USING ANALYTIC RUBRICS TO SUPPORT SECOND LANGUAGE WRITING DEVELOPMENT IN ONLINE TASKS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

SECOND LANGUAGE STUDIES

DECEMBER 2021

By

Kristin Naomi Rock

Dissertation Committee:

Marta González-Lloret, Chairperson
James D. Brown
Betsy Gilliland
Kristopher Kyle
Seongah Im, University Representative

Keywords: Assessment, rubrics, second language writing

For my mother, Susan Patricia Rogers, and my father, James Rock Riccabona

in loving memory

ACKNOWLEDGEMENTS

I would like to thank the members of my dissertation committee for their guidance and thoughtful feedback. Specifically, I would like to thank Marta González-Lloret for her enthusiasm and for never ceasing to believe that I would reach the end of this journey. I am grateful to James Dean Brown for sharing his knowledge and experience and for doing so in a way that enabled me to learn through practice. I would like to thank Betsy Gilliland for her insightful comments on earlier drafts of this dissertation and Kristopher Kyle for his patience as I worked to understand the intricacies of corpus linguistics research. I would like to thank Seongah Im for inspiring me to use many of the advanced statistics appearing in this project and for teaching me about the importance of inferential statistics in social science research. Though not a member of my committee, I am also grateful to Gabriele Kasper for expanding my capacity to think critically and for giving me the opportunity to commence and, in turn, to complete this academic journey.

In addition to the members of my committee, several individuals have helped me to grow as a language professional and as an Applied Linguistics researcher. First, I would like to thank the brilliant John S. Hedgcock, whose support and candid commentary gave me the courage to persist in moments of doubt. I would also like to thank the inspirational Rosa Maria Manchón Ruiz for taking me under her wing, for being an excellent model of a strong, successful female academic, and for connecting me with a wonderful group of researchers in Spain. Along that line, I would like to thank the members of the research team at the University of Murcia, including Lourdes Cerezo García, Yvette Coyle, Raquel Criado, Aitor Garcés Manzanera, Belén González Cruz, Joaquín Gris Roca, Sophie McBride, María Dolores Mellado Martínez, José Ángel Mercader Domínguez, Flori Nicolás Conesa, Irene Oñate Ruiz, Alba Pérez Velasco, Julio

Roca de Larios, Lucía Roth Martínez, Ester Sánchez Fajardo, Alberto José Sánchez López, Juan Solís Becerra, Arturo Valera García, and Lena Vasylets. Thank you for welcoming me into your classrooms, your team meetings, and your homes! Thank you also for the various roles that you played in enabling this project to arrive at its current form. I would especially like to thank the incredible Sophie McBride for enthusiastically supporting me in tasks both big and small, from summoning the motivation to write to locating the *concerje* to sharing a dryer. In addition to the research team at the University of Murcia, I would like to thank several individuals based at the *Servicio de idiomas* in Murcia, including Juana Rosario Sanmartín Vélez, José Saura Sánchez, and Lola Vidal. I also want to thank the many anonymous individuals who participated in various phases of this research. I would also like to thank the J. William Fulbright Foreign Scholarship Board, the Bilinski Educational Foundation, the *Fundación Séneca Región de Murcia*, and the *Ministerio de Ciencia, Innovación y Universidades* for their financial support.

I'd like to express my gratitude to various individuals who have supported me over the years. First, I would like to acknowledge several professors who inspired me in the early stages of my academic career. In addition to John S. Hedgcock, I want to thank Kathi Bailey, Lynn Goldstein, and Jean Turner for their encouragement and kind words. I would also like to thank Nicole Ziegler for believing in me and for introducing me to the genre of the statement of purpose. I would like to thank Joel Weaver for creating a caring, supportive work environment, and for advocating on my behalf. I am also grateful for the support of Theres Grüter throughout the dissertation writing process. I would like to show my appreciation to Jamie Simpson Steele for being the first professor to provide a rubric for an online discussion board post, and in turn, for creating a magnificent learning environment. I would like to thank Luke Harding for his comments on participant incorporation of emojis and Paula Winke for her encouragement and

assistance on an earlier version of the abstract. I'd like to acknowledge the "Word Wizard" Lynne Jaynes. I would also like to thank the lovely Deborah Crusan for welcoming a young and less experienced Kristin into the greater academic community. In addition, I wish to recognize the amazing individuals who were part of the International Language Testing Association's Graduate Student Assembly during its first year of existence, including Natalia de Andrade, Jorge Beltrán Zuniga, John Dylan Burton, Wenjun Elyse Ding, Hyunah Kim, Yunjung (Eunice) Nam, Haoshan (Sally) Ren, Olena Rossi, and Liberato Silva dos Santos. Of course, our Graduate Student Assembly would not have been possible without the mentorship of Gerriet Janssen, and I am grateful for his leadership and friendship.

