would be found for %EFS, while significant negative correlations would be found for all other measures. The results (Table 6.28) confirmed these predictions for the full corpus and the L2 corpus but revealed that accuracy measures did not have a significant relationship with scores in the L1 corpus. Given that there were very few errors of any kind in the L1 essays, it was not surprising that their effect was negligible.

Table 6.28. Spearman correlations: Total scores and Accuracy measures

	Full corpus		L2 Corpus		L1 Corpus	
Measure	r_s	N	r_s	N	r_s	N
%EFS	.688**	118	.447**	90	.240	28
%TotE	-.774**	118	-.624**	90	-.305	28
%Gre	-.668**	118	-.481**	90	-.208	28
%Lex	-.648**	118	-.448**	90	-.347	28
%Prag	-.456**	118	-.216*	90	.083	28
%Sp1	-.515**	118	-.389**	90	*n/a*	28
%Sp2	-.230*	118	-.252*	90	-.056	28

** $p < .01$; * $p < .05$f

The relationship between accuracy and perceived quality was explored further in the L2 corpus by looking at correlations between accuracy measures and the 5 component scores. These revealed that all measures of accuracy had the strongest relationship with Vocabulary scores, except for pragmatic errors, which were more strongly related to Content scores (Table 6.29).

Table 6.29. Accuracy correlations in the L2 corpus (n = 90)

Score	%TotE	%Gre	%Lex	%Prag	%Sp1	%Sp2
Con	-.526**	-.380**	-.340**	-.263*	-.272**	-.225*
Org	-.504**	-.370**	-.388**	-.212*	-.331**	-.198
Voc	-.653**	-.555**	-.474**	-.172	-.512**	-.321**
Lang	-.583**	-.490**	-.421**	-.110	-.373**	-.221*
Mech	-.476**	-.379**	-.347**	-.070	-.379**	-.228*
Tot	-.624**	-.481**	-.448**	.216*	-.389**	-.252*

** *p* < .01; * *p* < .05

When the accuracy of high and low-scoring L2 essays was compared with Mann-Whitney U-tests, the expected differences were found, with significant differences seen for all measures except for %Prag and medium to large effect sizes across the board, as reported in Table 6.30.

Table 6.30. Mann-Whitney tests: Accuracy differences between high and low scoring L2 essays

	Mean Rank					
	Low scorers	High scorers	*U*	*Z*	*Sig.*	*r*
%EFS	15.05	28.64	85	-3.554	**<.001**	0.54
%TotE	30.81	13.59	46	-4.495	**<.001**	0.69
%Gre	28.76	15.55	89	-3.450	**.001**	0.53
%Lex	28.14	16.14	102	-3.138	**.002**	0.48
%Prag	24.26	19.84	183.5	-1.164	.244	0.18
%Sp1	28.05	16.23	104	-3.185	**.001**	0.49
%Sp2	25.79	18.39	151.5	-2.716	**.007**	0.41

6.4.1.3 Syntactic complexity and variety

We then explored the relationships between qualitative writing scores and measures in the domains of syntactic complexity and variety. For syntactic complexity, it was predicted that the clause-level measures would have significant positive relationships with writing scores, based on the observations made in section 6.1.2; however no similarly strong predictions were made for sentence-level measures. Spearman correlations, reported in Table 6.31, revealed that the predictions for clause-level measures were confirmed for the full corpus, but that no such correlations were found in either the L1 or L2 corpuses independently. This suggested that clausal complexity primarily differentiated between L1 and L2 essays, which is explored further in the between-groups comparisons below. The same result was found for DC/S, which had a negative relationship to scores in the full corpus but no significant relationships in the L1 or L2 samples.

Table 6.31. Spearman correlations: Total scores and complexity measures

	Full corpus		L2 Corpus		L1 Corpus	
Measure	r_s	N	r_s	N	r_s	N
MLS	-.080	118	-.063	90	-.432*	28
DC/S	-.243**	118	-.070	90	-.058	28
MLC	.331**	118	.059	90	-.308	28
SYNNP	.241**	118	.023	90	-.303	28

The correlations between these measures were then re-explored when controlling for the number of words and sentences produced, in order to determine whether the significant relationships remained, as several of these measures were significantly related to overall fluency. Partial correlations, controlling first for #W and then for #S are reported in Table 6.32. When the number of words was controlled, all significant relationships remained significant, indicating that complexity had a relationship with writing quality independently of text length. Additionally, a new relationship was found between SYNNP and scores in the L2 corpus, indicating that in L2 essays of similar length, this particular measure of clausal complexity played a role in raters' judgments. Similarly, for sentence-level measures the previously documented significant relationships remained strong. When the number of sentences was controlled, the relationship for clause-level measures remained significant; however significant correlations between scores and all sentence-level measures disappeared. This indicated that, as suspected, the relationship between perceived quality and sentence-level complexity was more likely a side-effect of the relationship between these measures and the number of sentences produced. That is, the number of dependent clauses per sentence was negatively correlated with the number of sentences produced, which was in turn correlated with qualitative scores, but DC/S was not found to have any independent relationship to essay quality.

Table 6.32. Partial correlations between qualitative scores and complexity measures when controlling for text length

		Full corpus		L2 Corpus		L1 Corpus	
	Measure	*r*	*df*	*r*	*df*	*r*	*df*
Controlling #W	MLC	.366**	115	.164	87	-.380	25
	SYNNP	.348**	115	.212*	87	.095	25
	MLS	-.154	115	-.186	87	-.400*	25
	DC/S	-.321**	115	-.214*	87	-.121	25
Controlling #Sentences	MLC	.399**	115	.197	87	-.347	25
	SYNNP	.321**	115	.157	87	-.341	25
	MLS	.147	115	.198	87	-.211	25
	DC/S	-.114	115	.076	87	.040	25

** $p < .01$; * $p < .05$

For syntactic variety, the prediction was that both measures would be negatively correlated with writing scores, because higher scoring essays would use a wider variety of syntactic structures, resulting in lower levels of overlap and repetition. These predictions were not confirmed in analysis of bivariate correlations (although the negative correlation for TEMP was near significance, r_s = -.178, p = .054); however both measures were significantly correlated with #S, and when this measure was controlled, significant negative relationships were indeed documented, both for the full corpus and for the L2 corpus, as reported in Table 6.33. This indicated that, among essays with similar numbers of sentences, a greater variety of syntactic structures (less repetition of tense and aspect, less structural similarity between adjacent sentences) was associated with higher quality writing. No significant relationships were found in the L1 corpus, where the degree of syntactic variety was relatively more homogenous.

Table 6.33. Partial correlations: Total scores and variety measures

		Full corpus		L2 Corpus		L1 Corpus	
	Measure	*r*	*df*	*r*	*df*	*r*	*df*
Controlling #S	StrutA	-.254 **	115	-.233*	87	.300	25
	Temp	-.192*	115	-.164	87	.092	25

** $p < .01$; * $p < .05$

When the high and low scoring L2 essays were compared on complexity and variety measures using Mann-Whitney tests, the results of the correlations were confirmed and no significant differences were found between groups, as reported in Table 6.34. Based on the partial correlations above, the assumption is that any differences were obscured by the effect of text length. As expected following correlation analysis, no significant between-groups differences were found for syntactic variety, and the assumption is that these differences were being obscured by the difference in the number of sentences produced by the two groups.

Table 6.34. Mann-Whitney test: Complexity and variety differences, high and low L2 groups

	Mean Rank Low (n = 21)	Mean Rank High (n = 22)	*U*	*Z*	*Sig.*
MLC	21.43	22.55	219	-.292	.771
SYNNP	21.64	22.34	223.5	-.182	.855
MLS	22.71	21.32	216	-.364	.715
DC/S	22.88	21.16	212.5	-.450	.653
StrutA	24.36	19.75	181.5	-1.203	.229
Temp	24.76	19.36	173	-1.412	.158

6.4.1.4 Cohesion and pronoun density

The final measures explored were those in the domain of cohesion, along with our measure of pronoun density (DenPr). Again, we began by exploring correlations between these measures and qualitative scores. For cohesion, the prediction was that all measures would be negatively correlated with qualitative scores, both because of the nature of the content (personal opinions and common knowledge) and because of the characteristics of our raters (expert readers with high prior knowledge). That is, we expected that an overuse of cohesive devices would be perceived as redundant and that texts with fewer connectives and less referential overlap would be perceived as higher quality. These predictions were confirmed for all measures, as seen in Table 6.35, although only %Con and %CauseCon were seen to have significant relationships in both the L1 and L2 corpuses as well. In the L2 corpus, an additional negative relationship was found between RefC (content word overlap) and qualitative scores that would not found in the L1 corpus, suggesting that the correlation in the full corpus was predominantly due to differences in the L2 texts. Finally, in the L1 corpus the negative correlation between qualitative scores and connectives appeared to be entirely due to causal connectives, while in the L2 corpus it was due to an overuse of the full range of connectives (additive, causal, temporal and logical). For pronoun density (DenPr), the prediction was met for the corpus as a whole, in that qualitative scores were negatively correlated with high numbers of pronouns; however given that these relationships were not found in either the L2 or L1 corpuses when examined separately, it appears that this measure was primarily discriminating between L1 and L2 texts.

Table 6.35 Correlations between qualitative scores and cohesion measures

	Full corpus		L2 Corpus		L1 Corpus	
Measure	r_s	N	r_s	N	r_s	N
RefP	-.293**	118	-.174	90	-.203	28
RefA	-.225*	118	-.194	90	.012	28
RefC	-.251**	118	-.266*	90	-.029	28
%Con	-.466**	118	-.541**	90	-.486**	28
%AdCon	-.208*	118	-.328**	90	-.328	28
%CausCon	-.468**	118	-.421**	90	-.532**	28
%LogCon	-.547**	118	-.515**	90	-.320	28
%TempCon	-.199*	118	-.243*	90	-.152	28
DenPr	-.377**	118	-.181	90	.048	28

** $p < .01$; * $p < .05$

As with syntactic complexity and variety measures, these correlations were re-explored through partial correlations controlling for text length. For measures of referential overlap, the number of sentences was controlled while for pronoun and connective density the number of words was controlled. As seen in Table 6.36, all significant relationships between referential overlap and qualitative scores remained significant when text length was held constant, and an additional relationship arose between pronoun density and qualitative scores in the L2 corpus. While the overall relationship between connectives and qualitative scores remained significant when the number of words was controlled, the relationship between additive and temporal connectives disappeared for the corpus as a whole, suggesting that these two measures were closely related to text length. Similarly, in the L2 corpus, the significant relationships between additive, causal, and temporal connectives disappeared when text length was controlled, and the significant relationship between connectives and qualitative scores was seen to primarily result from differences in the use of logical connectives. In the L1 corpus, the strong negative correlation between causal connectives and qualitative scores remained strong when text length was controlled.

Table 6.36. Partial correlations: Total scores and cohesion measures

		Full corpus		L2 Corpus		L1 Corpus	
	Measure	*r*	*df*	*r*	*df*	*r*	*df*
Controlling #S	RefP	-.237*	115	-.120	87	-.339	25
	RefA	-.185*	115	-.111	87	-.162	25
	RefC	-.197*	115	-.220*	87	-.089	25
Controlling #W	%Con	-.256**	115	-.280**	87	-.447*	25
	%AdCon	.053	115	-.080	87	-.264	25
	%CausCon	-.331**	115	-.187	87	-.492**	25
	%LogCon	-.401**	115	-.269*	87	-.375	25
	%TempCon	-.160	115	-.176	87	-.075	25
	DenPr	-.423**	115	-.218*	87	-.005	25

** $p < .01$; * $p < .05$

When measures in this domain were compared in high and low-scoring L2 essays, high scoring learners were found to use significantly less referential overlap than low-scoring learners, and significantly fewer connectives of all types except for temporal connectives, while no difference was found for pronoun density.

Table 6.37. Mann-Whitney test: Variety and cohesion differences, high and low L2 groups

	Mean Rank Low (n = 21)	Mean Rank High (n = 22)	*U*	*Z*	*Sig.*	*r*
RefP	24.83	19.30	171.5	-1.449	.147	--
RefA	26.00	18.18	147	-2.047	**.041**	**.31**
RefC	26.29	17.91	141	-2.188	**.029**	**.33**
%Con	30.21	14.16	58.5	-4.191	**<.001**	**.64**
%AdCon	26.38	17.82	139	-2.235	**.025**	**.34**
%CausCon	29.14	15.18	81	-3.644	**<.001**	**.56**
%LogCon	29.90	14.45	65	-4.033	**<.001**	**.62**
%TempCon	25.10	19.05	166	-1.584	.113	--
DenPr	23.10	20.95	208	-.559	.576	--

6.4.2 Relationship between FLACC and LGP

Finally, we considered the relationship between FLACC measures and lexoci-grammatical proficiency. We first considered this via correlations between grammar and cloze scores and FLACC measures in the full L2 corpus, in response to *RQ4b. Which FLACC measures are significantly correlated with grammar and cloze scores?* We looked at baseline correlations and also looked at correlations when essay scores (our proxy for writing ability) were controlled. Next we considered which measures discriminated between learners and native speakers with similar qualitative scores—two groups with presumably similar levels of writing ability but inherently different levels of lexico-grammatical competences—in response to *RQ4c. Which FLACC measures discriminate between learners and native speakers with similar qualitative scores?* These two subquestions are addressed in subsections 6.4.2.1 and 6.4.2.2, below.

6.4.2.1 Correlations in the L2 corpus

Correlations between FLACC measures and grammar and cloze scores were analyzed using Kendall's Tau coefficient, due to the many tied ranks in the latter, as explained above in section 6.2. We considered the results for all domains together, in Table 6.38.

Table 6.38. Correlations between grammar and cloze scores and quantitative measures (N = 89)

Fluency & Lexical		Accuracy		Complexity & variety		Cohesion	
	τ		τ		τ		τ
#W	.087	%EFS	.333**	MLS	-.042	RefP	-.155*
#S	.072	%TotE	-.512**	DC/S	-.097	RefA	-.154*
GI	.355**	%Gre	-.422**	MLC	.184*	RefC	-.297**
AG1k	.212**	%Lex	-.283**	SYNNP	.124	%Con	-.174*
HyN	-.032	%Prag	-.166*	StrutA	-.047	%AdCon	-.064
HyV	-.078	%Sp1	-.257**	Temp	-.200**	%CausCon	-.101
		%Sp2	-.144			%LogCon	-.122
						%TempCon	-.188*
						DenPr	-.190**

Analysis revealed that LGP was not correlated with overall fluency, but was positive correlated with lexical diversity and sophistication, as measured by AG1k. In the domain of accuracy, high scores on the grammar and cloze tests were correlated with all measures except for Type 2 spelling errors, further confirmed assumptions about this measure. In the domain of syntactic complexity, the only relationship found was a positive relationship between clause length and LGP, with higher scoring learners producing longer clauses. LGP was negatively correlated with tense and aspect repetition, and with all 5 cohesion measures indicating that learners with higher scores on the grammar and cloze tests used more varied sentence structures, a greater variety of tense and aspect, and wrote less repetitive sentences, with fewer pronouns and connectives per word.

We also considered the correlations between LGP and FLACC measures when writing ability was held constant, conducting partial correlations using qualitative scores (Total PROFILE scores) as a control variable. These results are reported in Table 6.39.

Table 6.39. Partial correlations between grammar and cloze scores and quantitative measures, controlling for essay scores (N = 86)

Fluency & Lexical		Accuracy		Complexity & variety		Cohesion	
	τ		τ		τ		τ
#W	-.271**	%EFS	.312**	MLS	-.080	RefP	-.162
#S	-.234**	%TotE	-.528**	DC/S	-.181	RefA	-.115
GI	.239**	%Gre	-.485**	MLC	.268*	RefC	-.315**
AG1k	.111	%Lex	-.124	SYNNP	.192	%Con	.112
HyN	.006	%Prag	-.127	StrutA	.031	%AdCon	.146
HyV	-.128	%Sp1	-.198	Temp	-.275**	%CausCon	.070
		%Sp2	-.002			%LogCon	.152
						%TempCon	-.058
						DenPr	-.198

In the domain of fluency, a significant negative correlation appeared for both measures (#W and #S), which suggested that for essays with similar scores, greater fluency was associated with lower levels of LGP. In the domain of lexical diversity and sophistication, the positive correlation seen between GI and LGP remained strong; however the correlation between AG1k disappeared and no new relationships were observed in the domain of lexical sophistication. In the domain of accuracy, the correlation between LGP and the umbrella measures (%EFS, %TotE) remained strong, as did the correlation for %Gre; however the correlations between LGP and the other error subtypes (%Lex, %Prag, %Sp1) disappeared, suggesting that the correlations between these measures and LGP may have resulted from the overlap between LGP and writing competence. The same was true for two measures of referential overlap, both measures of cohesion, and pronoun density. The only measure in the domain of cohesion that appeared to have an independent relationship with LGP was RefC, or referential overlap of content words between sentences. The negative correlation may have reflected the fact that less overlap of content words was related to more lexical diversity (RefC and GI were correlated (at $\tau = -471, p < .001$).

6.4.2.2 Differences between learners and NS with similar qualitative scores

To answer RQ4c, we compared the 22 high-scoring L2 essays identified above in section 6.4.1 to the 23 L1 essays that fell within a highly comparable range (86-95). The 5 highest scoring L1 essays were excluded from analysis so that scores between the groups would be more comparable. Although the native speakers scored higher on average, differences were minimal in comparison to differences seen in the full corpus based on language background (above in section 6.3). The distribution of scores in these two groups is reported in Table 6.40.

Table 6.40. L1 essays vs. high-scoring L2 essays

Group	N	Mean	Median	Min	Max
L2	22	89.9	90	85.5	95
L1	23	91.1	91.5	86	95

When L1 and L2 essays with comparable scores were compared on fluency and lexical measures, no significant differences were found behind groups, although the difference in lexical sophistication as measured by AG1k neared significance (p = .054) and represented the greatest difference between the two groups (see Table 6.41).

Table 6.41. Mann-Whitney test: Fluency and lexical differences between high scoring L1 and L2 essays

	Mean Rank L2 (n = 22)	Mean Rank L1 (n = 23)	*U*	*Z*	*Sig.*
#W	25.77	20.35	192	-1.385	.166
#S	25.80	20.33	191.5	-1.403	.161
GI	24.41	21.65	222	-.704	.482
AG1k	19.14	26.70	168	-1.930	.054
HyN	20.27	25.61	193	-1.362	.173
HyV	20.41	25.48	196	-1.294	.196

In the domain of accuracy, we found that despite obtaining similar scores, there were dramatic differences in accuracy found between the L1 and L2 essays, with significant differences and large effect sizes found for all measures except for Type 2 spelling errors, as reported in Table 6.42. The lack of between-groups differences for this measure further confirmed the suspicion that Type 2 spelling errors more likely resulted from the pressure of writing under timed conditions than from a lack of linguistic knowledge.