I would like to thank a few of the wonderful language teachers who have impacted my life, including Hala Abdelal, Julio Baena, Antonio Carreño, Fernando Operé, Andrés Prieto, John Slater, and Hye Young Smith. I would also like to thank a close group of friends who have stuck with me no matter the country in which I was living or the progress I had made on a particular book club novel. Thank you Christina Dunn, Alexandra Hay, Heather Park Meek, Stephanie Rengulbai, and Chelsea Skovgaard. I am grateful for my friends and colleagues Sophia Durbin, Jessica Fast Michel, Keith C. Haller, Vera Hanaoka, Kendi M. Ho, Beth Godley, Nicole Miller, Andrea Peña Vasquez, Brigitte Rabie, Stacy L. Rodgers, Therese Tishakov, Myra Tran, Kristen Urada, and Laura Sardagna Viana. I wish to express my heartfelt thanks to Melanie Smith and the rest of the Smith family, as well as to Valerie Edrington and the extended Dressler, Rock, and Rodgers families. In closing, I convey my most sincere gratitude to Brian J. McKee and Dee Dee Riccabona for their unconditional love and support.

ABSTRACT

With the accelerated move to online learning, writing skills have become increasingly important for managing digital genres, such as educational blogs and discussion forums. Although effective written communication via such media is important for student success, many university-level second language learners navigate these unfamiliar tasks without access to guidelines concerning content, structure, and language use. Researchers in Applied Linguistics have suggested communicating teacher expectations through descriptive rubrics (Crusan, 2010; Ferris & Hedgcock, 2014; Weigle, 2002), and this dissertation investigates the effects of sharing an analytic rubric on learners' written development.

The first phase of this sequential mixed-methods research involved the expert review of academic blog posts written by learners of English as a foreign language (EFL). Quantitative and qualitative data gathered during the review led to the identification of five categories around which learners' written performance was assessed, including (a) genre-specific features (i.e., use of hyperlinks), (b) task fulfillment and relevancy, (c) content, (d) organization and balance, and (e) language use. On the resulting analytic rubric, each category was assessed on a 1- to 6-point scale. In Phase II, six raters used the rubric to score the posts written by 163 EFL learners. A many-facets Rasch analysis revealed that the rubric categories were functioning appropriately; however, the raters were not using the full 6-point scale.

In the final phase, written data from the blog entries of 31 learners were collected over two years, with 15 participants having access to the revised (4-point) rubric. After data collection, raters who were unaware of the order of composition scored three posts per participant according to the revised scale. A two-way repeated measures analysis of variance showed that the presence of the rubric, time, and the interaction of the rubric and time had a

significant positive impact on average scores ($p < 0.001$). Participants' longitudinal written development was also analyzed via nine linguistic indices covering lexical diversity, lexical sophistication, and syntactic complexity. The mean values for two variables, noun-adjective and verb-direct object dependency bigrams, demonstrated a significant change over time, while the moving-average type-token ratio (MATTR) and lexical decision time contributed to a regression model predicting 16% of variance in language use scores. A subsequent rhetorical moves analysis revealed a sequence of optimal steps for constructing an academic blog post. The results of this study are of use to pedagogues and researchers interested in digital genres, technology-mediated tasks, and second language writing assessment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Rationale for Research

With the exception of practices connected to the grammar-translation method of language teaching, the oral mode has held a privileged position in the field of Applied Linguistics for more than half a century. The primacy of the oral mode reaches back to descriptions of audiolingual classrooms in which emphasis was placed on mastering spoken forms of the second language and writing was "withheld until reasonably late in the language learning process" (Richards & Rodgers, 1986, p. 51). Cognitivist perspectives of second language acquisition have also stressed the oral mode (Manchón, 2014), hypothesizing, for example, that spoken language was the primary vehicle through which researchers could analyze variation in second language learners' *interlanguage*, or systematic linguistic behavior, which includes imperfect reflections of some norm (Selinker, 1972). According to Byrnes and Manchón (2014), since the publication of Selinker's seminal piece, "the language field has generally assumed that oral performance offers a privileged window into the nature of language acquisition precisely because of its 'spontaneous' and largely 'unmonitored' quality" (p. 2). Indeed, many of the theoretical principles underlying communicative language teaching, and by extension, task-based language teaching, were galvanized by questions surrounding the development of oral proficiency, i.e., negotiating for meaning in face-to-face interactions (Byrnes & Manchón, 2014).

In response to this prioritization of oral performance, several researchers have called for more nuanced investigations into the role of written production in promoting second language acquisition (Harklau, 2002; Manchón, 2009; Manchón & Roca De Larios, 2011). In Harklau's (2002) observations of high school language classrooms, the researcher discovered that written output consumed more class time than learner-learner interaction via spoken output. Harklau

commented that "interactions through writing and reading seemed pivotal in these particular learners' acquisition processes" (p. 332), a point echoed by researchers who have investigated the language learning potential of writing in English as a foreign language (EFL) contexts (Manchón, 2009; Manchón & Roca de Larios, 2011). In other words, both participating in meaningful oral communication with speakers and learners of the target language (Chaiklin, 2003; Ohta, 2000; van Lier & Matsuo, 2000) *and* producing purposeful, genre-sensitive written texts (Hyland, 2004) can lead to successful second language acquisition. Furthermore, several unique features of the written mode, including the permanence of written text (Byrnes & Manchón, 2014), the problem-solving nature of the composing process (Manchón et al., 2009), the ability to control the pace of production (Harklau, 2002), and the affordance of additional reflective time (Verspoor et al., 2012), contribute to an argument in favor of focusing attention on the language learning potential of writing. Therefore, in response to Manchón's (2011a) call to assign "written language learning a more central place in [Second Language Acquisition] studies" (p. 6), this research originates from the conception that writing is an essential component of English as a second language (ESL) and EFL curricula.

Writing in today's world is highly mediated by technology. The Internet and other technological innovations have expanded the number of contexts in which learners are expected to write in the target language, whether for personal, professional, or educational purposes. Although in some academic contexts computers and word processing software serve merely as an extension of pen-and-paper assignments (Bloch, 2008), in other situations, instructors have leveraged Web 2.0 technologies to provide safe, input-rich environments in which traditional students as well as distance learners are able to advance their digital literacy (González-Lloret & Ortega, 2014). For example, instructors at the tertiary level in the United States now ask students

to contribute written or multi-modal texts to discussions held via learning management systems such as Blackboard and Canvas (Sotillo, 2016). Across the Atlantic in Spain, EFL instructors have experimented with educational blogs to enhance students' writing skills (Vurdien, 2013). Irrespective of the rationale behind incorporating such tasks in ESL and EFL classrooms (i.e., to prepare students for participating in a more digital world, to stimulate creation of a community of practice, etc.), their frequency of use has heightened the necessity for analyses of the nature of these text-types. According to Elola and Oskoz (2017), educational stakeholders "need to acknowledge the profound shift that is occurring from traditional notions of literacies to digital literacies" (pp. 52-53); however, full descriptions of many of these emerging digital genres have yet to be developed (Elgort, 2017). It is critical that educators and researchers uncover the characteristics and discursive patterns of these texts in order to discern the means by which individuals utilize the medium to communicate effectively to a designated audience. In turn, this knowledge could be harnessed to help second and foreign language learners participate successfully in twenty-first century academic contexts. Rather than assuming that second language (L2) learners who have grown up in the digital age are competent in manipulating the affordances of a range of digital tools, educators and researchers should be cognizant that some students will benefit from explicit instruction in the capabilities of various Web 2.0 technologies and in the ways learners can utilize them effectively (Elola & Oskoz, 2017). In other words, through rich descriptions of the features of valued digital genres, we can make available the components of that cultural capital, eventually enabling learners to exploit and to push the boundaries of those genres creatively (Hyland, 2004; Tardy, 2016).

As writing within digital genres assumes a more central role in ESL and EFL curricula, methods for evaluating those texts should also come under the spotlight (Elola & Oskoz, 2017).