Table 6.42 Mann-Whitney test: Accuracy differences between high scoring L1 and L2 essays

	Mean Rank L2 (n = 22)	Mean Rank L1 (n = 23)	*U*	*Z*	*Sig.*	*r*
%EFS	12.91	32.65	31	-5.044	**<.001**	**0.75**
%TotE	34.14	12.35	8	-5.563	**<.001**	**0.83**
%Gre	33.09	13.35	31	-5.044	**<.001**	**0.75**
%Lex	30.55	15.78	87	-3.919	**<.001**	**0.58**
%Prag	29.91	16.39	101	-3.672	**<.001**	**0.55**
%Sp1	28.23	18.00	138	-3.588	**<.001**	**0.53**
%Sp2	21.52	24.41	220.5	-1.352	.176	0.20

When L1 and high-scoring L2 essays were compared in the domains of syntactic complexity and variety, the expected differences were found between the two groups for both clausal complexity measures, with native speakers producing significantly longer clauses, with more coordinate phrases and more noun phrase modification, than their L2 peers. For sentence-level complexity measures, significant differences were found for C/S and DC/S, with L2 writers producing more of each. No differences were found in the domain of syntactic variety.

Table 6.43. Mann-Whitney test: Complexity and variety differences between high scoring L1 and L2 essays

	Mean Rank L2 (n = 22)	Mean Rank L1 (n = 23)	*U*	*Z*	*Sig.*	*r*
MLC	14.77	30.87	72	-4.110	**<.001**	**.61**
SYNNP	16.57	29.15	111.5	-3.213	**.001**	**.48**
MLS	22.36	23.61	239	-.318	.751	--
DC/S	28.00	18.22	143	2.503	**.012**	**.37**
StrutA	24.27	21.78	225	-.636	.525	--
Temp	26.43	19.72	177.5	-1.715	.086	--

Finally, we considered differences in cohesion and pronoun density. The only significant differences between L1 essays and L2 essays that received similar qualitative scores was for the measure of pronoun density, as seen in Table 6.44. It was revealed that the L2 essays used significantly more pronouns per word than their L1 counterparts (p <.001, r = .59).

Table 6.44. Mann-Whitney test: Cohesion differences between high scoring L1 and L2 essays

	Mean Rank L2 (n = 22)	Mean Rank L1 (n = 23)	*U*	*Z*	*Sig.*
RefP	26.05	20.09	186	-1.522	.128
RefA	23.16	22.85	249.5	-.080	.937
RefC	21.98	23.98	230.5	-.511	.609
%Con	20.11	25.76	189	-1.442	.149
%AdCon	19.18	26.65	169	-1.907	.056
%CausCon	23.05	22.96	252	-.023	.982
%LogCon	24.64	21.43	217	-.817	.414
%TempCon	20.45	25.43	197	-.1272	.203
DenPr	30.91	15.93	79	-3.951	**<.001**

Overall, it appeared that the high scoring learners differed primarily in terms of accuracy and syntactic complexity, and not in terms of cohesion, fluency, lexical diversity or sophistication.

6.5 Summary: Results

In this chapter we presented the results of our analysis of argumentative essays gathered from 28 L1 and 30 L2 participants, looking at both the perceived quality of these essays and their quantitative characteristics. We considered the longitudinal progress of the L2 learners—in relation to learning context (AH vs. SA) and 'initial level'—differences between L1 and L2 essays, and relationships between perceived quality, lexico-grammatical proficiency, and quantitative measures. In the following chapter, we will provide a discussion and interpretation of the results obtained in relation to our four research questions.

Chapter 7

Discussion

In the previous chapter we presented the results gathered in relation to our four research questions, analyzing writing collected from 30 EFL learners before and after AH and SA learning contexts, and from 28 native speakers of English. In this chapter, we elaborate further on the results obtained and consider them in relation to previous research, returning to many of the theoretical arguments and empirical studies reviewed in Part I of this dissertation. We also supplement the statistical analysis presented in the previous chapter with some qualitative analysis, looking at the performance of one individual in detail, and looking at extracts from a handful of essays in order to improve our understanding of the measures used and the progress made. This chapter is organized around the four main research questions outlined in Chapter 4 and discusses the results obtained in the order that they were presented in the previous chapter.

7.1. Longitudinal changes after AH and SA contexts

Our first research question asked whether or not the L2 participants improved their writing over time and after the AH and SA learning contexts, respectively:

RQ1. *Does learners' writing improve over time, and after AH and SA, in terms of perceived quality and in terms of FLACC? Is one context relatively more beneficial than the other?*

After the analysis reported in the previous chapter, we can offer affirmative answers to both parts of this question: the 30 participants' writing did improve over time and the SA context was found to be relatively more beneficial that the AH context. We will review the main findings and consider how they compare with previous SA research in our discussion of the two sub-questions (RQ1a and RQ1b).

7.1.1. Changes in writing quality

Sub-question RQ1a (*Do qualitative writing scores improve significantly after either the AH or SA learning contexts?*) addressed improvement as a function of the perceived quality of learners' writing. As described in Chapter 5, writing quality was measured using the popular and well-validated analytic scale developed by Jacobs et al. (1981). We had two raters use the scale to evaluate each essay in our corpus in terms of its communicative effect, providing scores for 5 weighted components (Content, Organization, Vocabulary, Language Use, and Mechanics), which were then summed and averaged across raters; Total scores were used as the primary dependent variable, though the component scores were also considered to enrich our interpretations of changes in Total scores. We will first summarize the results found the then discuss their implications.

Before measuring longitudinal changes, we examined participants' qualitative scores in relation to the normative data published by the authors of the PROFILE. The large majority of essays in the 90-essay corpus fell into one of the three "Advanced" categories, in line with our initial assessment of participants' proficiency levels; however a small portion of essays scored in the High Intermediate range (n = 12), and this included essays written at T1 (n = 6), T2 (n = 5), and T3 (n = 1). The group of lower-scoring learners did not lead us to question our classification of the learners as having advanced proficiency, since research indicates formal writing often lags behind other aspects of proficiency. That is, participants with high-intermediate writing skills may well be more advanced in other areas, particularly in passive skills like reading and listening comprehension. Indeed, the authors of the PROFILE report this discrepancy between skills in their description of the validation process:

> "In our testing program we have observed a substantial correlation among and between all language skills in our ESL students, but we have also noticed that many students who demonstrate strong proficiency in listening and reading, and occasionally speaking, have often *not* developed an equivalent level of proficiency in writing English." (Jacobs et al., 1981, p.58, emphasis theirs)

After establishing participants' levels of proficiency and looking at descriptive statistics, we explored longitudinal changes and observed that mean total scores increased steadily from T1 to T3. Statistical analysis revealed that this improvement was significant and that, as a group, the learners received significantly higher writing scores at the end of the study than at the beginning. Pairwise comparisons revealed that the improvement observed after AH did not reach statistical significance while the improvement observed after the SA context did. When we explored the effect sizes for each of the 5 component scores, to improve our understanding of why writing improved after the SA context, we found that the component with the largest effect size was Language Use, which evaluates the frequency and accuracy of complex constructions and the presence of grammatical errors, particularly those that obscure meaning (see Chapter 3, Figure 3.4); in contrast, the Vocabulary component had the smallest effect size, indicating that although Vocabulary scores did improve significantly, they remained more stable over time than those of the other components.

We also observed a bimodal distribution in three of the 5 component scores (Content, Organization, and Vocabulary) at T3, which led us to explore how the pattern of longitudinal improvement varied for different participants. We found that, as expected, the majority of participants improved their scores over the course of the study, with 26 participants receiving higher scores at T3 than at T1 and only 4 participants showing a lack of improvement over time; however when we explored the gains that occurred after AH and SA respectively, we found there was a negative correlation between improvement in the two contexts, suggesting that participants who improved in the AH context were less likely to improve in the SA context, and vice versa. When we looked at the number of individuals who improved in each context, we saw that a substantial number improved in the AH context as well but that the SA context was relatively more beneficial to the group: 21 participants improved after the SA context while only 18 improved after the AH context; 11 of the 30 participants improved *only* during the SA context, in comparison to 8 participants who improved only in the AH context, and 10 participants who improved in both.

Overall, despite the variation among individual participants, our analysis provided us with a clear response to RQ1a. *Do qualitative writing scores improve significantly after either the AH or SA learning contexts?* We found that qualitative writing scores improved significantly after the SA context but did not improve significantly after the AH period of classroom study. At first glance, the findings with regard to the AH context are discouraging; however given that the learners wrote only 1-2 essays per term and were given very little writing instruction or feedback during these courses (which, as discussed in Chapter 5, were primarily focused on grammar and linguistic analysis), it is unsurprising that they did not make much progress during this time. Furthermore, our results with regard to the AH context are consistent with the findings of Sasaki (2004, 2007, 2009, 2011). Sasaki (2004) is the only study of the four to report substantial improvement during EFL classes in the AH context. In that study, all 11 participants made process over the course of their freshman year at home in Japan, and the group mean increased from 66.8 on the Jacobs scale, in the "High intermediate" range, to 75.5, in the "Low Advanced" range[33]; however she highlights that they all took a year-long course focused on process-writing and on improving their meta-cognitive knowledge of the writing process. As Sasaki describes:

> "The instructor (the researcher) taught the participants process-writing strategies such as planning and revising, based on Bereiter and Scardamalia's (1987) ideas of ''promoting the development of mature composing strategies'' (p. 245), using Hashiuchi's (1995) Paraguraphu Raitingu Nyuumon [Introduction to Paragraph Writing], a composition textbook with special emphasis on process writing. In the first class, the students were told that writing is an interactive process between what they write and what they want to write, and that such a process is cyclical, starting with planning and followed by writing and revising. Furthermore, in each chapter of the textbook (the class covered a total of nine chapters), the students first learned rhetorical patterns such as comparison, classification, and expressing opinions, and then were instructed to write a similar paragraph themselves. The 11 participants in the present study received this instruction in a class of 25 students once a week for 90 min." (p. 535-536)

In Sasaki's (2007) study, on the other hand, she found that the AH participants who attended regular EFL classes but did not have any focused writing instruction made little improvement over a full academic year. Similarly, in Sasaki (2011), the 9 participants who remained at home reportedly did not see significant, sustained improvement over the course of the study, despite the fact that this period included 3.5 years of EFL classes: 8.8 hours per week in the 1st year, 6.2 hours/week in second year, and 6.1 hours/week in the third year. Considering that our own participants had only 2.5 hours/week of English class and that the AH period lasted for only 6 months, it becomes less surprising that their writing scores did not significantly improve during this context.

[33] Calculated based on values reported in Sasaki (2004, p. 545)

With regard to the SA context, as we discussed in Chapter 1, relatively little SA research has focused on writing and even less research has considered improvement in the perceived quality of that writing. Freed, So, and Lazar (2003) considered qualitative improvement in the writing of 15 North-American students who spent a semester abroad in France and found that native-speaking judges did not perceive the students' writing as being more "fluent" after SA; however their results are not easily comparable to our own due to the methods used and the heterogeneous proficiency levels of participants (who had studied French "from a few intensive months to nine years" p .4). Sasaki's (2004, 2007, 2009, 2011) four studies in the Japanese context are much more easily comparable to our study, given that she used the same analytic scale (Jacobs et al., 1981) and given that writing ability was assessed by means of a timed, argumentative essay. Her studies pointed to similar results as those obtained here: that SA experiences have a positive impact on L2 writing skills. Her later studies suggested that SA experiences lasting less than 4 months did not have the same lasting impact; however the group with shorter stays spent only 1.5-2 months abroad. Our results indicate that 3-months is long enough to lead to improvement, perhaps because 3 months (at least in the fall term) allows participants to complete an entire semester at the host university and thus to be more integrated in the community (1.5-2 months would not be enough time to enroll in regular semester-length courses).

One interesting difference between our study and Sasaki's, which puts our results in a different light, is that most of the participants in Sasaki's studies were at North American universities where they were enrolled in composition courses or ESL writing courses during their SA, whereas our participants were mostly in British and Irish universities and did not take composition courses or have process-writing instruction, though they may have done some writing in language courses, similar to those taken in the AH context. Sasaki's first two studies suggested that the improvement of the SA groups may have been due to their coursework while abroad, as opposed to immersion in the target language community more generally. For example, in Sasaki (2004), where all 11 participants received the intensive process-writing instruction described above, both groups improved and there did not seem to be any comparative advantage of the SA learning context. In Sasaki (2007), on the other hand, she compared 7 SA participants who *did* have writing instruction with 6 AH participants who did *not* receive intensive writing instruction. As she describes: "All study-abroad students completed at least one ESL writing class or a regular writing class (for English-speaking students) during their stay abroad. In those writing classes, all students learned how to prepare texts for what Johns (1997) called "the pedagogical genres"" (Sasaki, 2007, p.

607). In contrast, she reports that the AH participants took part in regular EFL classes (3.3 EFL classes per week) but makes no mention of any writing-specific instruction. In Sasaki (2007), the SA group improved considerably over the year spent abroad (from 75.15 to 82.43) while the AH group's scores actually decreased slightly (from 72.25 to 70.75). These results, alongside those from Sasaki (2004) suggested that the improvement seen in the SA context may have resulted from the instruction and not from the SA experience per se; however our own study has shown that even when there is no intensive writing instruction during the SA, simply being immersed in the L2 and attending degree-related courses in the target language may have a significant positive effect of the perceived quality of participants' writing, most likely due to improvement in general linguistic competence (as manifested in dramatic improvement in the Language Use component).

7.1.2. Changes in quantitative measures

After answering RQ1a, we explored writing improvement in more depth by considering changes in the quantitative characteristics of learners' essays. We selected measures in the domains of fluency, lexical diversity and sophistication, accuracy, complexity and cohesion (FLACC), all of which had theoretical relationships to writing quality and/or language proficiency based on previous research. We considered longitudinal changes for each domain in turn and also compared the two learning contexts, in response to subquestion RQ1b. *Are there significant changes in FLACC measures after either the AH or SA learning contexts?*

In the domain of fluency, we found that participants' essays increased in fluency over time, and that the group mean for both fluency variables (#W and #S) was substantially higher at T3 than at T1; however the means appeared to have a U-shaped pattern, decreasing after the AH context and increasing again after the SA context. When we analyzed these changes statistically we found that the observed decrease in fluency after the AH context was not statistically significant but that the improvement in fluency after the SA context was. The observation that #W increased was in line with the data reported by Perez-Vidal and Juan-Garau (2009) for a different sample of the SALA corpus and with the results reported by Sasaki (2004, 2007), who reported that SA participants' essays increased in length after their time abroad. That is, after spending time abroad participants are able to produce more content in the same amount of time, suggesting either that they have a greater linguistic repertoire on which to draw or that they are able to access their linguistic repertoire more efficiently (Wolfe-Quintero, et al., 1998). While our analysis of RQ4 suggested that the increase in sentences may have been closely related to changes in syntactic complexity, which we will discuss further below, both changes are clear evidence of progress after the SA context, and a lack of progress after the AH context.

The results in the domain of lexical diversity followed a similar pattern to those observed in the domain of fluency, which is predictable given that there tend to be strong correlations between these measures (Malvern & Richards, 1997). The mean scores for lexical diversity (as measured by Guiraud's Index) decreased after the AH context and then increased after the SA context, and both changes were found to be statistically significant, suggesting that as a group, the learners were using less variety of word types in their T2 essays and a greater variety of word types at T3. There were no significant changes in lexical sophistication, as measured by either the use of rare words (AG1k) or by noun or verb hyponymy, despite the fact that they clearly had not converged with native speakers on these measures (except potentially for verb hyponymy, which fluctuated very little between groups). The gains in both fluency and lexical diversity after the SA period are in line with predictions based on previous research and with the findings reported by Perez-Vidal & Juan-Garau (2009) for a different sample of the SALA corpus; however the lack of progress in lexical sophistication is somewhat surprising since lexical sophistication measures based on the relative frequency of words, like AG1k or the lexical frequency profile, are theoretically held to be measures of a writer's productive vocabulary size (Laufer & Nation, 1995). Since previous research suggested that *receptive* vocabulary is likely to grow substantially during SA learning contexts (Ife, et al., 2001; Milton & Meara 1995), we expected that this might have an impact on productive vocabulary and that there would be a substantial increase in lexical sophistication after SA. The fact that there was no such change may be due to the uncertain link between receptive and productive vocabulary (Meara, 2010) or to the demanding nature of argumentative writing, which requires attention to many other factors. Murphy and Roca de Larios (2010) report that searching for lexical items that adequately convey their meaning is one of the most difficult problems L2 writers face, and it might be particularly difficult to retrieve newly acquired lexis when writing under time pressure and restricted by the prompt.

In the domain of accuracy, we found that the proportion of errors (%TotE) in the learners' essays decreased over time and the proportion of error-free sentences (%EFS) increased over time but that learners were still far below native speakers in terms of accuracy and that the observed changes were not statistically significant for either of these two global measures. The only domain of accuracy that did show significant improvement over time was spelling (%Sp1): the number of spelling errors per word decreased slightly but not significantly after the AH context, and then decreased significantly after the SA context. Although spelling errors are frequently eliminated from consideration in studies of L2 writing (see Polio, 1997), this decision is rarely justified and appears to be more a matter of convention. As Mollet et al. (2010) argued, "misspellings may be precisely what separates out one writer from another" (p. 434); in our own study, taking spelling errors into account has provided us with evidence that the learners improve in at least one aspect of accuracy after the SA, and correcting them would have deprived us of important information. The lack of significant improvement in overall accuracy as measured by %TotE and %EFS, frequency-based measures, may be contrasted with the significant improvement in participants' Language Use scores as reported in the results for RQ1a, which raises the question of whether frequency measures miss an important component of accuracy, as argued by Engber (1995) and others. That is, the Language Use component on the PROFILE (see Chapter 3, Figure 3.4) asks raters to reflect on the number of errors in an essay only to the extent that "*meaning is confused or obscured*". The fact that Language Use scores improved significantly after the SA context (and indeed, showed the largest effect size of all components) suggests that although participants' essays had a similar *number* of errors before and after the SA context, there was likely a decrease in the number of "serious" errors, or those that impeded the writers' ability to communicate. In this study we did not consider the relative gravity of errors in our coding scheme but reanalyzing our accuracy data using a system for ranking severity (such as that used in Kuiken & Vedder, 2008) would presumably reveal that the SA had a significant positive impact.