One advantage of approaching writing assessment from a genre-based perspective is that within this paradigm, students investigate and become acquainted with specific features of distinct text-types. In turn, descriptions of these features could serve as guidelines for improving writing performance or as evaluative criteria for inclusion on a scoring rubric. Hyland (2004) explained that scoring rubrics are common to all genre-based approaches to assessing writing, and in the L2 writing classroom, a scoring rubric likely refers to a grid that includes some combination of categories, descriptive criteria, and scoring bands (Brown, 2012b). Assessment by means of an analytic rubric can aid teachers in diagnosing the strengths and weaknesses of their students' written production, and it provides students with a more nuanced view of their progress toward specific learning goals. Hyland (2004) maintained that "as a result of such explicit criteria, students know how they will be assessed and what they have to do to be successful" (p. 164). Furthermore, analytic rubrics provide language learners with targeted feedback on various dimensions of their performance (Brown, 2012a; Ferris & Hedgcock, 2014; Hyland, 2004), thereby supporting them in the achievement of course outcomes. If second language writing instructors are going to teach the written genres that modern-day students must manage successfully, it would be beneficial to harness the power of evaluation tools like analytic rubrics. Scoring rubrics based on actual samples of target texts can help demystify the nature of these genres and scaffold learners' entrance into their desired academic communities. Thus, the primary goal of this research is to examine evidence in support of the use of rubric-based measures of student writing performance, and in turn, to document the longitudinal development of second language learners' performance on an online, academic writing task.

1.2 Research Gaps

1.2.1 Overuse of the Five-paragraph Essay

Over the past 50 years, a great deal of research on writing in U.S. high schools and universities has been conducted on the basis of the well-known five-paragraph essay (Elola & Oskoz, 2010; Ruiz-Funes, 2014). After the 1959 publication of an academic article recommending instruction in the five-paragraph essay (Caplan, 2019), the teaching of this particular text-type eventually became entrenched in second and foreign language writing curricula. The formula for the five-paragraph essay, which has been described as a "prescriptive statement of how *all* school writing should be done" (Caplan, 2019, p. 7), has also assumed a prominent role in many ESL and EFL writing textbooks and in various large-scale assessment instruments (i.e., TOEFL, IELTS, etc.). Crusan (2010) has cautioned that focusing solely on the five-paragraph essay in second language academic preparatory programs misrepresents the gamut of writing assignments encountered in university coursework. Indeed, the range of written work students are expected to produce in tertiary institutions of learning around the world extends far beyond the bounds of the formulaic five-paragraph essay (Melzer, 2014). While not ignoring certain attractions of this restrictive formula, including the availability of relevant teaching materials and the familiarity found in a structure that can be explained using five fingers, research aimed at increasing our understanding of the processes involved in writing in another language needs to move to include task types more representative of genres students will encounter in target use situations. For many academically-oriented English language learners, these future contexts may include study or research at an English-medium university and participation in the global economy, contexts in which the five-paragraph essay is only *one* of a large number of potential writing tasks. Therefore, it is time to shift attention away from this

frequently encountered form toward the analysis of additional academic genres, so that these text-types will also become part of mainstream ESL and EFL writing curricula and assessment paradigms.

In many international standardized tests of English, such as the International English Language Testing System (IELTS) and the Cambridge Certificate in Advanced English (CAE), writing tasks that require a response in the form of a five-paragraph essay have made this formulaic structure a tempting focal point for English language instruction. However, emphasizing solely this structure limits learners' exposure to other genres and fails to equip them with a broad repertoire of skills with which to confront different writing tasks and assessments. Even though the five-paragraph essay emerged as a solution to the time and word limits imposed by large-scale assessments (Crusan & Ruecker, 2019), Youn (2013) cautioned that an unbalanced emphasis on a test instrument's practicality as opposed to its authenticity could create a potential threat to validity. Seeing as decisions made by designers of large-scale tests have impacted classroom teaching around the world, leading occasionally to the phenomenon described as "teaching to the test," selecting a representative sample of writing tasks for these assessments should be a primary stakeholder concern. For both high-stakes and more localized assessments, Norris (2016) explained that task selection highlights the values held by stakeholders in a particular context, thereby "raising awareness about what is to be learned, why, and how" (p. 241). Therefore, careful consideration of the purposes for which a given student population will need to write ought to be present in the early stages of assessment design. It is also important for researchers to recognize that the Internet and other technological innovations have expanded the number of contexts in which learners will write in the target language (Elola

& Oskoz, 2017), and as a result, this dissertation responds to the need for explorations of ways to assess these new genres effectively.