In the domain of syntactic complexity, we observed no significant changes after either context for the sentence-level measures, which included a global measure of sentence complexity (MLS) and a measure of the use of subordination (DC/S). For MLS, the lack of changes were unsurprising given that the learners' average scores were highly similar to those of native speakers at the beginning of the study, and did not vary by more than 1 word at any data collection time. Subordination appeared to decrease steadily over time, in the direction of native speaker norms, but variation was very minimal (from 1.5 to 1.4 to 1.3, over the course of the study). In general, these high values showed that our learners were already making extensive use of subordination, and using a greater proportion of complex sentences than simple sentences; given their advanced proficiency, we would expect that any improvement in complexity would be seen in a slight decrease in the use of subordination in favor of increased phrasal elaboration or nominalization following the developmental theories of Norris and Ortega (2009) and Halliday and

Matthiessen (1999). When we examined clausal complexity, this prediction was confirmed, as we saw a significant increase in clause length (MLC). Our measure of noun phrase modification (SYNNP) did appear to increase over time, in the direction of native speaker norms; however these changes were not significant and the improvement observed (from .71 to .76) still left them well below the level of native speakers (.89), who appeared to use much more noun phrase modification. MLC decreased slightly after the AH context, though this change was not significant, and then increased significantly after the SA context. Again, the improvement seen from T1 (9.2 words per clause) to T3 (9.6 words per clause) was still well below the values recorded for native speakers (11.6 words per clause); however it was evidence that the SA context had a significant positive effect on the complexity of our participants' writing. In the two previous studies to consider changes in complexity after SA contexts, both Freed, So, and Lazar (2003) and Perez-Vidal and Juan-Garau (2009) reported that SA experiences had no effect on syntactic complexity; however they only examined indices of subordination and coordination at the sentence level (Freed et al. used the number of words per T-unit, while Perez-Vidal and Juan-Garau used the Coordination Index proposed by Bardovi-Harlig (1992)). Our own study has suggested that sentence-level measures are not sufficient to measure progress (for example, our learners were already performing quite similarly to native speakers in terms of MLS and clearly making greater use of subordination than the native speakers) and confirm Norris and Ortega's argument that measures of phrasal elaboration and nominalization must be used in order to accurately measure development in complexity.

In addition to syntactic complexity measures, we looked at two Coh-Metrix indices associated with syntactic variety: structural overlap between adjacent sentences (StrutA) and tense and aspect repetition between adjacent sentences (Temp). We assumed that such measures, while not related to proficiency or included in the CAF model, might be evidence of development in composing competence, since syntactic variety is theoretically associated with higher quality writing, as seen on the rubrics reproduced in Chapter 3. There were no significant changes in StrutA; however Temp decreased significantly over time, in the direction of native speaker norms, which suggested improvement in the domain of syntactic variety. Examination of changes between learning contexts revealed that the significant improvement occurred after the AH context and not after the SA context; however, since the learners appeared to converge with the native speakers after the AH period, it was predictable that no further changes occurred (that is, this measure had a clear ceiling effect). It is coherent with our understanding of writing development that Temp, a feature associated with composing competence and not necessarily with linguistic proficiency, was observed to change after the period of formal instruction. Furthermore, the reduced repetition of verb tense and aspect between sentences may have reflected a tendency to use a greater variety of tense and aspect in the text as a whole, which would be coherent given that their EFL classes were grammar-focused and aimed to improve their command of syntax, as discussed in Chapter 5. Indeed, 2 of the 9 units on the syllabus for the first-year English course are specifically related to tense and aspect. The decrease in structural overlap from T2 (.083) to T3 (.078) in the direction of native speaker norms (.071) suggested improvement but showed us no significant changes and thus cannot be interpreted. In general, very few writing studies have examined syntactic variety in a systematic way, and we lack precise and descriptive quantitative measures of variety; designing a method of quantifying syntactic variety, such as Huw Bell's (2008) system for quantifying syntactic complexity by assigning point values to different layers of embedding, would be a potentially interesting and useful area for future research.

Finally we considered changes in the domain of cohesion, looking at changes in the amount of referential overlap (RefA, RefP, RefC) and in the use of connectives (%Con, and subtyles); we also considered the use of personal pronouns (DenPr). We expected that improvement would lead to decreases in explicit cohesion, given the subject matter and expertise of our readers, based on the arguments of McNamara (2001) and McNamara and Kintsch (1996), as discussed in Chapter 3. We did not observe any significant changes in referential overlap, suggesting that cohesion across sentences remained stable throughout the study; for the first two measures (RefP and RefA) differences between the learners and native speakers suggested that their may have been room for improvement, while RefC appeared to have a ceiling effect, with learners performing similarly to native speakers from the beginning of the study. No changes were found for the measure of pronoun density, despite the qualitative improvement in learners' scores and the fact that the overuse of personal pronouns has been found to differentiate between learners and native speakers (see Shaw & Liu, 1998). There did appear to be an increase after the AH period and then a decrease after the SA period, and learners at T3 produced fewer pronouns than learners at T1; the fact that this change did not reach significance (*p* = .084) may have been because in this study we did not differentiate between first and second person pronouns or between singular and plural forms. In Shaw and Liu's analysis of development over a 2-3 month EAP (English for Academic Purposes) course focused on writing, they examined first person pronouns and found that the use of singular forms (*I, me, my*) decreased significantly but that the use of plural forms (*we, us, our*) did not, nor were there any changes in the use of *it*. In general, the fact that learners used substantially more pronouns overall that native speakers suggested that there was room for improvement, and the SA period did not appear to induce this improvement.

As for the use of connectives, we found that the proportion of connectives in learners' essays (%Con) significantly decreased over time and in the direction of native speaker norms, presenting further evidence of improvement in the quality and sophistication of their texts. There were significant changes in additive, causal, and logical connectives, but not in temporal connectives, where learners appeared to perform like native speakers from the beginning of the study. Again, the observed changes were significant after the SA context but not after the AH context. Several previous studies (Granger & Tyson, 1996; Milton & Tsang, 1993) have found that non-native speakers tend to overuse connectives in comparison to native speakers, such that a decrease was a clear sign their writing was becoming more "nativelike". Milton and Tsang explored the use of connectives (which they refer to as "logical connectors" (p. 222) in a very large corpora of learners in Hong Kong, in comparison to large L1 corpora. They found that the learners both overused and misused connectives, often inserting a connective where none was required and unintentionally confusing the meaning of their utterance. We found similar evidence of misused connectives in our own corpus, with frequent doubling of connectives e.g. "*But, nevertheless*"; "*Besides, furthermore,*" and other redundant uses that suggested the learners were using connectives as "ornamentation" and not as methods of conveying precise logical relationships between ideas. To use Milton and Tsang's words: "Cohesion becomes an end in itself rather than a means to achieve the ultimate goal: coherence" (p. 229). They ascribe this misuse of connectives to the nature of EFL instruction in secondary schools in Hong Kong and the fact that cohesion may be promoted in classrooms with too little attention to its communicative purpose: "A brief survey of English textbooks currently used in secondary schools in Hong Kong reveals that discourse markers, including logical connectors, are often taught in an absence of situational or cultural context" (p. 231). While connecting such a survey is outside the scope of the present study, it rings true anecdotally and may explain why our participants made progress in the SA context, after a break from formal EFL materials of this nature.

7.1.2.1 A closer look at progress made by one learner

In order to enhance our understanding of all changes in all domains, we chose to explore all FLACC measures in the writing of one participant across the three data collection times. We selected the first L2 participant in our corpus who saw an increase in Total writing score after *both* the

AH and the SA learning contexts so that we might have the opportunity to observe steady progress in quantitative characteristics. Table 7.1 shows the full transcripts of the three essays written by this participant (Participant #5), with both errors and connectives highlighted for illustrative purposes (Table 7.2 gives descriptive data for the set of FLACC measures).

Table 7.1 Essays by Participant #5* at T1, T2, T3

T1 (Total score: 79)	T2 (Total score: 85.5)	T3 (Total score: 95)
In the last few years, the pretty massive arrival of immigrants has brought up /$LexE:cho/ a serious debate about the matter of their integration. **On the one hand**, some people think that immigrants should totally abandon his or her /$GrE:ag/ own customs **in order to** adopt those of their new country **and** to "fit in" /$GrE:fw/ the society **and** the culture who /$GrE:fw/ receives them. **On the other hand**, **however**, some people disagree with that point of view. They thing /$Sp2/ that, **rather than** completely giving up one's customs -which are **actually** a big part of one's identity-, immigrants, **and**, in general, anyone who moves to a foreign country -whatever the reason behind the move is- should adopt some of the customs of their new culture **while** also keeping some of their old one /$PragE:ref/.	Moving to a foreign country, away from all the people **and** customs that are familiar to you, must not be easy: you have to deal with adopting the way of life of your new country **but**, **at the same time**, you probably feel the need to keep some aspects of your own culture. **So**, what should one do? In /$Sp2/ think the best option is to find a balance. **Firstly**, I think it is essential that you get used to your new country's culture. You must learn their language, you don't have to be afraid to use it **even if** you are sure you are making tons of mistakes **and** you must try to meet new people from the country you have moved to. You must let them show you how they live **and** how they act, **and** you have to adopt some of these customs **as if** they were your own. **However**, you can't or must /$GrE:neg/, forget your identity. You are who you are **because** of the culture	One may move to a foreign country for many different reasons **and**, **therefore**, one may want or have to adjust to the new surroundings in many different ways. A student who, **for example**, goes abroad for a while, will probably try to adopt as many as /$GrE:fw/ the local customs as possible, **for** a student who decides to move to a foreign country is generally speaking willing to learn about new cultures, to meet new people **and** is there mainly to discover the world. **And**, of course, a student knows, **in most cases**, that his or her stay will eventually end **and** that he or she will be able to go back home. **On the other hand**, someone who must move to a foreign country as an immigrant, wether /$Sp1/ it is for political or economical /$LexE:cho/ reasons, will be a lot less pleased at /$GrE:fw/ the need of adjusting to a new way of life. An immigrant probably feels forced to accept new customs **and** habits, **and** the usually difficult situations they very often found /$GrE:v/ themselves in don't make it any easier. **However**, these people

This way, there would still be integration **but** people would still be able to be themselfs /$Sp1/ **and** maybe, with time, some parts of their original culture would become a part of the new one. **Because** this is, actually, something that has happened all **through** human history. **Summing up**, we can all have different opinions about this topic **and**, **although** I better agree /$GrE:adv/ with the second one, only time will tell.	**because** of the culture you were raised in, **and** nothing will change that. You can continue using your language **when** talking to your friends **and** family back at /$GrE:fw/ your country **and**, why not, you can even teach it to your new friends. You may want to celebrate your favourite holidays **and** share those special traditions you love with them **and also** cook for them to show them how good food is where you were born. **In conclusion**: I think you should keep the best of both worlds to build a new identity for yourself **and** maybe even for those around you. **Because**, **afterall** /$Sp1/, every culture is a mix of other cultures that just happened to meet.	should definitely try to fit in: learning the local language, taking part in the events organized by the community **and** even maybe joining some societies or organizations will certainly help them to start feeling a bit "at home" **and** will also make it easier for the locals to get to know **and** accept them. This does not mean, of course, that someone who moves to a foreign country has to forget absolutely everything about his or her own culture, customs or way of life. It's this background what /$GrE:fw/ makes us be who we are /$ PragE:idio /, **and** denying or forgetting this aspects /$GrE:ag/ own customs **in order to** adopt of our personality would be giving up our identity, which is just not good! **In conclusion**, I think that while one should try to adjust as quickly as possible to the new country, one's identity must never be forgotten.

*Connectives highlighted in bold. Errors marked in red; see error codes in Appendix 4.

As we can see, the length of Participant #5's essays increased steadily over time, suggesting steady progress in fluency, while the number of sentences produced increases at T2 but then decreases at T3, suggesting that #W and #S did not have a clear correlation across the writing. The use and organization of paragraphs does not appear to change in any systematic way: at both T1 and T3, the writer opts to write 5 paragraphs with very short (1 sentence) introductory and concluding paragraphs, while at T2 the writer opts for 4 paragraphs with slightly longer introduction and conclusion.

Table 7.2 Frequency counts: FLACC measures in essays by Participant #5

Fluency/Lexical				Accuracy			
	T1	T2	T3		T1	T2	T3
#W	201	284	324	EFS	2	10	5
#S	7	12	9	TotE	6	2	7
GI	8.11	8.19	9.39	Gre	4	2	5
AG1k	.92	1.01	1.11	Lex	1	0	1
HyN	4.15	4.35	4.29	Prag	1	0	1
HyV	1.33	1.31	1.28	Sp1	1	1	1
				Sp2	1	1	0

Complexity/variety				Cohesion			
	T1	T2	T3		T1	T2	T3
MLS	28.7	23.7	36	RefP	.33	.73	.63
DC/S	1.14	1.42	1.44	RefA	.50	.73	.75
MLC	11.2	8.11	12.0	RefC	.078	.180	.096
SYNNP	.893	.579	1.015	%Con	.37	.30	.31
StrutA	.069	.065	.041	Con	*17*	*22*	*17*
Temp	.750	.773	.750	DenPr	59.7	172.5	64.8

In terms of vocabulary use, both GI and AG1k increase in a linear fashion from T1 to T3, indicating that Participant #5 is using a greater variety of words as well as more sophisticated words; however the increase, particularly for GI, is more dramatic after the SA period. At T1, Participant #5 only uses words from the first two bands of the BNC (1K and 2K words, see Heatley, Nation, & Coxhead, 2002), with the exception of the word "*customs*", which is in the prompt. The same is true at T2; however the proportion of 2K types to 1K types increases (16/121 at T2, in comparison to 9/102 at T1). At T3, Participant #5 uses a handful of words that are not on the 1K or 2K list: in addition to *customs*, we have *abroad*, *habits*, *personality*, and *surroundings*). Noun hyponymy appears to increase from T1 to T2, indicating that the writer uses more concrete, specific words after the AH context, and there is a slight decrease at T3 but it remains substantially higher than at T1. This seems coherent when considering that the T1 essay expresses opinions in entirely general terms while by T2 the writer uses adds some concrete examples (e.g., "*You must learn the language…*" "*you can celebrate your favourite holidays and share those special traditions… cook for them…*"), and this trend is continued at T3. Verb hyponymy seems to mostly remain constant, though the minor fluctuations are in the direction of decreasing hyponymy from T1 to T3. In general, Participant #5 seems to make use of a roughly similar set of verbs in all three essays (high frequency verbs and auxiliaries along with topic-related verbs like *adopt*, *adapt*, *keep*, *abandon*). This observation along with the observation that the group means do not change much over time nor vary much from the NS mean suggest that this performance was typical and that HyV was not a descriptive measure in our corpus.

In the domain of accuracy, we see a marked improvement from T1, where there were only 2 error-free sentences (or %33 of all sentences), to T2, where there were 10 EFS (83%). The total number of errors also dropped, from 6 to 2, and lexical and pragmatic errors entirely disappeared. Although the percentage of error-free sentences decreases somewhat at T3, it remains well above T1 values, and the total number of errors at T3 represents a substantial improvement over T1 when text length is taken into account. Furthermore, qualitative analysis supports our intuition that error gravity may play a larger role than error frequency, and that the learners may improve significantly in the former category even if not in the latter. In Participant #5's writing, we do not find many errors that confuse the writer's meaning; however those that seem to be the most problematic are found in the T1 essay: e.g., the pragmatic error at the end of paragraph two, *"...[they] should adopt some of the customs of their new culture while also keeping some of their old one*", where the clumsy wording makes it unclear if the "one" refers to culture or if it refers to customs and contains an agreement error. None of the errors in the T2 or T3 essays create similar levels of ambiguity.

For syntactic complexity, we see that the length of sentences decreases between T1 and T2, from an average of 28.7 words/sentence to 23.7 words/sentence, but then increases again at T3, to 36 words/sentence, following an inverse pattern as that seen for #S and further suggesting that these two measures must be considered in tandem as developmental indices. The ratio of dependent clauses per sentence is well above 1 at all three data collection times, reflecting Participant #5's tendency to use very long sentences broken up with assorted punctuation such as hyphens, commas, and colons which, while not ungrammatical, are often quite clumsy. Given that the group means are all above 1 for this measure, it may be that this performance is typical. The pattern of development for MLC is quite similar to the pattern observed for the group as a whole; however Participant #5's MLC at both T1 and T3 is highly comparable to the mean observed for the native speakers (11.6), which was not the case for the group as a whole. Since the number of dependent clauses per sentence remains the same between T2 and T3 yet sentence length increases substantially, it is clear that the increase is primarily due to longer clauses and noun phrases, both of which are evidence of more sophisticated grammar. There is a decrease in structural overlap between

T2 and T1; however tense and aspect repetition remains mostly the same, with an increase at T2 (in the opposite direction of native speaker norms) and a return to T1 levels after the SA.

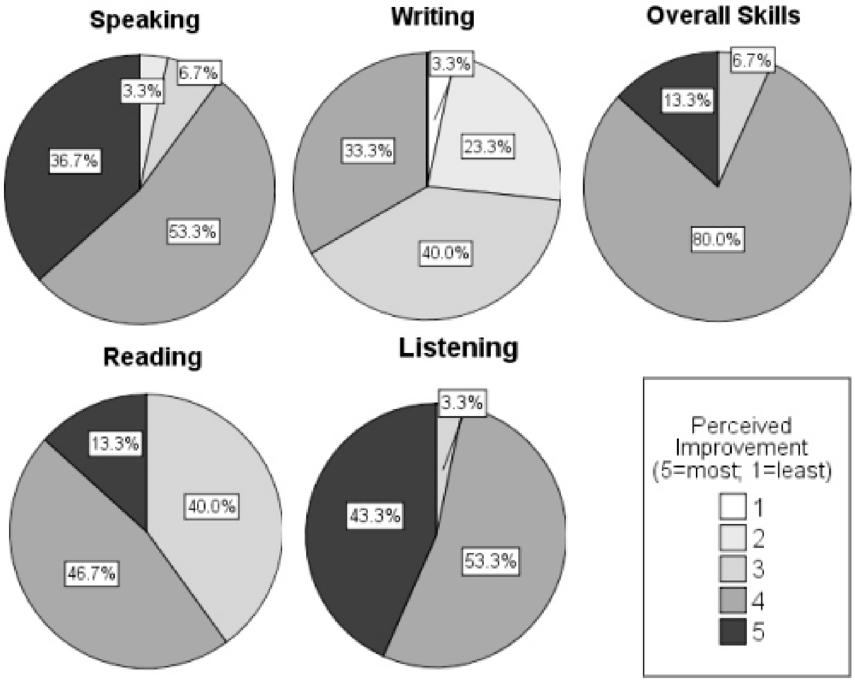
Finally, in the domain of cohesion, we see that referential overlap seems to increase and then plateau, while the use of connectives seems to increase and then decrease. As with the number of errors, although Participant #5 uses the same number of connectives at T1 and at T3, the substantial increase in fluency at T3 is such that connectives take up a proportionally smaller amount of the text. With a %Con score of just .37 at T1, Participant #5 uses relatively fewer connectives than his or her peers (the group mean was .40) and is closer to the mean score of the native speakers (.35). When we consider the specific connectives used, we can see that although there is not a dramatic change in the proportional frequency of connectives, there is some progress in the types of connectives chosen. For example, the essay at T1 uses the more informal connective expression *summing up*, which is updated to the more formal, academic *in conclusion* in the T2 and T3 essays. Furthermore, at T2 and T3 the conjunction *and* is used as or more frequently than all other connectives combined (*and* is 45% of the connectives at T2 and 59% of the connectives at T3). Milton and Tsang (1993) highlight that *and* is the one category of connectives that is used with similar frequency in the very large L1 and L2 corpora examined in their study; thus the greater use of *and* in comparison to other connectives is further evidence that Participant #5 is performing similarly to native speakers in this domain but does make some subtle progress (*and* is only 23% of connectives at T1, where a greater variety of logical connectives is used, e.g., *because, rather than*). Finally, we see a dramatic increase in the proportion of personal pronouns used at T2, which then decreases to near original values at T3. Interestingly, we can see that the T2 essay differs from the other two in its use of the second person “you”. While in this specific instance, the use of “you” is reasonably effective, and the use of concrete examples represents an improvement over the more abstract and undetailed essay in T1, in general “you” is associated with informal, spoken registers (Biber, 1988) and thus some of the improvement at T3 is likely due to the more formal third-person perspective adopted at this time.

7.1.3 Real and perceived improvement

Overall, despite evidence that individual participants, like Participant #5, made progress in both learning contexts, evidence at the group level suggests that for our participants the SA context was relatively more beneficial to writing. Not only did they significantly improve their scores on the Jacobs scale, indicating that after their 3-month stay they wrote texts that were perceived as being of higher quality in terms of Content, Organization, Vocabulary, Language Use, and Mechanics, but there were a number of objective changes in their texts that suggested progress and that affirmed the validity of these subjective ratings. After returning from the SA, we found that participants' writing showed signs of improved fluency, lexical diversity, improved spelling accuracy, increased syntactic complexity at the phrasal level, and a more native-like use of connectives. In contrast, after the AH context, the only gains at the group level were an increase in syntactic variety as measured by a decrease in tense and aspect repetition. Again, although many individual participants made progress during the EFL classes at home (remember, 18 of the 30 participants did see improved qualitative scores), the lack of systematic changes in this context—and the presence of so many clear changes in the SA context—clearly suggests that SA is a particularly beneficial learning context for learners at advanced levels of proficiency and will serve to increase the fluency, complexity, accuracy, and cohesion of their writing. This finding is in line with those reported by Perez-Vidal and Juan-Garau (2009) and can be contrasted with the findings of Freed, So, and Lazar (2003), which is frequently cited in the SA literature as evidence that writing does not benefit from SA in the same way as speech (e.g. DeKeyser, 2007). It is particularly interesting given that the learners reportedly do not spend much time practicing writing, which DeKeyser's (2007) theories suggest might be essential for improvement. On the other hand, given our understanding of writing ability and the role of proficiency, it seems likely that an overall increase in proficiency—potentially gained through practicing oral language or simply through exposure to high quality input—allowed learners to perform better in writing and to demonstrate more of the composing competence acquired in formal education contexts in their previous education, both in their first and foreign languages.

Finally, our findings may be contrasted with the participants' perception of their own improvement, taking advantage of a questionnaire administered by SALA researchers at T3, which asked participants to reflect on their SA experience and the extent to which they had improved their different language skills. Participants were asked to rate their progress in different linguistic areas on a 5-point scale, with 5 indicating the most improvement and 1 indicating the least improvement. They rated progress in speaking, listening, reading, writing, and overall skills. For the 30 participants in our study, their self-assessment of improvement in overall skills was overwhelmingly positive: 93.3% of participants rated their improvement with either a 4 or a 5, and the remaining 6.6% rated their improvement with a 3, indicating that all participants felt that they had improved their language skills at least moderately. When the perceived improvement was evaluated separately for different skills, however, we found that writing was the skill that received the lowest rating: only 33% of participants indicating they felt their writing had improved considerably (either a 4 or 5) as compared to 90% for speaking. (The distribution of participants' ratings for all skill areas is provided below in Figure 7.1.) The fact that participants did not perceive themselves has having improved much in writing confirms early findings by Meara (1994) that participants are more pessimistic about writing than other skills; however the simultaneous finding that writing does indeed improve significantly after the SA context supports DeKeyser's (2007) argument that self-report data is not a reliable measure of progress and that empirical analysis may refute results obtained from such studies.

Figure 7.1 Participants' perceived improvement in speaking, writing, reading, listening, and overall skills



7.2 The effect of initial level

In the next section of our results we considered whether longitudinal improvement, over time and after each learning context, was different for participants with different initial levels of proficiency, in response to RQ2. *Is writing improvement different for learners with different initial levels of* *proficiency?* We found that improvement *was* different, but that the nature and extent of differences depended upon how initial level was measured. We will discuss the nuances of these results in relation to our two subquestions (RQ2a and RQ2b).

7.2.1 Relationship between IWP and IGP

We elected to evaluate initial level using two different grouping variables: writing scores at T1 (IWP) and grammar and cloze scores at T1 (IGP). We began this section by exploring the relationship between these two variables, which was of theoretical interest as well given the body of literature reviewed in Chapter 2, which proposes that linguistic proficiency and writing proficiency are overlapping but separate competences (e.g., Cumming, 1989; Krapels, 1990; Sasaki & Hirose, 1996). Our own results confirmed this: we found that the correlation between writing score and grammar scores at T1 was significant but only moderate (τ = .328), suggesting that the two dimensions of proficiency were distinct (the same was true when we considered correlations in the larger L2 corpus, at τ = .478). That is, these correlations indicated that a non-trivial portion of our L2 participants presented proficiency profiles that were "mixed" to a certain degree (i.e., they had high IWP and low IGP, or high IGP and low IWP). The two learners profiles in Table 7.3 exemplify these mixed profiles.

Table 7.3 Comparison of T1 essays of two L2 participants with mixed LGP and writing scores

Participant #6 at T1	Participant #27 at T1
Total Score: 81 *(Rank 23 of 30)*	Total Score: 73.5 *(Rank 7.5 of 30)*
Gram/Cloze: 16% *(Rank 4 of 30)*	Gram/Cloze: 76% *(Rank 29 of 30)*
Human is an animal of customs. It means that people can adapt themselves to everything. When someone moves to another country gets influences from it; from television, newspapers....It has advantatges and disadvantatges. *First of all, it is necessary to state the everyone has a different personality. No one is like other person. You would think the same everywhere you be but on the other hand, there are some aspects that can't be kept in another country. For intance the dinner time or bed time is different all around the world. These should be adopted.*	*If you adopt a strictly rational point of view, you can hardly find a strong argument to defend the statement above. In fact, one should be free to live the way one decides, according to the customs of any culture, without any constraints. But, nevertheless, if you regard things from a more realistic point of view, you get to the conclusion that this inconditional individual freedom can lead to social problems. You just need to turn on the TV and watch the news to understand that the current situation in*

Secondly, there are other points than are usually adapted but not as a whole. It is clearly shown when we talk about food. People adapt new types of food and recipes, although they maintain the typical meals of their land. Both ways of cooking are combained without any problems. *Language is a problematic point. There are people who live in a place, where another language is spoken, for a long time and they don't understand it.* *The manners of someone are reflected on this. Maybe you don't want to speak that language but you should understand it in order to make people easier to talk to you and is a sign of respect.* *Consequently, when somebody is in a different place for a long time he/she feels no so confident as he/she was in his/her land. This is a reason why adopting other's customs can make us be happier and it make our lives easier.* *To sum up, there are some points such as beliefs that do not change, some others changes in some way like food or timetable and finally some others that changes obligatory and sometimes without we notice them such as learning a bit of everything, language, food, culture, history. Everyone should learn about other cultures and adopt them without forgetting where they come from.*	*Europe is quite conflictive. It is a matter of fact that an immigrant who adopts the culture of his new country will be much better integrated than one who chooses to preserve his way of life in a place where it may be seen as strange by the local inhabitants. But no one can be demanded to forget or quit his way of life.*

Participant #6 was in the top third of writers at T1, but also received one of the lowest grammar and cloze scores. Participant #27, in contrast, was in the bottom third of writers at T1, but received the second highest score on the grammar and cloze test. When we compare their essays side by side we can see that Participant #6, despite a weaker command of grammar and lexis, wrote a much longer essay than Participant #27, broken into paragraphs and showing awareness of the appropriate structure of an argumentative essay, including an introduction and a conclusion. Although there are many basic errors (e.g., the opening sentence: "*Human is an animal of customs*"), and the writer clearly struggles to express complex ideas with a limited grasp of the language, the relatively high score indicates that these errors probably did not confuse or obscure the intended meaning for raters. Participant #6 provides a number of concrete examples related to customs that can or should be adopted (e.g., food, language) and then recovers these examples in the conclusion and ends with a final persuasive note that expresses a clear opinion (everyone should adopt customs without forgetting their own origins), thus responding to the task requirement to agree or disagree. Participant #27, in contrast, despite showing a firm grasp of the grammatical rules of English and an impressive vocabulary, produced a text in a single paragraph, with no clear introduction or conclusion, thus showing little understanding of the requirements of the argumentative essay genre. Furthermore, Participant #27's opinion on the topic is somewhat unclear: the opening and concluding statements suggest that the writer disagrees with the prompt (that a person traveling to a foreign country should adopt the culture); however the single concrete example given (that an immigrant will become more integrated if they adopt the new culture) seems to contrast with this opinion. Both participants exhibit the 'uneven' profiles that Weigle (2002) argues are typical for L2 learners, and which the PROFILE was designed to evaluate: Participant #6, for example, received a very high score in the Content component (in the "Excellent to Very Good" range), and a lower score for Language Use (at the low end of the "Good to Average range"); Participant #27 scored at the high end of the "Good to Average" range for Language Use but at the bottom of this range for Content.

7.2.2 Effect of IWP on improvement

After exploring the relationship between our grouping variables, we conducted analysis in response to RQ2a. *Are changes in perceived quality and FLACC different for participants with higher and lower initial writing proficiency (IWP)?*

First we looked at improvement in the perceived quality of writing (Total PROFILE scores), using mixed-design ANOVA. We found that there was a statistically significant interaction between time and IWP, suggesting that groups with different initial levels of writing ability did have different patterns of development over the two contexts. The learners with high-IWP, those classified as "Advanced" at T1, did not significantly improve their scores after either context: they saw a slight decline in scores after the AH context and then an increase in scores after the SA context;

however as a group, their scores after SA were not significantly different from their initial scores. The group with mid-IWP, those classified as "Low Advanced" based on their T1 scores, showed no change in scores after the AH context but made significant gains after the SA context. Finally, the group that was classified as "Upper Intermediate" based on their T1 scores made significant progress after the AH context, pushing them into the "Low Advanced" category, but then seemed to plateau, making no further progress after the SA context, and their scores remained well below the other two groups at T3. *F*-values for the RM-ANOVAs conducted for each group revealed that the mid-IWP group made the greatest gains over time, followed by the low-IWP group. Although the low-IWP group made substantially more progress than the high-IWP group, their scores at T3 still remained well below the scores received by the other two groups.

We then considered the interaction between time and IWP on progress in the domains of FLACC. We found that there were significant interactions for 5 variables: #W, #S, GI, DC/S and DenPr, showing that there were qualitative differences in the amount and nature of progress across the three groups. The high-IWP group began the study with significantly greater fluency and lexical diversity than the other two groups, but that after the AH context they were overtaken by the mid-level group, and after the SA context the mid-level group was outperforming them on all three measures. Both groups maintained their advantage over the low-IWP group throughout the duration of the study, although the degree of differences between the groups decreased across contexts. The high level group appeared to increase the number of dependent clauses used, while the low and mid-level groups decreased their usage. After the AH context, the low-IWP group saw a substantial decrease in the number of dependent clauses produced, which corresponded to their improvement in qualitative scores. After the SA context, the mid-level group saw a substantial decrease, again in line with qualitative improvement. These observations corroborated the observation that higher quality writing used less subordination; however the finding that the low-IWP actually produced fewer dependent clauses per sentence than the high-IWP group at T3, was somewhat surprising and suggested that DC/S has a more complex relationship with both writing quality and linguistic proficiency, as we explore in relation to RQ4. Finally, the high level group seemed to increase their use of personal pronouns from beginning to end of the study, while the lower two groups increased their usage after the AH context but then saw a dramatic decrease after SA. The decrease observed for the lower two groups is coherent with our understanding that excessive use of personal pronouns is associated with weak or informal writing; however the increase seen for the high-IWP group, especially after the

SA, is somewhat surprising and, along with the increase in DC/S suggests that their writing became more informal and showed more evidence of features associated with spoken registers in Biber's (1988) model. This may lead us to speculate that interaction with the spoken variety of the language influenced their writing in line with the strain of research focused on ways in which the oral language may "interfere" with writing (see Sperling, 1996, for a review); since only the learners with high levels of initial writing proficiency were influenced in such a way, we may further speculate that perhaps these more advancer learners interacted more with native speakers and thus had more contact with spoken English (following Segalowitz & Freed, 2004; Freed, et al., 2004).

Overall, in relation to RQ2a, we found that there were clear differences in development in relation to IWP, and that the mid-level group (those in the "Low Advanced" range) improved the most in comparison to their peers. These results match up well with the 'Threshold Hypothesis' discussed in Chapter 1. That is, previous research has suggested that students need to have a certain threshold level of proficiency in order to take advantage of the opportunities for SLA in SA contexts, but that once participants are over this threshold, higher level learners will improve relatively less due to the normal learning curve (Brecht, et al., 1993; Milton & Meara 1995). Our findings that the most advanced learners make comparatively less progress in qualitative scores may thus easily be interpreted using this threshold hypothesis. (The reduced progress of the advanced group was not due to a ceiling effect, since the high-IWP group's mean score of 86.5 at T3 (SD = 6.2) was still well below the mean score received by the native speakers and below the "Upper Advanced" score range.) It may be that the only way to progress from advanced levels to "upper advanced" levels is through only extensive practice, which is in line with theories about the development of writing expertise offered by Flower (1979), Bereiter & Scardamalia (1987), and other important scholars of L1 writing. On the other hand, less proficient writers who still make basic errors may benefit from mere instruction or exposure, in ways that will increase their scores. This may be particularly the case for features of style and structure that may be easily corrected or taught but have a large impact on readers. Consider the case of Participant #27, profiled above as a writer who, at T1, demonstrated high levels of lexico-grammatical competence but received a very low writing score. If we look at Participant #27's essay at T2, in comparison to the essay from T1 (both are reproduced in Table 7.4), we can see that the practice and instruction given in the EFL classes AH were sufficient to alert the writer that an academic essay requires paragraphing, and should provide an introduction and conclusion. Meeting these basic requirements of the genre was enough to earn Participant #27 (who already had strong Language Use and Vocabulary scores) a T2 Total score of 88, now well into the "Advanced" range.

Table 7.4 Participant #27 essays at T2 and T3

Participant #27 at T1	Participant #27 at T2
If you adopt a strictly rational point of view, you can hardly find a strong argument to defend the statement above. In fact, one should be free to live the way one decides, according to the customs of any culture, without any constraints. But, nevertheless, if you regard things from a more realistic point of view, you get to the conclusion that this inconditional individual freedom can lead to social problems. You just need to turn on the TV and watch the news to understand that the current situation in Europe is quite conflictive. It is a matter of fact that an immigrant who adopts the culture of his new country will be much better integrated than one who chooses to preserve his way of life in a place where it may be seen as strange by the local inhabitants. But no one can be demanded to forget or quit his way of life.	*Lately, immigration has become one of the major worries of the countries in the EU. Every year, thousands of people quit their countries escaping from poverty, any kind of political dictatorship or wars, in order to find the life they have been longing for in Europe.* *Thus, Western societies have to face the arrival of thousands of people to whom the Western values and way of life are strange, and integrate them efficiently. Obviously, this is not an easy task, as both immigrants and "hosts" suffer what is called a "cultural shock".* *As it has been stated above, the integration of the immigrants always has to face this "culture shock" which gives pace to multiple difficulties derived from cultural differences. Therefore, a way to overcome these difficulties should be found. Some say that the immigrants should be the ones to adopt the culture of their new country. Others argue it should be the native people who should help the immigrants.* *In my opinion, no one has the right to demand someone to abandon their own culture and embrace a new one, but some small concessions can be done in order to improve cohabitation. Both immigrants and natives can help by trying to understand each other's culture, but, if it were to be chosen, the native culture would have preference beyond the others.*

7.2.3 Effect of IGP on improvement

Next we regrouped participants in order to answer RQ2b. *Are changes in perceived quality and FLACC different for participants with higher and lower initial lexico-grammatical proficiency (IGP)?*

Due to the many tied ranks in grammar and cloze scores we were not able to form 3 coherent groups based on IGP and thus we had to eliminate the mid-level group, which unfortunately meant we could not directly compare the different patterns between our two 'initial level' variables. When we compared the high and low scoring participants, we found that both groups made significant progress over the course of the study and that the high level group appeared to make steady progress in both learning contexts while the low-level group plateaued in the AH learning context and saw a significant jump in scores after the SA context. This differed from the findings observed for IWP, where the low-level participants improved in both contexts and appeared to improve more in AH than in SA. The interaction between time and IGP was not significant for Total scores but was significant for scores on the Vocabulary component: the group with lower levels of IGP did not make any progress in the domain of Vocabulary while the higher-level group did. Exploration of the other component scores suggested that improvement in the lower level group was primarily due to Language Use and Mechanics scores, and that the other areas did not improve significantly over the course of the study. In contrast, the group with high IGP made significant progress in all areas, and particularly in Vocabulary, Organization, and Content. These results are in line with our hypotheses about the ways in which lexico-grammatical proficiency may influence L2 writing ability, as explored in Chapter 2, and with observations made by Manchón (2009) and colleagues in the Murcia research group. That is, participants with adequately high levels of lexico-grammatical proficiency may have more time available to focus on solving "higher order" problems related to discourse organization and style, while the participants with lower levels of lexico-grammatical proficiency may dedicate the majority of their time and attention to "lower order" features related to language use and mechanics. This interpretation was further supported by exploration of between-groups differences in the component scores at each of the three writing times: we found that, as expected, the high-IGP group received significantly higher scores than the low-IGP group, but only in the domains of Vocabulary, Language Use, and Mechanics. This suggested that these learners were similar in terms of their control of Content and Organization at T1. At T2, the group with high-IGP outperformed the group with low-IGP in all domains, suggesting that they were able to focus more attention on improving features of Content and Organization during the period of AH study. At T3, the higher IGP group maintained their advantage in all domains except Language Use, suggesting that the low-IGP group was able to catch up in this respect after the SA period. Interestingly, the low IGP group's scores at T3 were highly similar to the high IGP group's scores at T1, in terms of group means and medians and in terms of the distribution of scores. Finally, an exploration of individual differences revealed that a larger number of participants in both groups made gains during SA than during the AH context, and that the relative benefits of the SA context were more pronounced for the higher level group than for the lower level group.

Overall, when we consider the results in relation to IWP and IGP together we have seen that we may tentatively answer *yes* to RQ2 (*Is writing improvement different for learners with different initial levels of proficiency?*) but that the differences are not clear-cut and depend upon how we define proficiency. When we consider IWP, it appears that the mid-level learners benefit the most over the course of the study, and that the most advanced learners benefit the least by a substantial margin. In contrast, when we consider IGP, though both groups improve and differences are less dramatic, it appears that the high-scoring learners are those who improve the most. Not only that, but the high-scoring learners improve in the AH context, which contrasts with the findings for IWP. Clearly these two types of proficiency are capturing groups of learners with different profiles: it may be that the participants who score high on the grammar and cloze tests are more academically oriented in general and are more motivated during the period of AH study, which would explain their improvement during this context, though both groups improve about the same amount in the SA. It may also be that, as we discussed above, participants who scored especially low on the T1 writing test suffered from basic errors in structure or style that could be easily corrected in the AH period, while the higher level learners needed more time to improve the nuances of their writing styles. In general, the significant interactions between time and IWP, and between Vocabulary scores and IGP indicate that the different patterns of development for participants grouped by level are not accidental, and that students with different initial levels respond differently.