1.2.2 Treatment of Scoring Rubrics

Scoring rubrics often play a role in research on the assessment and development of productive language skills, such as speaking and writing (Brown et al., 2002; Norris et al., 2002b). However, very few studies have utilized rubrics based on actual performance data, leading to doubts surrounding the quality of those assessment instruments. Since gatekeepers and language teachers frequently make decisions about students on the basis of results on performance assessments, Brown (1995) asked, "How good are the tests upon which such decisions are based?" (p. 112). A simple Internet search will uncover a plethora of scoring rubrics that include descriptors of effective five-paragraph essays, yet these rubrics are not likely to be viable for assessing student performance in a different writing task. In fact, when raters are asked to assign scores to written products utilizing a rubric that does not describe adequately the nature of the texts under consideration, raters will rely on a more personal understanding of the types of performance representative of different scoring levels (Lumley, 2002). As a result, evaluations of learners' written products may not be consistent across raters or across scoring periods. One way to respond to this concern involves developing rating scale criteria based on a systematic genre analysis of learner performance on a particular writing task.

While rating scales form part of assessment instruments at both the local and global levels, this dissertation establishes that rubrics can also serve as a point of reference for language learners who wish to strengthen their genre knowledge and to monitor changes in their genre-specific writing over time. In other words, rubrics constructed on the basis of actual performance data can be used both as a valid method for assessing second language writing *and* as a tool for

self-assessment and instructor feedback. However, there appears to be an absence of studies within the field of second language writing that explore the value of analytic rubrics for promoting second language writing development in particular genres. Certainly, to encourage students' optimal performance on writing tasks, teachers should provide learners with detailed scoring criteria in advance of a particular assignment (Crusan, 2010; Weigle, 2002). However, to my knowledge, in the field of Second Language Studies there are no studies that investigate empirically the language learning potential of providing students with these rubrics and instructing them on how to assess their writing performance in relation to the scoring criteria. Ultimately, teaching and assessment are not value-neutral, and since research in Applied Linguistics is often underpinned by social justice goals (Ortega, 2019), the current project aims to empower English language learners and first-generation college students by providing them with the criteria upon which an academic, genre-specific text is judged.

1.3 Research Questions

In light of the research gaps previously identified, four research questions guided this dissertation. In addition to the primary foci, sub-questions were added to the first two questions in order to support the collection and analysis of relevant data. The complete list of research questions is as follows:

RQ1: Based on examples of actual student performance on an online genre-based task, what features should be included in analytical rating criteria used to assess written performance on that task?

1.1: What aspects of student performance do expert reviewers find salient for making quality judgments of individual texts?

1.2: How can qualitative data provided by the reviewers support the inductive formation of rubric categories and level descriptors?

RQ2: To what extent does evidence support the use of the analytic rubric as a measure of participant second language writing performance on the online genre-based task?

2.1: How are the performances of examinees, raters, and categories related when they are placed on the same logit scale?

2.2: To what degree are the raters biased toward particular categories?

2.3: To what degree are the rating scale levels functioning within each category?

RQ3: How does the longitudinal written performance of English as a Foreign Language university students on an online genre-based task differ between learners who are provided with the detailed analytic rubric and learners in the same context who do not have access to the rubric?

RQ4: In what ways does a linguistic analysis of participants' longitudinal performance, as well as a rhetorical moves analysis of their genre-specific texts, contribute to our understanding of the written development of both groups of participants and the nature of this emerging genre?

1.4 Structure of the Dissertation

In order to guide the reader through this doctoral dissertation, the text has been divided into eight chapters. The current chapter has presented the rationale for the study, and it has highlighted gaps in Applied Linguistics research pertaining to the teaching and assessment of second language writing. The research questions listed in Section 1.3 serve as the unifying mechanism for the review of literature, the method, the presentation of results, and the discussion of the implications of those findings. The next two chapters cover a review of relevant literature.