7.3 Comparisons with native speakers

The next research question addresses differences between learners and native speakers: RQ3. *How do learners' essays compare to those of native speakers, in terms of the perceived quality of their essays and in terms of FLACC?* We had already observed a number of differences when looking at the descriptive statistics reported in relation to RQ1, but wanted to test these differences statistically and determine whether the learners converged on certain FLACC measures. Again, we explored the two sets of dependent variables (quantitative and qualitative) in two stages.

First we considered differences in perceived quality, to respond to RQ3a. *How do learners' essays compare to those of native speakers in terms of* *perceived quality?* All 28 of the native speakers' essays fell into the "Advanced" or "High Advanced" range, with the majority scoring between 92-100 points, which was expected given the design and goals of the PROFILE and the fact that 3 of the 5 components are largely focused on linguistic competence. The fact that not all of the native speakers fell into the "Upper Advanced" range are further evidence of the variability in the L1 population noted by McNamara, Crossley, and McCarthy (2010) and many others. That is, although the native speakers were all university students, the majority in their 3rd year, had presumably passed entrance exams for entrance to university and were experienced with the type of writing expected in English-speaking university contexts, about a third of them (n = 10) did not score in the highest possible category. That said, the native speakers' mean scores were still significantly higher than those of the L2 learners at T3, primarily due to their greater linguistic competence: when we considered differences in component scores, we found the largest effect size was found for the Language use and Vocabulary components, and the smallest effect size for the Content component. Overall, the findings in relation to RQ3a suggested that the learners' essays were perceived as being of significantly lower quality, and were in line with previous research, which consistently arrives at this finding, even for advanced learners. For example, in Silva's (1993) review of research on L1/L2 differences, he cites 11 different studies that considered overall quality of texts, all found of which reported that the L2 texts were perceived as being of lower quality (or, "less effective" p. 663) than the L1 texts.

In order to better understand the differences between learners and native speakers, we next looked at quantitative differences, in response to RQ3b. *How does learners' writing compare to NS writing in terms of FLACC?* We found that the learners converged with the native speakers in terms of both fluency and lexical diversity (GI) after the SA context, but that the native speakers still used significantly more sophisticated word types, as measured by AG1k. The latter finding is consistent with previous research that has shown that lexical sophistication is a key difference between L1 and L2 writing (Hinkel, 2003). Hinkel looked at 1,083 argumentative essays written by learners and native speakers (NS n = 206); he reported that learners used the class of words that he defined as "vague" nouns (e.g. *thing, stuff, people, man*) far more frequently than native speakers and tended to overuse the same vague nouns. In our case, there did not appear to be a systematic difference in the use of vague nouns, as this would have been captured by HyN; however the learners clearly used more frequent and therefore less sophisticated words, even after the SA. In the domain of accuracy, we found, in line with the many previous studies reported in Silva (1993), that the native speakers performed significantly better than the learners on all measures, and that even after the SA the learners produced significantly more errors and fewer error-free sentences. We found that there were significant differences for all indices of syntactic complexity except for mean length of sentence (MLS). As we said before, the means for MLS were similar across all three data collection times in the L2 corpus and this measure did not appear to be a meaningful descriptor of either writing or linguistic proficiency in our corpus, at least when considered independently from the number of sentences produced. The native speakers produced significantly fewer dependent clauses per sentence than the learners at either T2 or T3, again suggesting that at advanced levels of proficiency one should interpret a decrease in the amount of subordination as a sign of improvement, particularly when seen alongside a corresponding increase in clausal complexity. Despite the improvement made in clausal complexity (MLC) over the course of the study, the learners at T3 still produced significantly shorter clauses than the native speakers, and used significantly less noun phrase modification as well. In terms of syntactic variety, there were no differences between the learners and native speakers in terms of structural overlap and the learners appeared to have converged with the native speakers in terms of tense and aspect repetition (Temp) since there were significant differences observed between native speakers at T1, before making improvement in the AH context, but not after that. The lack of differences in structural overlap suggests that there was not enough variation in StrutA to make this an informative measure of syntactic variety in our corpus, and again we argue that more precise syntactic variety measures would be a practice and interesting topic of future research. In the domain of cohesion, there were significant differences between the learners and native speakers at T2, for both referential overlap (RefP) and for connectives (%Con, %CausCon, %LogCon), but these differences disappeared after the SA when significant progress was made in this domain. The finding that the learners at T2 overused connectives in comparison to native speakers was consistent with the previous findings of Milton and Tsang (1993) discussed above, in that the difference in connectives was reflected more in causal and logical connectives than in additives, which included *and*, the most frequent connector. Schleppegrell (1996) also documented an overuse of connectives in her study of ESL writing, focused particularly on the connective *because*, which falls into both the causal and logical categories, and which she argues is overused because it is so common in spoken English. The finding that our learners were able to reduce their overuse of connectives and converge with the native speakers is especially encouraging in light of all the previous studies that have documented NS/NNS differences in this domain, and suggest that our learners are nearing the extremely high levels of proficiency to which they aspire. For the final measure of personal pronoun usage, however, the learners continued to use more personal pronouns than the native speakers even after the SA, indicating that they were still relying on features associated with informal, spoken registers and had room for improvement in other domains of style and register.

Overall, the results in relation to RQ3 confirmed the large-scale meta-analysis conducted by Silva (1993) who reported on a wide variety of differences between L1 and L2 writers and a persistent difference in perceived quality. On the other hand, the convergence with native speakers on a number of FLACC measures, including fluency, lexical diversity and cohesion, do indicate that the learners began to write more "native-like" texts after the SA period, and provide further evidence that they make considerable progress after this context and over the course of the study.

7.4 Relationship between FLACC, perceived quality, and LGP

Our final research question inquired into the relationships between FLACC measures and our measures of essay quality and linguistic proficiency: *RQ4. Do FLACC measures have the predicted relationships with a) writing quality and b) lexico-grammatical proficiency?* We looked at this question in two parts, focusing first on qualitative writing scores, as measured by the PROFILE, and then focusing on grammar and cloze scores (and on comparisons with native speakers).

7.4.1 FLACC and perceived writing quality

First we analyzed the relationship between FLACC measures and qualitative writing scores, in response to RQ4a. *Which FLACC measures are significantly correlated with qualitative writing scores and which discriminate between high and low scoring learners?* To answer this question we conducted correlations between quantitative measures and writing scores for the full corpus, and for the L1 and L2 corpuses independently, and also considered which measures discriminated between the 10 highest scoring learners (with a mean Total score of 89.9) and the 10 lowest scoring learners (with a mean Total score of 72.5).

In the domain of fluency, we found significant relationships for both measures. The number of words produced (#W) was strongly correlated with writing scores in the full corpus (r_s = .432, p < .01) and in the L2 corpus (r_s = .492, p < .01), but no such significant correlation was found among the L1 essays. The positive correlation observed for the L2 corpus was in line with our predictions and in line with previous research exploring the relationship between essay length and holistic scores in timed writing (e.g., Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000; see also a number of early studies cited in Weigle, 2002). For example, Ferris (1994) looked at variation in 160 ESL compositions written in under highly similar conditions to our own (argumentative essays, on the topic of culture shock, written in a 35-minute period and evaluated by three raters on a 10-point holistic scale) and considered which of Biber's (1988) 62 features best discriminated between the high and low scoring essays. In her regression analysis of the 28 features associated with essay quality she found that essay length was the single largest contributor to variation, by a significant margin: the number of words contributed 37.6% to variation in R^2, as compared to synonymy/antonymy, the next largest contributor, which contributed 6% (p. 418). Similarly, Grant and Ginther (2000) found that essay length increased in a linear fashion between essays receiving scores of 3, 4, or 5 on the TOEFL TWE scale (also argumentative essays written in a 30-minute time frame).

Our finding that #W did not correlate with the native-speakers' scores was somewhat surprising given the results of previous research. For example, Frase et al. (1999) looked at variation in 1,737 TWE essays written on two topics in the standard 30-minute time period considering compositions written by 3 non-native L1 groups (Arabic, Chinese, and Spanish) and also by a fourth group of native-English speakers, both within and outside of the US. They found that #W correlated strongly with holistic scores for all four groups, ranging from $R^2 = .65$, for the native English speakers, to $R^2 = .82$, for the native Spanish speakers (far stronger than our correlation of $r_2 = .492$ for the L2 corpus). One possible explanation for the lack of significant correlations in our L1 corpus is the small sample size: we had only 28 NS participants, and although the correlation was not significant ($r_2 = .281$, p = .147), it was large enough to suggest the possibility that a correlation might have been found in a larger corpus. Alternately, it may have been that 30 minutes was an overly generous time limit for the native speakers. That is, perhaps the prompt used was open-ended and general enough that all the native speakers were able to complete it in well under 30 minutes, without racking their brains for further content and examples. While this explanation is somewhat unlikely given that 30 minutes is typical for essay exams in both L1 and L2 contexts, studies of L1 writers in *untimed* conditions have reported no correlation between essay length and perceived essay quality. For example, McNamara, Crossley, and McCarthy (2010) considered 120 essays written by undergraduate students at Mississippi state which were classified as either high or low proficiency based on scores received on a holistic scale designed for the

SAT (the standard exam used for admissions to US universities). In their study essays were written in untimed conditions and the authors found that there was no significant difference in the number of words used by high and low proficiency writers (p. 67). Unfortunately, although the native speakers, like all participants, were required to finish their essays in less than 30 minutes, we did not record the exact amount of time spent writing or gather any information on the writing process that might have shed more light on this issue.

The number of sentences (#S) produced also significantly correlated with qualitative scores in both the L1 and L2 corpuses. Initially our adoption of #S as a fluency measure was made following Wolfe-Quintero et al.'s (1998) logic that: "fluency means that more words and more structures are accessed in a limited time…Fluency is not a measure of how sophisticated or accurate the words or structures are, but a measure of the sheer number of words or structural units a writer is able to include in their writing within a particular period of time" (p. 25). That is, we felt that evaluating the number of sentences produced might capture something about the number of different ideas expressed in a text, and thus we considered it alongside the number of words in the domain of fluency. Our analyses of this variable in relation to our different research questions suggests that it is inappropriate to consider #S as an index of fluency since is too closely related to sentence length and affected by punctuation choices and errors. That is, although there was a significant correlation between sentences and qualitative scores in the L1 corpus, there was also a significant correlation between sentence-length and scores, discussed below, which cautioned us to view these measures as inextricably related.

When we considered the correlations between essay scores and lexical diversity and sophistication measures we found that both GI and AG1k were significantly correlated with essay scores in the L2 corpus, but that noun and verb hyponymy measures were not. In the L1 corpus, on the other hand, neither lexical diversity or overall sophistication (AG1k) were significantly correlated with scores but noun hyponymy was (p < .01). The results for the L2 corpus confirmed predictions based on previous research which has routinely found that lexical diversity and sophistication are related to L2 essay scores (Engber, 1995; Grant & Ginther, 2000; Yu, 2009), while the lack of significant relationships in the L1 corpus suggested that these variables may be more closely related to issues of language proficiency (e.g., vocabulary size) than essay quality or composing competence. The fact that hyponymy, or the degree of specificity vs. generality of lexis, did not appear to have a significant relationship on essay scores in the L2 corpus suggests that these measures of lexical sophistication may be too fine-grained for differentiating between essays written by non-native speakers; however the fact that noun hyponymy did have a significant relationship to essay quality in the L1 corpus suggests that this variable should not be discounted when analyzing variation in L2 texts written by writers with very high levels of proficiency. Noun hyponymy was positively correlated with L1 essay scores, suggesting that L1 writers who made use of highly specific words were evaluated more positively than those who used more general, abstract words.

Qualitative analysis suggested that high hyponymy values may have been related to the use of concrete, detailed examples, as can be seen in the extracts below in Table 7.6. As we can see, Participant #57, who had the highest HyN score, wrote a highly specific essay with many personal details while Participant #54, who had the lowest HyN score, wrote an entirely abstract essay without such concrete examples (Participant #57 was not one of the highest scoring native speakers, with a Total score of 91, but still slightly outperformed Participant #54, who scored an 86). To provide a further example, many of the L1 participants made references to situations in which not adopting customs might be perceived as rude (as in Participant #54's text); however none of the writers with low HyN values gave specific examples. In contrast, Participant #43, who was another of the higher HyN essays (ranked 22 out of 28) gave the following: "*A good example is when one visits Italy it is customary for women to wear skirts below the knees and shirts that cover their shoulders in church. To disobey these customs and standards would be very offensive to other churchgoers and their is no reason that this custom should not be followed. It most likely does not go against what one is accustomed to; therefore, it is not a stretch to expect the rule to be followed.*" If HyN values indeed indicate the use of concrete examples or details, it is logical that they have been linked to essay quality in the present study, since the Jacobs scale explicitly rewards this trait on the rubrics for the Content component: essays in the "Excellent to Very Good" and "Good to Average" ranges are expected to be "relevant to the assigned topic", but the lower range amends this description with the phrase "but lacks detail".

Table 7.5 Examples of L1 texts with high and low HyN

Participant #57 (HyN rank 28/28)	Participant #54 (HyN rank 1/28)
After having lived in Mallorca for over eleven months now I believe this statement is very true, and that it needs to be repeated to more residents on the island. When I say 'residents', I am referring to those of English and German nationality who have chosen to live on the island for its agreeable	*When one moves to a foreign country, he or she must enter the new culture with an open mind, prepared to adapt to the new customs and way of life. At first, one does not, or probably cannot, immerse his or herself entirely in the environment of the new country. Even neighboring nations*

weather. These foreigners, in many cases have lived for 15 and 20 years on the island and cannot even hold a basic conversation in Spanish with another (Spanish-only speaking) Mallorcan. *In many ways these non Spanish speaking English and German residents have imported their own culture to the island. Although I am American and in many ways I share this culture with them (the English, I mean), I do not think that they should have their "Little Britain" grocery stores and their German-only bars. If the English and German people are so worried about losing their culture due to not living in their home country, well I don't think that they should have even left.* *On the other hand, this importation of foreigners to the island is extremely profitable for a number of different job sectors. Those who choose to purchase homes and cars need English or German speaking lawyers, "funcionarios", and car dealers/real estate companies to serve them. In this way I would say that adaptation is not necessary because it increases the demand of cars/transportation services, housing, and other jobs such as translators.*	*have great differences, and different regions within one nation embrace diverse lifestyles, so the process of integrating into a new society must be a gradual one. One must be tolerant of other people's actions because people may behave differently. What may be considered rude in Spain might be seen as appropriate in the United States. An individual may feel nostalgia for their place of birth or where they grew up, but the best part of living in another country is the enriching experience of living a life that is probably unlike your own. Nostalgia is natural and one should retain his or her native customs to keep them connected to their roots.*

In the domain of accuracy the results were straightforward: all accuracy measures were positively correlated with writing scores in the L2 corpus, indicating that essays with fewer errors received higher scores, and the high scoring learners produced significantly fewer errors than the low scoring learners, with significant differences found for all measures except the percentage of pragmatic errors, which was similar across groups. The L1 participants produced very few errors overall, and thus there were no correlations observed between errors and essay scores in the L1 corpus. These results are in line with previous research showing that learners' errors have a negative impact on holistic ratings of L2 writing (e.g., Engber, 1995), and are coherent based on the criteria of the ESL Composition PROFILE, which explicitly asks raters to consider the frequency and gravity of errors when awarding scores. Among the 5 components, accuracy measures had the strongest correlations with Vocabulary scores (r_2 = -.653) followed by Language Use scores (r_2 = -.583), the two components that ask raters to focus on accuracy.

In the domain of syntactic complexity, we found that the two clause-level measures were positively correlated with writing scores in the full corpus, but not in either the L1 or L2 corpuses independently, suggesting that the L1 writers wrote longer clauses, with more noun phrase modification, than the L2 writers. No correlations were found between sentence length (MLS) and essay scores, while the number of dependent clauses per sentence was negatively correlated with qualitative scores in the full corpus, but not in either subset, suggesting that the L2 writers used significantly more dependent clauses than the L1 writers but that this measure was not relevant for describing differences in L1 or L2 writing quality. When we controlled for essay length (#W), the same relationships remained significant, actually increasing in strength, indicating that the relationship between syntactic complexity and the perceived quality of writing was independent from features related to essay length. When we controlled for sentences, on the other hand, we found that only the clause-based measures (MLC and SYNNP) continued to have significant relationships with scores. That is, when comparing essays that used the same number of sentences, differences in MLS and DC/S disappeared. This may have been due to the fact that writers tended to use very few long sentences, or many short sentences, and this that these variables changed in tandem. In general, the negative correlation between MLS and scores (along with the finding that #S, but not #W, had a significant positive relationship to scores in the L1 corpus) suggested that low-scoring L1 essays used fewer, longer sentences than high-scoring essays, a finding illustrated by our comparison of writing from Participant #39 and Participant #51 (Table 7.6 shows the introductory paragraph of each essay).

Table 7.6 Comparison of introductory paragraphs from L1 essays of similar length with differing numbers of sentences.

Participant #39	Participant #51
Total score: 88.5 *(Rank 5 of 28)*	Total score: 96.5 *(Rank 26.5 of 28)*
Words/sentences in full essay: #W = 258 ; #S = 5	Words/sentences in full essay: #W = 276 ; #S = 13
This is very true, I think nowadays there are far too many people moving or emigrating (as it is referred to), and ignoring the fact that there is a whole different and new culture, language and way of life to learn about. This is not to say that these people should forget their original customs and culture, I just personally feel that they should make more of any effort to try to for example taste local dishes or learn a few phrases in the new language.	*I agree with this statement but I think that this is something that is easier said than done. Upon coming to Spain this past summer, I wasn't entirely sure what to expect in terms of customs and the way people lived their lives from day to day. I had a general idea from what I had studied previously about the country, but nothing prepared me for how hard it would be to adjust.*

These two participants were selected as they wrote similar numbers of words but Participant #51 wrote more than double the number of sentences; however we can also see that Participant #51 was one of the highest scoring native speakers while Participant #39 was one of the lowest scoring. The use of run-on sentences in Participant #39's essay provides a clue as to why longer sentences, though theoretically associated with greater complexity, may have been penalized by raters when assigning scores. We did not consider punctuation errors in the present study but would include this variable in future analysis so that it could be examined alongside MLS with more precision.

Neither measure of syntactic variety was found to have any correlation with qualitative scores when considered independently, but significant relationships were found when we controlled for the number of sentences. Structural overlap (StrutA) was negatively correlated with qualitative scores in the full corpus and in the L2 corpus, suggesting that essays with less structural overlap received higher scores. Tense and aspect repetition (Temp) was negatively correlated with qualitative scores in the full corpus, though not for either the L1 or L2 samples independently, suggesting that while decreased tense and aspect repetition was associated with perceived essay quality, this aspect of variety was primarily relevant for distinguishing between L1 and L2 essays. These findings are supported by data reported in Hinkel (2003), who used a large corpus (N = 1,083) of academic writing to examine the features that characterize L1 and L2 texts. He found that L2 texts were characterized by a handful of features associated with syntactic and lexical simplicity and that, with regards to verbs, the L2 texts relied heavily on the *be*-copula and on the a small set of verbs classified as "public" (e.g. *say, state talk*), "private (e.g. *feel, learn*), or "expecting/tentative" (*like, try, want*). When we compared the high-scoring and low-scoring L2 learners, no significant differences in either complexity or variety were found, which was predictable since this piece of analysis did not control for the total number of sentences produced.

Finally, we looked at 8 measures in the domain of cohesion and 1 measure of personal pronoun use and found that all 9 measures were negatively correlated with essay scores, in line with our predictions, given the personal, non-complex nature of the essay topic and the characteristics of our raters (McNamara, 2001; McNamara & Kintsch, 1996). For two measures of referential overlap (RefP, pronoun overlap, and RefA, argument overlap) there were no significant correlations in either the L2 or L1 corpus independently, suggesting that these aspects of cohesion varied primarily between L1 and L2 writers. Content-word overlap (RefC) was found to have a significant relationship to qualitative scores in the L2 corpus but no relationship to qualitative scores in the L1 corpus. Indeed, in the L1 corpus there was no relationship between referential overlap and qualitative scores, which suggested that these measures may have a stronger relationship with proficiency than with writing ability or that they may not be sufficiently precise to differentiate between skilled and unskilled native speakers. In the L1 corpus, the only measures of cohesion that were found to have a relationship with qualitative scores were the umbrella measure of connectives per word (%Con), and the more specific measure of causal connectives per word (%CausCon). As with the complexity measures, we opted to control for text length and the number of sentences to determine if the relationship between cohesion and essay scores was an artifact of the relationship between cohesion and the amount of text produced. We found that the relationships between perceived quality and referential overlap remained robust. The correlation between the overall frequency of connectives and essay scores decreased from r_2 -.466 to r_2 = -.256 but remained significant at p < .01; however the relationship between additive connectives (%AdCon) and temporal connectives (%TempCon) disappeared, indicating that only the use of causal and logical connectives had a significant impact on scores when essay length was controlled. The relationship between pronoun density and essay scores increased in strengh (from r_2 = -.377 to r_2 = -.423) when the number of words was controlled. Finally, when we compared the high and low scoring L2 learners we found significant differences for argument and content word overlap and for all measures of connectives except for temporal connectives, indicating that high scoring learners used significantly less explicit cohesion in their essays than did low scoring learners.

Our results are similar to those of previous empirical studies considering the same specific dimensions of cohesion (i.e. those measured with Coh-Metrix). For example, in Crossley and McNamara (2010) they used 1200 essays from a large corpus of high school students in Hong Kong (L2 learners of English) to conduct discriminate analysis and discover which of the 600 total Coh-Metrix indices discriminated between essays receiving grades from A to F (they used an internal version of their tool, which included many of the same measures used in the present study as well as a much larger and more specific selection of indices). They found that content word overlap (RefC) was among the 14 best variables for discriminating between essays of different levels of proficiency, and they also reported a negative relationship, suggesting that greater content word overlap was associated with lower essay scores. They also reported that higher proficiency writers used less aspect repetition and fewer logical operators, a category that partially overlaps with our measure of connectives. Like Crossley and McNamara, we also found that essays evaluated as high quality contained fewer explicit markers of cohesion and less referential overlap. Furthermore, the finding that explicit cohesion was also negatively related to scores in the L1 corpus suggested that this negative relationship was not related to improper or incorrect use of cohesive devices (such as that reported in Milton and Tsang, 1993).

Our qualitative analysis also suggested that the use of connectives might be related to other changes, for example in syntactic complexity. For example, if we reconsider the final essay of Participant #5 (see Table 7.1), we find that the additive connective *and* is used 10 times, and that in all but 2 instances it is used to introduce a coordinate clause. Therefore the negative relationship between additive connectives and essay scores may reflect a negative relationship between coordinate clauses and essay scores, which would be in line with our understanding of syntactic complexity and the notion that the use of coordinate clauses decreases with increasing proficiency, as writers learn to express more relationships with subordinate clauses or with phrasal elaboration. The fact that there is no relationship between additive connectives and essay scores among the L1 writers, who are all presumably fully proficient and beyond a phase during which coordination might be used at the expense of other types of complexity, further supports this interpretation; however since we did not include an explicit measure of coordination in our analysis of complexity, believing all of our learners to be of higher proficiency and thus beyond this developmental stage, a definitive understanding of this relationship is beyond the scope of this study and left as a topic of future research.

7.4.2 FLACC and LGP

Next we considered the relationship between FLACC measures and LGP, in response to RQ4b. *Which FLACC measures are significantly correlated with grammar and cloze scores?* and RQ4c. *Which FLACC measures discriminate between learners and native speakers with similar qualitative scores?* First we will discuss the results found for RQ4b, which looked at correlations in the L2 corpus (independently, and when controlling for essay scores).

With regards to fluency, we found that neither measure (#W or #S) correlated significantly with scores on the grammar and cloze test. The finding in relation to #S was not surprising given our previous observations—such as the significant correlation between #S and essay scores in the L1 corpus—since we assumed that the number of sentences (along with MLS) might tell us something about perceived writing quality but was not related to linguistic development. These findings were also consistent with previous research. In Wolfe-Quintero et al.'s (1998) book they review 6 empirical studies that considered the sentence as an index of fluency and report that none found any relationship between the number of sentences and proficiency level (p. 32).

The finding that #W had no relationship to grammar and cloze scores was somewhat more surprising, since we assumed that both would be related to linguistic development; however this finding was in line with the CAF theories proposed by Skehan (1996, 1998), Robinson (2001), and others, and the notion that development may occur unevenly across components or across individuals. Interestingly, when we controlled for essay quality, we observed a significant negative correlation between #W and LGP, suggesting that there was a tendency for learners with greater lexico-grammatical competence to write longer essays and vice versa. This again confirmed CAF theories holding that learners with different psychological profiles will be more likely to focus on one element of the language than another (Ellis & Barkhuizen, 2005). These findings were also supported by our previous observation, in relation to RQ2, that essay scores and LGP were only correlated at $\tau = .478$, and thus that many participants had mixed profiles, with high levels of competence in only one area. If we reconsider the two essays compared in Table 7.3 above, we can see that Participant #6 demonstrates greater fluency and strategic competence, despite having shortcomings in the domains of complexity and accuracy; Participant #27, on the other hand, clearly prioritizes form over meaning and makes an effort to search for sophisticated words and expressions, thus exhibiting less fluency. The correlations observed suggest that this type of tradeoff may have been common in our corpus.

In the domain of accuracy, unsurprisingly, both umbrella measures (%EFS and %TotE) as well as Type 1 spelling errors (%Sp1) were significantly correlated with grammar and cloze scores, as were all three sub-types of errors captured under %TotE (grammatical, lexical, and pragmatic errors). The grammar and cloze tests required participants to produce grammatically accurate responses and thus an attention to accuracy was crucial to high scores on these tests, so it is logical that the learners who were able to perform accurately in this context were also largely more accurate in writing. The fact that type 2 spelling errors (Sp%) did not correlate with grammar and cloze scores further supported our assumption that these errors did not reflect a lack of knowledge but merely lapses in attention due to time pressure. These results are consistent with previous research reported in Wolfe-Quintero et al. (1998), Polio (1997) and elsewhere, which indicate that accuracy may be reliably viewed as a developmental measure and discriminate between learners with different degrees of linguistic proficiency. It appeared that the error count measure (%TotE) was a more powerful measure than %EFS, though both had strong significant correlations. When essay scores were controlled, the relationships between the two umbrella measures and the measure of grammar errors (%Gre) remained strong; however the relationships disappeared for lexical, pragmatic and spelling errors, suggesting that these features were more closely related to writing proficiency than to lexico-grammatical proficiency. The fact that lexical and pragmatic errors were not linked to scores on the grammar and cloze test was somewhat odd; however it may simply be because these errors types were relatively infrequent in our corpus (much more infrequent that grammar errors).

Of the syntactic complexity measures, only mean length of clause (MLC) was found to have a significant relationship to grammar and cloze scores, and this relationship remained when essay scores were held constant, confirming our prediction that MLC is a reliable developmental index, at least for learners with advanced levels of proficiency. Neither MLS nor DC/S had any relationship to lexico-grammatical proficiency, further supporting our findings in the previous sections, which indicated that these two measures were related to punctuation and writing style/perceived quality, but not to grammatical ability or linguistic proficiency per se. These results are supported by a reassessment of the 7 empirical studies reviewed in Wolfe-Quintero et al. (1998). In general, the studies that aimed to relate sentence length to holistic ratings found a relationship, but those that attempted to use sentence length to discriminate between learners across grade levels or over periods of time (i.e. learning contexts). Therefore, while MLS is a potentially useful variable and may tell us something about the quality of the writing used, it should not be considered an index of proficiency more generally.

In the domain of syntactic variety, StrutA did not have any significant relationship to LGP, confirming our predictions that this was a measure of writing quality but not of linguistic proficiency in isolation. Tense and aspect repetition (Temp), on the other hand, did have a significant negative correlation to LGP, suggesting that a greater variety of verb tense and aspect (and hence less repetition) was associated with high grammar and cloze scores. This was predictable given our understanding that less tense and aspect repetition was likely related to an increased grammatical repertoire, and the findings for Temp support our interpretation of the decrease in Temp after the AH learning context (in relation to RQ1b). That is, the improvement seen after the AH context was likely not due to any improvement in composing competence per se, and instead due to an increased syntactic repertoire, cultivated in the grammar-focused EFL classes.

In the domain of cohesion, we found that all three of the referential overlap measures, as well as the umbrella measure of connectives, the measure of temporal connectives, and the measure of pronoun density were negative correlated with LGP; however when we controlled for essay scores the relationships disappeared for all measures except for RefC (content word overlap). This suggested that the use of cohesion was primarily related to writing quality and not to LGP, and confirmed our predictions. That is, the overuse of connectives observed in studies such as Milton and Tsang (1993) suggest that learners do not understand the requirements of good writing in relation to cohesion, but that they have the linguistic repertoire to produce a range of cohesive devices.

When we compared the native speakers and the high scoring learners (in response to RQ4c), we found that there were no differences in terms of either fluency or lexical diversity or sophistication. The latter was somewhat surprising given that native speakers are presumed to have larger and more accessible vocabularies and suggest that lexical diversity and sophistication measures are more related to essay quality than development. This is interesting since both measures are frequently used as indices of linguistic development, and given the previously discussed links with productive vocabulary size. Given the large body of previous studies that have reported evidence of GI and AG1k (or related measures) as developmental indices, our own findings may be taken as evidence that this particular group of L2 learners have achieved very high levels of proficiency in English, and may have vocabulary sizes that are comparable to those of native speakers. This assumption is plausible given the number of years our participants have spent studying English and given that research indicates that the lexicon is one area in which L2 language users may converge with L1 users (Nation, 2001); furthermore as speakers of Latin-derived languages (Spanish and/or Catalan), our learners have access to a large portion of the academic vocabulary and thus have an advantage in the domain of lexical acquisition.

In contrast to fluency and lexical characteristics, all of the accuracy measures except for %Sp2 discriminated between the two groups. The learners produced many more errors per word, across all error classes, and significantly fewer error-free sentences that did the native speakers, despite receiving similar essay scores. In the domain of complexity, both clausal measures (MLC, SYNNP) and the measure of subordination (DC/S) discriminated between the two groups but, as predicted, MLS did not. The native speakers produced more complex clauses (longer clauses with more noun phrase modification) and fewer dependent clauses per sentence, further confirming our previous arguments about the nature of development in complexity at advanced levels of proficiency and indicating that our learners still have a way to go before converging with native speakers and must learn to express complex relationships at the phrasal level, as opposed to through subordination, before their writing will confirm to the norms expected in academic registers. Neither measure of syntactic variety discriminated between the two groups, which was in line with our understanding of these two measures and the previous observation (in relation to RQ1b) that Temp had a clear ceiling effect.

There were no differences between the high scoring learners and native speakers in terms of connectives or referential overlap; however there was a significant difference in the pronoun density measure, showing that the learners used significantly more personal pronouns than their native-speaking counterparts. All of these findings confirmed our predictions based on previous observations and based on previously reported research with the exception of pronoun density (DenPr). That is, we expected that DenPr would be related to writing quality/writing proficiency and not to LGP, which was confirmed by our finding that the significant relationship in the L2 corpus disappeared when essay scores were held constant. It was interesting that learners with presumably similar levels of lexico-grammatical competence and who wrote essays that were judged to be of similar quality as those of native speakers, continued to use significantly more personal pronouns in their writing. It may be that the L2 writers used more first and second person pronouns but that this usage did not detract from the perceived quality because it was done effectively (as in the second essay written by Participant #5, seen above in Table 7.1). The imperative to avoid second-person pronouns (or not) may simply reflect differences in the writing instruction received by the L1 and L2 participants in their respective contexts (that is, English-speakers are usually taught to avoid second-person pronouns in formal writing, but perhaps our L2 writers do not receive this instruction). This is purely speculative, and analysis of first and second person usage was beyond the scope of the present study, but is one possible interpretation for the patterns observed. Overall it appears that when writing ability is held constant, the differences between learners and native speakers, who inherently differ in their linguistic proficiency, are primarily explained through differences in accuracy and complexity.

In sum, in response to RQ4, we found that a number of the measures used were relevant for discriminating between essays of different perceived quality, and that others were relevant for discriminating between writers with different levels of lexico-grammatical proficiency, and that others overlapped. It is worth pointing out that because so many measures overlap, it is important to reevaluate previous studies of writing ability which claimed to be evaluating one measure but may have been inadvertently measuring a larger combination of constructs.

Chapter 8

Conclusions and Future Research

In this dissertation we have evaluated the development of writing skills in a 3-month study abroad period and found that participants' writing improves significantly in terms of perceived quality and in terms of a variety of linguistic characteristics associated with proficiency in writing. Using a combination of qualitative and quantitative evaluation techniques, we have shown that after the SA learning context, participants receive higher evaluations on an analytic scale, they make measurable progress in the domains of complexity, accuracy, and fluency (CAF) and lexical diversity, and they show a more appropriate and native-like use of cohesion. These findings are an important contribution to the body of SA research given that previous studies claiming a lack of writing progress are widely cited in the literature (i.e., Meara, 1994; Freed, So, & Lazar, 2003).

In order to determine whether improvement in the SA context was due to the specific nature of this learning context, (i.e., to the massive exposure to input in the target language that learners presumably receive while abroad), we also evaluated improvement in writing after a previous period of classroom instruction at the home institution (AH) and compared progress in the two contexts. In contrast to the improvement seen during the SA context, very few significant changes occurred during the AH context, despite the fact that it was twice as long (6-months). The learners' writing was not perceived as being of higher quality, and there were very few significant changes in quantitative measures: the only domain in which we observed improvement was that of syntactic variety, where we saw a decrease in tense and aspect repetition after the AH context. This was in line with previous research indicating that traditional EFL classrooms are ineffective when it comes to cultivating writing skills, and that only writing-intensive courses and extensive practice are likely to have a measurable impact. Overall, the comparison of the changes that occurred in the two contexts made it clear that it was the SA experience, and not simply the natural course of development, that led to the observed improvement in our participants' writing.

In our analysis of longitudinal improvement we also explored the influence of 'initial level', as measured by both writing ability and lexico-grammatical proficiency at the beginning of the study. As in previous research, we found that the learners' initial levels of proficiency had an impact on the amount of progress they made. All of our learners had advanced proficiency prior to the onset of the study; however we found that the learners who were relatively more advanced improved more than their peers. Thus our study provides further support for the arguments of Brecht et al. (1991), who measured SA gains in reading, speaking, and listening and concluded that "communication skills are most effectively built upon a solid grammar/reading base" (Brecht, et al., 1991, p. 16). Given these findings, we must assume that our conclusions about the benefits of SA on writing should be generalized only to learners who embark on the SA with upper-intermediate to advanced levels of proficiency. While this is more common in the European context, particularly for learners of English, it is less common in the US and may be one of the reasons why fewer gains have been reported in this context.

Both before and after the SA, we compared learners' writing to that produced by native speakers of English with similar educational backgrounds. We found that the learners were already 'native-like' on a several quantitative measures and that they became more native-like over the course of the study. Although the learners' essays still received lower qualitative scores and exhibited differences in most domains (especially accuracy), they converged in the domains of fluency, lexical diversity, and cohesion after the SA context. In addition to highlighting the advanced proficiency of our participants and the benefits of the SA, gathering baseline data from the native speakers allowed us to scrutinize our selection of quantitative measures and ensure that we were interpreting progress accurately. For example, for the specific writing topic and task assigned in our study, explicit cohesion was associated with less proficient, non-native writing. If we had assigned a more complex topic, or one that drew on domain-specific knowledge outside our readers' expertise, explicit cohesion might have been evaluated more positively. Thus it was useful to evaluate the learners' writing in comparison to the native speakers and to confirm that the decrease in cohesion over time brought them slowly closer to these norms. The same benefit was observed for quantitative measures such as MLS or Temp, which had clear ceiling effects. For Temp, for example, because the learners converged with native speakers after the AH context we were able to see that the lack of further progress in the SA context was not evidence that this context did not benefit syntactic variety, but evidence that this measure was not sufficiently descriptive to evaluate continued improvement at more advanced levels.

We were also able to improve our understanding of quantitative measures by exploration in relation to our final research question, where we aimed to identify the characteristics associated with writing quality and lexico-grammatical proficiency. We found, for example, that fluency measures were more closely associated with writing quality than with linguistic proficiency, which is coherent with the findings that fluency tends to vary depending on genre, register, or even topic, but is important to keep in mind given the role of fluency in the CAF triad. We also found that both sentence length (MLS) and subordination (DC/S) were negatively correlated with essay scores, suggesting that any CAF study which uses these measures as indices of development must be careful in interpreting results. In the case of DC/S, for example, at beginning levels of proficiency increases in subordination might be associated with linguistic progress and simultaneously associated with clumsier and less coherent writing, which presents something of a conundrum for learners who have high levels of composing competence (cultivated in L1) but low levels of linguistic proficiency. When essay quality is held constant, L1 and L2 writers seem to differ primarily in the domains of accuracy and clausal complexity; that is, in order to achieve native-like writing competence, advanced learners must work to improve their accuracy and to increase their use of phrasal elaboration and nominalization. Given that the SA learning context began to push the learners in the appropriate direction in both domains, our exploration of the measures further highlighted the benefits of this learning context; additionally, an awareness of the most persistent areas of differences between high-level learners and native speakers may improve instruction and allow us to focus attention on these issues in EFL classrooms at the university level.