Chapter 2 begins with an examination of the notion of "task" and the emergence of task-based language teaching (TBLT). To strengthen the study's theoretical grounding, a discussion of the ways in which TBLT has grown to include second language writing tasks is followed by a review of genre-based approaches to teaching writing. Next, a description of characteristics that empirical studies have found to be indicative of effective second language writing tasks leads to an examination of the emergence of new digital genres. Chapter 3 begins with a consideration of the role of feedback and assessment in promoting second language writing development. This review then advances into an academic reflection on the utility of analytic rubrics and the procedures whereby validity evidence supporting the contextualized use of a given rubric can be collected.

After this review of literature, Chapter 4 introduces the methodological framework that underscores the research. Given the complexity of the study design, the utility of a sequential mixed-methods approach is discussed, and the relevance of design-based research to the iterative nature of the project is explained. The chapter continues with a chronological presentation of the three phases into which the research is divided. The first phase of the research covers the creation of an analytic rubric for a specific, online genre-based task. The second phase documents the process of revising the initial rubric, and the third phase reports on the portion of the project that necessitated the collection of longitudinal data. Within Chapter 4, coverage of each phase is divided into four sections detailing the participants, the materials, the procedures, and the analyses of data pertinent to the stage in question. After this detailed account of the manner in which data were collected and analyzed, Chapter 5 initiates the presentation of results. The first half of the chapter covers the creation of the analytic rubric. Quantitative and qualitative results are discussed in turn, and the contribution of the different data types to the construction of

the genre-based rating scale is reviewed in detail. The second half of Chapter 5 presents data obtained from the systematic collection of validity evidence in support of the rubric's use as a measure of second language writing performance. The results described in this section include a discussion of the ways in which the rating scale was revised before it was employed in Phase III. Chapter 6 relates the results of the third phase of the research. Quantitative and qualitative data collected during this longitudinal stage assist in answering the third and fourth research questions. Specifically, the information obtained in the final phase of the project enable an evaluation of the impact of the application of the genre-specific analytic rubric on second language writing performance. Chapter 7 presents a discussion of the results from all three phases of the research. The chapter follows the order of presentation of the study's research questions and sub-questions to interpret the findings, with subsequent sections addressing the study's implications for theory and practice. The final chapter of the dissertation, Chapter 8, provides a summary of the research. In addition, it addresses certain limitations in light of various possibilities for future research.

CHAPTER 2: GENRE-BASED WRITING TASKS IN A DIGITAL WORLD

2.1 Tasks and Task-based Language Teaching

In North American contexts in the middle of the twentieth century, behaviorist theories of learning exerted a heavy influence on the practices of foreign language teachers (Lightbown & Spada, 2006). Audiolingualism—a popular method at that time—emphasized the role of habit formation in the learning of a second language (L2), whereby memorization of short dialogues and engagement in oral pattern drills reduced the possibility of a speaker making mistakes (Richards & Rodgers, 1986). In the 1960s, critics of the decontextualized, drill-based mastery of linguistic structures called for an approach to language learning centered on promoting learners' communicative competence, an approach later labeled Communicative Language Teaching (CLT). With roots in CLT and aligned with contemporary perspectives "on the learning of complex functional abilities," Task-based Language Teaching (TBLT) emerged as a response to what researchers viewed as only "partial incorporation of communication work within the field of language education" (Van den Branden et al., 2009, p. 5). TBLT theorists promoted a model of second language education that was meaning- and learner-focused, and in which the notion of *task* simultaneously encompassed an educational goal, a pedagogic activity, and an informative assessment (Van den Branden et al., 2009).

Within literature on task-based language teaching, researchers have proposed a wide variety of definitions of "task," from information gap activities to authentic communication with other speakers of the language. Some of the frequently cited characterizations of *task* are the definitions provided by Samuda and Bygate (2008), Skehan (1998), Van den Branden (2006), and Willis (1996). Samuda and Bygate (2008) defined a task as "a holistic activity which engages language use in order to achieve some non-linguistic outcome while meeting a linguistic