While conducting and completing the empirical study reported in this dissertation, we noted a handful of issues and potential limitations that were left aside to be addressed in future research. As is typical, the number and range of questions we could ask were limited by the data available, by the institutional context, and of course by practical considerations of time and cost. For example, in the present study we had no information about the writing process, about participants' L1 writing abilities or their meta-cognitive knowledge, all of which might have enriched our intepretations of the observed changes in their writing products. Therefore, we may speculate that the increased proficiency acquired during the SA context allowed participants to dedicate more time and attention to higher-order features of content and style, or to spend more time on problems of 'upgrading', in line with previous process-oriented research (Manchón, 2009); however without think-aloud protocols or follow-up interviews at our disposal, we cannot prove that this explanation is the best one. On the other hand, writing process studies, though extremely valuable, are often limited by the time-consuming nature of data analysis; this leads them to focus on a small number of subjects, which makes it difficult to generalize (Krapels, 1990). The goal of the SALA project was to gather basic data on each of the major EFL skills from as large a number of participants as possible, so that generalizations could be made about the larger population of ERASMUS learners. For the study of writing we aimed to choose a robust sample and to conduct a careful and methodical examination of the available data; thus, although we can only speculate about the writing process, we may generalize that SA leads to improved writing products and measurable progress in many domains associated with proficiency.

One of the limitations of the design of the study is that is does not allow us to consider the possibility that the two learning contexts might have had different effects if the order had been reversed. That is, it may be that our participants increased their meta-cognitive knowledge about writing while studying at home (qualitative analysis suggested at some participants improved basic paragraphing skills, for example), but that they needed to further increase their proficiency and free up cognitive resources in order to put this knowledge into practice. The increase in proficiency in the SA context thus would have allowed them to demonstrate skills acquired in the previous AH context. Although it was impossible given the institutional context, a research design in which half the students completed the two learning contexts in the reverse order (first SA, then AH), would have allowed us to evaluate this possibility, and the impact of each context. That is, we would have been able to see whether improvement in SA was as pronounced without the prior EFL classroom study, and whether the EFL classroom study would have been more beneficial after the SA.

On the other hand, although we isolated this particular period of formal instruction, it is safe to assume that the majority of SA participants have had a significant amount of classroom instruction prior to embarking on their SA, particularly in the case of EFL students in Europe. Our own participants had received an average of 10-years of formal instruction prior to beginning at the university, and many of them had also taken privately funded extracurricular language classes during this time. That is, we assume that the advanced proficiency the participants had obtained prior to embarking on the SA were not the result of any single period of prior formal instruction, but the result of the cumulated formal instruction received over the many years of formal education, in both L1 and L2. While reversing the order of the learning contexts would have allowed us to see whether the SA period magnified the benefits of later formal instruction—perhaps giving the learners greater motivation for language learning—it would have given us little novel information about the SA context, which was the primary aim of this study.

Although beyond the scope of this dissertation, the SALA project did collect data from participants at "T4", a period one year after the SA and after an additional 6-months of formal instruction. In a pilot study looking at a small selection of quantitative measures, we found that although the gains made during the SA period are maintained in the longer term, there are no further gains, suggesting that the second AH period is no more beneficial to participants' writing skills than the first (Perez-Vidal & Barquin, in press). Again, while classroom study serves to improve many aspects of learners' proficiency—in the case of the SALA participants, the AH classes were designed to improve students' abilities to analyze and dissect the language, and to provide them with tools they would use as translators and interpreters—it is reasonably well documented that without intensive, process-writing instruction classroom learning contexts are unlikely to lead to dramatic improvement.

Future research would address a number of methodological issues noted while exploring the quantitative measures associated with writing quality and language proficiency. An underlying methodological goal that ran parallel to this thesis was to identify the best computational tools and measures for analyzing proficiency in EFL writing, so that following this study we might analyze significantly larger corpora and consider development in more diverse contexts and conditions. While we found that the available tools are increasingly powerful and take us much of the way towards fully automatic analysis, some measures and domains are underdeveloped. In the domain of syntactic variety, for example, we found that the Coh-Metrix indices that best approximated this construct (those which informed us about overlap in structure and content across adjacent sentences) had ceiling effects and/or were too broad to give us a satisfying account of progress in this domain. Furthermore, despite the clear association between syntactic variety and writing quality, and the recognized importance of this characteristic on writing rubrics, there are currently no standard measures for manual analysis, which precludes the development of accurate computational measures. One interesting and practical area of future research would thus be to work on an index of syntactic variety that captures degrees of variety (for example, in the way that Huw Bell's (2008) system assigns values to different degrees of syntactic complexity) and to validate this measure with reliable measures of writing quality.

Finally, there were a number of additional data sources available through SALA that might have been interesting to explore in relation to our own findings, such as perceived satisfaction. That is, we observed in our discussion that most participants did not feel their writing had improved much, in comparison to other skills (in line with the findings reported in Meara, 1994). While this confirmed previous observations that self-report data is unreliable (DeKeyser, 2007), it would be interesting to explore the questionnaire data in more detail and determine if there was a correlation between perceived improvement and actual improvement. That is, it would be interesting to see if those students who made substantial improvement had more positive outlooks than their peers who did not improve or improved only minimally.

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alderson, J. C., & Crawshaw, R. (1990). Language needs and language preparedness of ERASMUS students. *Unpublished paper, University of Lancaster*.

Allen, H., & Herron, C. (2003). A mixed-methodology investigation of the linguistic and affective outcomes of summer study abroad. *Foreign Language Annals, 36*(3), 370-385.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (2nd release)[cd-rom]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. (1986). *Working Memory*. London: Oxford University Press.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390-395.

Bardovi-Harlig, K., & Bofman, T. (1988, March). *A Second Look at T-Unit Analysis.* Paper presented at the 22nd Annual TESOL Conference, Chicago, IL.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*(01), 17-34.

Barkaoui, K. (2007). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *Canadian Modern Language Review, 64*, 97-132.

Beacco, J., & Byram, M. (2007). *From Linguistic Diversity to Plurilingual Education: Guide for the Development of Language Education Policies in Europe (main version)*. Strasbourg: Council of Europe: Language Policy Division.

Becker, A. (2006). A review of writing model research based on cognitive processes. In A. Horning & A. Becker (Eds.), *Revision: History, theory, and practice* (pp. 25-48). West Lafayette: Parlor Press.

Bell, H. (2008, September). *Measuring the Syntactic Complexity of Embedded Clauses.* Paper presented at the Annual Conference of the British Association of Applied Linguistics (BAAL).

Bereiter, C., Burtis, P., & Scardamalia, M. (1988). Cognitive operations in constructing main points in written composition. *Journal of memory and language, 27*(3), 261-278.

Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition*. Hillsdale, NJ: Lawrence Erlbaum.

Bestgen, Y., & Granger, S. (2011). Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(2), 235-252.

Biber, D. (1988). *Variation across speech and writing*: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). Amsterdam/Philadelphia: John Benjamins.

Biber, D. (2012, March). *Using Multidimensional Analysis to Investigate Cross-linguistic Patterns of Register Variation.* Paper presented at the Georgetown University Roundtable on Language and Linguistics (GURT), Georgetown.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. New York: Wiley.

Brecht, R., & Davidson, D. (1991, March). *Language acquisition gains in study abroad: Program assessment and modification.* Paper presented at the NFLC Conference on Language Testing, Washington D.C.

Brecht, R., Davidson, D., & Ginsberg, R. (1995). Predictors of foreign language gain during study abroad. In B. Freed (Ed.), *Second Language Acquisition in a Study Abroad Context* (pp. 37-66). Amsterdam/Philadelphia: John Benjamins.

Brown, J. (1999). Standard error vs. Standard error of measurement. *Shiken JALT Testing & Evaluation SIG Newsletter, 3*(1), 20-25.

Brown, N. A., Solovieva, R. V., & Eggett, D. L. (2011). Qualitative and Quantitative Measures of Second Language Writing: Potential Outcomes of Informal Target Language Learning Abroad. *Foreign Language Annals, 44*(1), 105-121.

Bulté, B., & Housen, A. (2012). Defining and Operationalizing L2 Complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency -- Investigating Complexity, Accuracy, and Fluency in SLA*. Amsterdam/Philadelphia: John Benjamins.

Carroll, J. B. (1967). Foreign Language Proficiency Levels Attained by Language Majors Near Graduation from College. *Foreign Language Annals, 1*(2), 131-151.

Carson, J. (2001). Second Language Writing and Second Language Acquisition. In T. J. Silva & P. K. Matsuda (Eds.), *On Second Language Writing*. Mahwah, New Jersey: L. Erlbaum Associates.

Cenoz, J., & Jessner, U. (2000). *English in Europe: The acquisition of a third language* (Vol. 19). Clevendon: Multilingual Matters.

Chenoweth, N., & Hayes, J. R. (2001). Fluency in Writing: Generating Text in L1 and L2. *Written Communication, 18*(1), 80-98.

Coleman, J. A. (1996). *Studying languages: a survey of British and European students: the proficiency, background, attitudes and motivations of students of foreign languages in the United Kingdom and Europe*. London: CILT.

Coleman, J. A. (1998). Language learning and study abroad: The European perspective. *Frontiers, 4*(1), 167-203.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition, 26*(02), 227-248.

Collentine, J. (2009). Study abroad research: Findings, implications, and future directions. In M. H. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 218-233).

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*, 99-115.

Cooper, G., & Hamp-Lyons, L. (1988). *Looking in on essay readers*. Ann Arbor: University of Michigan English Composition Board.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-136.

Crystal, D. (1997). *English as a global language* (Vol. 2): Cambridge University Press.

Cumming, A. (1989). Writing Expertise and Second Language Proficiency. *Language learning, 39*(1), 81-135.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*, 67-96.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19*, 121-129.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics, 24*(2), 197-222.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. F. (1999). *Dictionary of Language Testing* (Vol. 7): Cambridge University Press.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill*. Princeton, NJ: ETS.

DeKeyser, R. (1990). From learning to acquisition? Monitoring in the classroom and abroad. *Hispania, 73*(1), 238-247.

DeKeyser, R. (1991a). The semester overseas: What difference does it make? *ADFL BULLETIN, 22*(2), 42-48.

DeKeyser, R. (1991b). Foreign language development during a semester abroad. In B. Freed (Ed.), *Foreign Language Acquisition Research and the Classroom*. Lexington, MA: D.C. Heath.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*: Cambridge University Press.

Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in English language test design and delivery*. Sydney, Australia: National Center for English Language Teaching and Research, Macquarie University.

Dewey, D. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers, 15*, 127-148.

Díaz-Campos, M. (2004). Context of Learning in the Acquisition of Spanish Second Language Phonology. *Studies in Second Language Acquisition, 26*(02), 249-273.

Dyson, P. (1988). *The Effect on Linguistic Competence of the Year Spent Abroad by Students Studying French, German and Spanish at Degree Level*. Oxford: Oxford University Language Teaching Centre.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*: Oxford University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

European Commission (2012). Erasmus -- Facts, Figures & Trend. The European Union support for student and staff exchanges and university cooperation in 2010-11. Luxembourg: Publications Office of the European Union.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Field, A. (2005). *Discovering statistics using SPSS* (2 ed.). London: Sage.

Flower, L. (1979). Writer-based prose: A cognitive basis for problems in writing. *College English, 41*(1), 19-37.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication, 32*(4), 365-387.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299-323.

Fotos, S. S. (1991). The Cloze Test as an Integrative Measure of EFL Proficiency: A Substitute for Essays on College Entrance Examinations?*. *Language learning, 41*(3), 313-336.

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English*: Educational Testing Service.

Freed, B. (1995a). *Second Language Acquisition in a Study Abroad Context*. Amsterdam/Philadelphia: John Benjamins.

Freed, B. (1995b). What makes us think that students who study abroad become fluent. In B. Freed (Ed.), *Second Language Acquisition in a Study Abroad Context* (pp. 123-148). Amsterdam/Philadelphia: John Benjamins.

Freed, B. (1998). An overview of issues and research in language learning in a study abroad setting. *Frontiers, 4*, 31-60.

Freed, B., Dewey, D., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition, 26*(02), 349-356.

Freed, B., So, S., & Lazar, N. (2003). Language Learning Abroad: How Do Gains in Written Fluency Compare with Gains in Oral Fluency in French as a Second Language? *ADFL BULLETIN, 34*(3), 34-40.

Fulwiler, T., & Young, A. (1982). *Language Connections: Writing and Reading across the Curriculum*. Urbana, IL: National Council of Teachers of English.

Galbraith, D. (2009). Cognitive models of writing. *German as a foreign language, 2-3,* 7-22.

Garson, G. (2010). Reliability Analysis. *Statistical Associates Blue Book Series* Retrieved September, 2011, from http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm

Ginsberg, R. B. (1992). *Language Gains During Study Abroad: An Analysis of the ACTR Data. National Foreign Language Center Working Papers*. Washington D.C.: Johns Hopkins University / National Foreign Language Center.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*: Longman New York.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, 36*(2), 193-202.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*(1), 17-27.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123-145.

Green, T. (2004). Making the grade: score gains on the IELTS Writing test. *Research Notes, 16, University of Cambridge ESOL examinations*, 9–13.

Groot, P. (1990). Language testing in research and education: The need for standards. *AILA Review, 7*, 9-23.

Halliday, M., & Matthiessen, C. (1999). *Construing Experience Through Meaning: A Language-Based Approach to Cognition*. New York: Cassell.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tubingen, Germany: Gunter Narr Verlag.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom*: Cambridge University Press.

Hamp-Lyons, L. (2001). Fourth Generation Writing Assessment. In T. J. Silva & P. K. Matsuda (Eds.), *On Second Language Writing* (pp. 117-125).

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189).

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Haswell, R. (2005). Researching Teacher Evaluation of Second Language Writing. In P. K. Matsuda & T. Silva (Eds.), *Second Language Writing Research: Perspectives on the Process of Knowledge Construction* (pp. 105-120). Mahwah, NJ: Erlbaum.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Levy & S. E. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Erlbaum.

Hayes, J. R. (2006). New directions in writing theory. In C. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 28-40). New York: Guilford.

Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J. F., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Reading, writing, and language processing* (Vol. 2, pp. 176-240): Cambridge University Press.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive Processes in Writing: An Interdisciplinary Approach* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY [Computer software]. Wellington, New Zealand: Victoria University of Wellington (Available from http://www.vuw.ac.nz/lals).

Hirose, K., & Sasaki, M. (1994). Explanatory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing, 3*(3), 203-229.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics, 30*(4), 461.

Howard, M. (2001). The effects of study abroad on the L2 learner's structural skills: Evidence from advanced learners of French. *Eurosla Yearbook, 1*(1), 123-141.

Hunt, K. W. (1965). A synopsis of clause-to-sentence length factors. *The English Journal, 54*(4), 300-309.

Hunt, K. W. (1970). Recent measures in syntactic development. In M. Lester (Ed.), *Readings in applied transformation grammar*. New York: Holt, Rinehert, and Winston.

Ife, A., Vives-Boix, G., & Meara, P. (2000). The impact of study abroad on the vocabulary development of different proficiency groups. *Spanish Applied Linguistics, 4*(1), 55-84.

Isabelli, C. A., & Nishida, C. (2005). Development of the Spanish subjunctive in a nine-month study-abroad setting. In D. Eddington (Ed.), *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages* (pp. 78-91). Somerville, MA: Cascadilla Proceedings Project.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*: Newbury House Rowley, MA.

James, C. (1998). *Errors in language learning and use*: Longman.

Johns, A. M. (1990). L1 composition theories: implications for developing theories of L2 composition. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 37-56): Cambridge University Press.

Jones, S., & Tetroe, J. (1987). Composing in a second language. In A. Matsuhashi (Ed.), *Writing in real time: modelling production processes* (pp. 34-57). Norwood, NJ: Ablex.

Juan, M., Prieto, J. I., & Salazar, J. (2007, April). *The effect of formal instruction context on the lexico-grammatical development of advanced learners of L2 English.* Paper presented at the XXV Congreso de AESLA, Murcia, Spain.

Juan-Garau, M., & Pérez-Vidal, C. (2007). The effect of context and contact on oral performance in students who go on a stay abroad. *VIAL, Vigo international journal of applied linguistics*(4), 117.

Kellogg, R. T. (1996). A model of working memory in writing. In C. Levy & S. E. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57-71). Mahwah, NJ: Erlbaum.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research*. New York: Palgrave Macmillan.

Knoch, U. (2009). *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. Frankfurt: Peter Lang.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural and rhetorical pattern and readers' background. *Language Learning, 46*, 397-437.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*, 81-112.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 37-56). Cambridge University Press.

Krashen, S. D. (1985). *The input hypothesis: issues and implications*. London: Longman.

Kroll, B. (1990). *Second Language Writing: Research Insights for the Classroom*: Cambridge University Press.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-242.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*(1), 48-60.

Lafford, B. (1995). Getting Into, Through and Out of a Situation: A Comparison of Communicative Strategies Used by Students Studying Spanish Abroad and 'At Home'. In B. Freed (Ed.), *Second Language Acquisition in a Study Abroad Context* (pp. 97-121). Amsterdam/Philadelphia: John Benjamins.

Lafford, B. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition, 26*(02), 201-225.

Larsen-Freeman, D. (2009). Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied linguistics, 30*(4), 579-589.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics, 16*(3), 307.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning, 40*(3), 387-417.

Levy, C., & Ransdell, S. E. (1996). *The science of writing: Theories, methods, individual differences, and applications*. Mahwah, NJ: Erlbaum.

Levy, R., & Andrew, G. (2006). *Tregex and Tsurgeon: tools for querying and manipulating tree data structures.* Paper presented at the 5th International Conference on Language Resources and Evaluation (LREC 2006).

Llanes, À. (2011). The many faces of study abroad: an update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism*(1), 1-27.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly, 45*(1), 36-62.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (Vol. 3). Mahwah, NJ: Erlbaum.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language. Papers from the Annual Meeting of the BAAL* (pp. 58-71). Clevendon: Multilingual Matters.

Malvern, D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*: Palgrave Macmillan New York.

Manchón, R. M. (2009). *Writing in Foreign Language Contexts: Learning, Teaching, and Research*. Bristol: Multilingual Matters.

Manchón, R. M., & De Larios, J. R. (2007). On the temporal nature of planning in L1 and L2 composing. *Language learning, 57*(4), 549-593.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 55*(1), 51.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction, 14*(1), 1-43.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes, 22*(3), 247-288.

McNamara, T. F. (1996). *Measuring Second Language Performance*. New York: Longman.

Meara, P. (1994). The year abroad and its effects. *Language Learning Journal, 10*(1), 32-38.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect. A Journal of Australian TESOL, 16*(3), 5-19.

Meara, P., & Miralpeix, I. (2004). D_Tools. *Swansea: Lognostics (Centre for Applied Language Studies, University of Wales Swansea).*

Meara, P. M., & Alcoy, J. C. O. (2010). Words as species: an alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language, 22*(1), 222-236.

Meisel, J., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition, 3*, 109-135.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL. Institut voor Togepaste Linguistik*(107-08), 17-34.

Milton, J. C. P., & Tsang, E. S. (1993). *A corpus-based study of logical connectors in EFL students' writing: directions for future research*. Studies in lexis. Proceedings of a seminar on lexis organized by the Language Centre of the HKUST, Hong Kong , 6-7 July 1992. http://hdl.handle.net/1783.1/1083

Minke, A. (1997, January). *Conducting repeated measures analyses: Experimental design considerations.* Paper presented at the annual meeting of the Southwest Educational Research Association, Austin.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N. R., & Wright, M. J. (2010). Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics, 15*(4), 429-473.

Mora, J. C. (2008). Learning Context Effects on the Acquisition of a Second Language Phonology. In C. Pérez-Vidal, M. Juan-Garau & A. Bel-Gaya (Eds.), *A Portrait of the Young in the New Multilingual Spain* (pp. 241-263). Clevendon: Multilingual Matters.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral Fluency, Accuracy and Complexity in Formal Instruction and Study Abroad Learning Contexts. *TESOL Quarterly.*

Murphy, L., & Roca de Larios, J. (2010). Searching for words: One strategic use of the mother tongue by advanced Spanish EFL writers. *Journal of Second Language Writing, 19*(2), 61-81.

Nation, I. S. P. (2001). *Learning vocabulary in another language*: Cambridge University Press.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics, 30*(4), 555.

O Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*(4), 557.

O'Loughlin, K., & Wigglesworth, G. (2003). Task design in IELTS academic writing Task 1: The effect of quantity and manner of presentation of information on candidate writing. *IELTS Research Reports, 4*, 89-129.

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College level L2 Writing. *Applied linguistics, 24*(4), 492.

Pérez-Vidal, C. & Barquin, E. (forthcoming) Measuring longitudinal progress in academic writing. In Pérez-Vidal, C. (ed) (forthcoming) *Study Abroad and formal instruction: Context effects on language learning*. Amsterdam/Philadelphia: John Benjamins.

Pérez-Vidal, C., & Juan-Garau, M. (2009). The effect of Study Abroad (SA) on written performance. *Eurosla Yearbook, 9*(1), 269-295.

Pérez-Vidal, C., Trenchs, M., Juan-Garau, M., & Mora, J. C. (2007, April). *El factor 'Estancia en el país de la lengua meta' en la adquisición de una lengua extranjera (inglés) a corto y medio plazo: Objetivos y metodología del proyecto S.A.L.A.* Paper presented at the XXIV Congreso Internacional de la Asociación Española de Lingüística Aplicada, Madrid (UNED).

Pineda Herrero, P., Moreno Andrés, M. V., & Belvis Pons, E. (2008). La movilidad de los universitarios en España: estudio sobre la participación en los programas Erasmus y Sicue. *Revista de educación*(346), 363-399.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. Language learning, 47(1), 101-143.

Polio, C. (2001). Research Methodology in Second Language Writing Research: The Case of Text-Based Studies. In T. J. Silva & P. K. Matsuda (Eds.), *On Second Language Writing* (pp. 91–115). Mahwah, NJ.

Polio, C. (2003). Research on second language writing: An overview of what we investigate and how. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 35-66). Cambridge University Press

Polio, C., & Williams, J. (2009). Teaching and Testing Writing *The handbook of language teaching* (pp. 486-517): Wiley-Blackwell.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays. *Journal of Educational Measurement*, 220-233.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach* (Vol. 8): Cambridge University Press.

Purves, A. C. (1992). *The IEA Study of Written Composition: Education and performance in fourteen countries* (Vol. 2): International Association for the Evaluation of Educational Achievement. Pergamon Press.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19*(2), 229-258.

Reid, J. (1990). Responding to different topic types. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 191-210): Cambridge University Press.

Rivers, W. P. (1998). Is being there enough? The effects of homestay placements on language gain during study abroad. *Foreign Language Annals, 31*(4), 492-500.

Rivers, W. P., & Golonka, E. M. (2009). Third Language Acquisition Theory and Practice. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 250-266): Wiley-Blackwell.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics, 22*(1), 27.

Roca de Larios, J., Manchón, R. M., & Murphy, L. (2006). Generating text in native and foreign language writing: A temporal analysis of problem-solving formulation processes. *The Modern Language Journal, 90*(1), 100-114.

Roca de Larios, J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language learning, 51*(3), 497-538.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1*.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing, 9*(3), 259-291.

Sasaki, M. (2004). A multiple data analysis of the 3.5 year development of EFL student writers. *Language learning, 54*(3), 525-582.

Sasaki, M. (2007). Effects of Study-Abroad Experiences on EFL Writers: A Multiple-Data Analysis. *Modern Language Journal, 91*(4), 19.

Sasaki, M. (2009). Changes in English as a Foreign Language Students' Writing Over 3.5 Years: A Sociocognitive Account. In R. Manchón (Ed.), *Writing in Foreign Language Contexts: Learning, Teaching, and Research*. Bristol: Multilingual Matters.

Sasaki, M. (2011). Effects of Varying Lengths of Study-Abroad Experiences on Japanese EFL Students' L2 Writing Ability and Motivation: A Longitudinal Study. *TESOL Quarterly, 45*(1), 81-105.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language learning, 46*(1), 137-168.

Schleppegrell, M. J. (1996). Conjunction in spoken English and ESL writing. *Applied linguistics, 17*(3), 271-285.

Schoonen, R., Gelderen, A., Glopper, K., Hulstijn, J., Simis, A., Snellings, P., et al. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language learning, 53*(1), 165-202.

Schoonen, R., Snellings, P., Stevenson, M., & Gelderen, A. (2009). Towards a Blueprint of the Foreign Language Writer: The Linguistic and Cognitive Demands of Foreign Language Writing. In R. Manchón (Ed.), *Writing in Foreign Language Contexts: Learning, Teaching, and Research*. Bristol: Multilingual Matters.

Segalowitz, N. (2011, September). *Cognitive mechanisms underlying fluency.* Paper presented at the 2nd Barcelona Summer School on Bilingualism, Barcelona.

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition. *Studies in Second Language Acquisition, 26*(2), 173-199.

Shaw, P., & TING-KUN LIU, E. (1998). What develops in the development of second-language writing? *Applied linguistics, 19*(2), 225.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 22*, 303-325.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Shuttleworth, M. (2009). Repeated Measures Design Retrieved October 20, 2011, from http://www.experiment-resources.com/repeated-measures-design.html

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*(4), 657-677.

Skehan, P. (1996). Second language acquisition research and task-based instruction. In J. Willis & D. Willis (Eds.), *The Challenge and Change in Language Teaching*. Oxford: Heinemann.

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics, 18*, 268-286.

Song, C. B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

Sperling, M. (1996). Revisiting the writing-speaking connection: Challenges for research on writing and writing instruction. *Review of Educational Research, 66*(1), 53-86.

Sperling, M., & Freedman, S. W. (2001). Research on Writing. In V. Richardson (Ed.), *Handbook of research on teaching* (Vol. 4, pp. 370-389). Washington D.C.: American Educational Research Association.

Stevenson, M. (2005). *Reading and writing in a foreign language. A comparison of conceptual and linguistic processes in Dutch and English*. Amsterdam: SCO-Kohnstamm Institut, University of Amsterdam.

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15*(3), 201-233.

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*(2), 123-143.

Teichler, U. (1997). *The ERASMUS experience: Major findings of the ERASMUS evaluation research project*. Luxembourg: Office for Official Publications of the European communities.

Trenchs-Parera, M. (2009). Effects of Formal Instruction and a Stay Abroad on the Acquisition of Native-Like Oral Fluency. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 65*(3), 365-393.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67.

Vähäpassi, A. (1982). On the specification of the domain of school writing. In A. C. Purves & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (pp. 265-289). Oxford: Pergamon.

Valls-Ferrer, M. (2010). *Language acquisition during a stay abroad period following formal instruction: temporal effects on oral fluency development.* Unpublished Master's Thesis, University Pompeu Fabra, Barcelona.

Valls-Ferrer, M. (2011). *The Development of Oral Fluency and Rhythm during a Study Abroad Period.* Unpublished doctoral dissertation, University Pompeu Fabra, Barcelona.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing writing, 6*, 145-178.

Weigle, S. C. (2002). *Assessing writing*: Cambridge University Press.

White, E. M. (1985). *Teaching and assessing writing*: Jossey-Bass San Francisco.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*: Univ of Hawaii Pr.

WordNet (2010). About Word Net. from Princeton University: http://wordnet.princeton.edu

Yang, W., Lu, X., & Weigle, S. C. (2012, March). *Syntactic Complexity of ESL Writing, Writing Performance, and the Role of Topic.* Paper presented at the Georgetown University Roundtable on Language and Linguistics (GURT), Georgetown.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied linguistics, 31*(2), 236.

Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly, 17*(2), 165-188.

Appendices

Appendix 1. List of data sets (transcribed essays) and changes made

Set	Changes made	Analysis	Software
1	None	Overall Quality	n/a
2	Spelling errors corrected; Non-words eliminated	Fluency & Lexical Characteristics	AntWordProfiler

3	Spelling & non-words corrected;
	Paragraphing standardized;
	Subject omission corrected
		Syntactic complexity & Cohesion	L2SCA & Coh-Metrix

4	Formatting and punctuation modified to CHAT conventions	Accuracy	CLAN

Appendix 2. Log of spelling errors and non-words in essays

1. Irregular usage that was corrected/standardized (not counted as errors)

FILE	USAGE IN TEXT	CHANGE MADE
IZGA01	6h	6
MACI01	every-day	everyday
QUES01	time table	timetable
CRGA02	open minded	open-minded
IZGA02	Catalunya	Catalonia
LELE02	openminded	open-minded
GAME03	open minded	open-minded
GUMA03	open minded	open-minded
LELE03	open minded	open-minded

2. Spelling Errors

T1 - FILE	SPELLING IN TEXT	CORRECTION MADE
ARDE01	exemple rabadam	example ramadan
BEGI01	costumes corea countrie's live (line 8) it's (line 5) shocking	customs korea country's life its shaking
BEPA01	weeker like (line 2) leaved	weaker life lived
BORO01	themselves thing (line 4)	themselves think
CAAM01	advantatges disadvantatges intance combained	advantages disadvantages instance combined
CATA01	asimilated	assimilated
ESAM01	habbit	habit
GORO01	el ramadan afirmation	ramadan affirmation
GUMA01	behabe	behave
LELE01	foreing	foreign
MACI01	allways	always
PAAR01	simbol were	symbol where ("a country where women")

PABE01	obligued analize	obliged analyze
QUES01	baicon adquire custums tradicions tee	bacon acquire customs traditions tea
REGA01	entirily a side it's	entirely aside ("to put roots aside") its ("its values and rules")
RIRA01	althoug atitudes	although attitudes
RUSH01	analize	analyze
VAMO01	inconditional	unconditional
VIBO01	belive live (line 9)	believe life
VIEM01	costums arribing incompetible simplier	customs arriving incompatible simpler
VIGO01	phisical routin costums	physical routine customs
T2 - FILE	**SPELLING IN TEXT**	**CORRECTION MADE**
BEGI02	althought	although
BLPA02	Tent abundancy	Tend abundance
BORO02	Afterall In	after all I (line 2)
CAAM02	custums live (line 11)	customs life
CATA02	xenophoby	xenophobia
GORO02	remaind	remain
GUMA02	addapt unconfortable stablish behabe	adapt uncomfortable establish behave
IZGA02	exemple wifes sump anormal	example wives sum (to sum up) abnormal
MACI02	loosing	losing
MIOD02	mantain	maintain
PAAR02	believes (line 9)	beliefs
PABE02	wether	whether
QUES02	disturbe	disturb
RIRA02	caractheristic	characteristic

RUSH02	decission	decision
VIBO02	diferent belive live (line 1) (line 2)	different believe life
VIGO02	loose life (line 6) ‘cause	lose live because
T3 - FILE	**SPELLING IN TEXT**	**CORRECTION MADE**
BEGI03	preety	pretty
BLPA03	lifes (line 2)	lives
BORO03	wether	whether
CAAM03	expirience adultlife abandone	experience adult life abandon
CRGA03	knew	new (your new country)
GAME03	like (line 19)	life
GORO03	lifes	lives
GUMA03	pleasent foreing (last time) chose	pleasant foreign choose
IZGA03	prayin sump	praying sum
MACI03	appart definetely is (line 8) sing	apart definitely it sign (“a sign of respect”)
MIOD03	adress	address
PAAR03	lifes (line 1)	lives
PABE03	desdain	disdain
QUES03	missunderstanding chose	misunderstanding choose
REGA03	deppending	depending
RIRA03	complet	complete
RUSH03	opportunitie	opportunity
VIBO03	belive intelectual thinks (line 5)	believe intellectual things
T0 - FILE	**SPELLING IN TEXT**	**INTENTION (ASSUMED)**
HADA00	Use	use
HIJE000	ones (line 3)	one’s
IMDI00	their (line 12)	there
LESA00	ones (line 4) form (last line)	one’s from
LICH00	you	your

3. Non-words

T1-FILE	**Spelling in Text**	**Version 2 Correction**	**Version 3 Correction**
BEGI01	Humil	*eliminated*	humble
CATA01	fastly	fast	quickly
QUES01	relationed	*eliminated*	related
T2			
GORO02	dures	*eliminated*	lasts
IZGA02	musulm extrange	*eliminated* *eliminated*	Muslim strange
VIBO02	apports	*eliminated*	brings
T3			
CAAM03	Overwent through	went through	went through
CATA03	recollector recollectors	collector collectors	collector collectors
MIOD03	discriminised	*eliminated*	discriminated against
VIBO03	apport	*eliminated*	contribute

Appendix 3. Example of score sheet given to raters

Composition #____

Content	**30-27** EXCELLENT TO VERY GOOD **26-22** GOOD TO AVERAGE **21-17** FAIR TO POOR **16-13** VERY POOR	
Organization	**20-18** EXCELLENT TO VERY GOOD **17-14** GOOD TO AVERAGE **13-10** FAIR TO POOR **9-7** VERY POOR	
Vocabulary	**20-18** EXCELLENT TO VERY GOOD **17-14** GOOD TO AVERAGE **13-10** FAIR TO POOR **9-7** VERY POOR	
Language Use	**25-22** EXCELLENT TO VERY GOOD **21-18** GOOD TO AVERAGE **17-11** FAIR TO POOR **10-5** VERY POOR	
Mechanics	**5** EXCELLENT TO VERY GOOD **4** GOOD TO AVERAGE **3** FAIR TO POOR **2** VERY POOR ***OR** not enough to evaluate	

Total Score
Comments:

Appendix 4. Error codes for analysis of accuracy

Grammatical	(all types were counted as Gre)
GrE:trans	Error due to L1 transfer. Direct translation of an L1 structure that is ungrammatical in English
GrE:so	Subject omission
GrE:fw	Function word error (pronouns, determiners, conjunctions): omission of an obligatory pronoun not in subject position; use of the article with uncountable nouns; article omission; double subjects; wrong preposition; omission of a preposition*(except phrasal verbs)
GrE:v	Verb error: tense, aspect; omission of a verb that is needed in sentence structure; auxiliary verb missing; wrong form (word formation, morphology); wrong modal verb (would instead of should)
GrE:n	Noun error: missing noun; wrong form (word formation, morphology); singular/plural
GrE:ag	Agreement: sub/verb agreement
GrE:adv/j	Adverb/adjective error: confusing comparatives and superlatives; wrong form (word formation, morphology); adjective placement; pluralizing adjectives
GrE:wo	Word order: when one or several elements are misplaced in a sentence, when not due to transfer; lack of inversion in questions
GrE:neg	Negation errors: double negatives; confusion between no/not; negative particles
Lexical	(all types were counted as Lex)
LexE:idio	Non-words resulting from creative morphology (e.g. *fastly*) or L1 transfer (e.g. *reflexed*)
LexE:trans	words directly borrowed from the L1 whether modified or not; false friends
LexE:cho	Wrong word choice (not due to transfer): mistakes with commonly confused words: make/do; words that are inappropriate in the context (misunderstood dictionary definition)
Pragmatic	(all types were counted as Prag)
PragE:ref	erroneous use of reference markers, including anaphoric and cataphoric reference; ambiguous references.
PragE:idio	Idiosyncratic usage not clearly ungrammatical; problems with formulaic language and idioms
PragE:con	wrong discourse connector

Appendix 5. List of connectives tagged by Coh-Metrix in each of the 4 categories.

(Source: http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm.)

Any of the words in parentheses may be used. An * indicates that no word is also an option.

Additive Connectives

after all
again
all in all
also
and
as a final point
as well
at least
besides
by the way
correspondingly
finally
first ¡ (next/second)
for example
for instance
fortunately
further
furthermore
in actual fact
in addition
in fact
in other words
in sum
incidentally
instead
it follows
moreover
next
on (the)* one hand
once again
secondly ¡
similarly
summarizing
summing up
that is (to say)*
thereupon
to (these/this) ends
to conclude
to return to ¡
to sum up
to summarize
to take an example
too
well, at any rate
alternatively
and conversely
anyhow
but
by contrast
contrasted with
except that
however
in contrast
notwithstanding that
on the (one/other) hand
on the contrary
or (else)*
otherwise
rather
whereas
yet

Causal Connectives

a consequence of
after all
arise from
arise out of
as a consequence
as a result
as soon as
because
by
Cause
conditional upon
consequently
due to
Enable
even then
follow that
For
for (the/these/that) purpose
hence
if
in case
in order that
it follow that
it follows
make
now that
on (the)* condition that
on condition that
only if
provided that
purpose (of/for) which
pursuant to
since
so
the consequence of
then again
therefore
thus
to (these/this) ends
to that end
to those ends
Whenever
Although
even though
nevertheless
nonetheless
though
unless

Logical Connectives

a consequence of
actually
all in all
also
anyway
arise from
arise out of
as a consequence
as a final point
as a result
as if
as well
at least
at this point

Logical Connectives (con't)

because
besides
cause
conditional upon
consequently
correspondingly
due to
enable
essentially then
even then
finally
first ¡ (next/second)
first ¡ then
follow that
For
for (the/these/that) purpose
for example
for instance
fortunately
further
furthermore
hence
if
in (short/brief)
in actual fact
in any (case/event)
in case
in conclusion
in fact
in order that
in other words
in sum
incidentally
instead
it follows that
likewise
moreover
Next
on (the)* condition that
similarly
since
so
summarizing
summing up
That is (to say)*
the consequence of
Then
Then again
therefore
thereupon
Thus
to conclude
to (these/this) ends
to return to ¡
to sum up
to summarize
to take an example
to that end
to those ends
Well, at any rate
while
admittedly x, but y
alternatively
although
and conversely
anyhow
But
by contrast
contrasted with
despite the fact
even though
except that
however
in contrast
nevertheless
nonetheless
Nor
notwithstanding that
on the (one/other) hand
on the contrary
or (else)*
otherwise
rather
though
unless
whereas
yet

Temporal Connectives

(an/one/two etc.) hour later
A consequence of
after (a/some) time
after (this/that/all)*
again
all this time
as
as (long/soon) as
as a consequence
at first .. in the end
at the same time
at first ¡ finally
at last
at once
at this (moment/point)
before
by this time
earlier
even then
finally
first ¡ (next/second)
first ¡ then
follow that
from now on
further
immediately
in the meantime
instantly
It follows (that)*
just before
later
meanwhile
next
now that
on another occasion
once again
once more
only when
presently
previously
secondly ¡
simultaneously
since
so far
soon
suddenly
the consequence of
the last time
the previous moment
then (again/at last)*
this time
throughout
to that end
up till that time
up to now
when
whenever
while
until (then)*