challenge, with the overall aim of promoting language learning through process or product or both" (p. 69). On the other hand, and not entirely incompatible with the first definition, Skehan (1998) described a *true* language learning task as an activity in which "meaning is primary; there is some communication problem to solve; there is some sort of relationship to comparable real-world activities; task completion has some priority; and the assessment of the task is in terms of outcome" (p. 95). Defining the concept of task in such explicit terms can assist teachers in designing instruction that maximizes learners' engagement of acquisitional resources. For Van den Branden (2006), a task is "an activity in which a person engages in order to attain an objective, and which necessitates the use of language" (p. 4), and for Willis (1996), a task is a "goal-oriented communicative activity with a specific outcome, where the emphasis is on exchanging meaning, not producing specific language forms" (p. 36). Ultimately, a task needs to have a clear objective, engage learners in using the target language, and provide for an outcome beyond the mere practice of a particular linguistic feature. Although the above definitions of task do not preclude writing activities, to date, most research on tasks has centered on oral communication. In response to this unbalance, several scholars have called for an expansion of the nature of research conducted under the TBLT umbrella to include writing tasks (Byrnes & Manchón, 2014).

2.1.1 Task-based Language Teaching and Second Language Writing

Parting from the dominant oral tradition, an increasing number of TBLT investigators have pursued research agendas focused instead on written output. Specifically, these scholars have looked to tasks as a means to address questions related to learners' literacy development. Byrnes (2014) explained that "our understanding of 'output' as facilitating language learning and language development is incrementally enriched when it is reframed as task, and in turn, when

task is reframed as an intellectually challenging writing task that fosters meaning making" (p. 87). In other words, by fusing the concept of task with second language writing contexts, scholars can gain additional insight on the role of TBLT in fostering second language development. Along this vein, Manchón (2011a) has pointed out that the teaching of second language writing serves not only to help students learn to communicate particular meanings in written form, but also to engage them in writing tasks that may support the development of other language skills and language knowledge.

One way in which researchers in the field of Second Language Writing (SLW) have connected their work to the field of TBLT is through theoretical and empirical investigations testing aspects of Robinson's (2011) Triadic Componential Framework and Skehan's (1998) Limited Capacity Model. Skehan's model hypothesizes that increases in task difficulty heighten the demands placed on an individual's cognitive resources. In turn, the learner, attempting to meet the exigency for greater complexity, improved accuracy, and enhanced fluency, must often make a sacrifice in at least one of those three areas. In other words, the mental capacity of a human being is not unlimited, and in an effort to meet the heightened demands of a more difficult second language task, learners may need to redirect some of their mental energy away from accuracy, for example, and toward fluency. In contrast to Skehan's model, Robinson's cognitive hypothesis makes a distinction between task *difficulty* and task *complexity*. Whereas the concept of task difficulty encompasses individual learner variables, including a learner's aptitude for handling a given task, task complexity addresses the cognitive demands that a particular task imposes on the learner. The third component of Robinson's framework concerns task *condition*, a feature intended to describe the type of interaction required by the task. Even though Skehan's model accounts for learner interaction with a given task, Robinson's cognitive

hypothesis compartmentalizes learner-task involvement into learner-, task-, and output-specific variables.

In reference to Robinson's framework, Manchón (2014) questioned whether or not manipulation of various task complexity factors, including the amount of planning time, the need for complex spatial reasoning, and the location of referenced events in a specific time and space, were relevant for investigations of writing tasks. Manchón noted that while planning time may or may not factor into learner performance on oral communication tasks, planning "is an intrinsic component of the writing process" at least in part due to "the offline nature of most forms of writing" (p. 32). Manchón also noted that the provision or withholding of feedback on writing as well as the collaborative or individual nature of a particular writing task would need to be added to the list of potentially relevant task conditions. In sum, as research within TBLT expands to include writing tasks of various dimensions, traditionally explored frameworks, such as the theories posited by Skehan (1998) and Robinson (2011) ought to be revisited in light of the differing demands posed by oral and written tasks.

2.1.2 Needs Analysis and TBLT

The previous theoretical discussion of the manner in which the notion of task might pertain to the study of second language writing provides a useful framework for addressing the centrality of needs analysis to TBLT. A needs analysis, as interpreted by Long (2015), should serve as the point of departure for curricular design at any level. In this way, task-based language programs can address widespread stakeholder demands for the incorporation of relevant content and language in applicable learning modules. Given the political nature of language (Gee, 2005) as well as the economic value attached to knowledge of multiple languages, especially English (McGuire, 1996), investing in a careful evaluation of the tasks that a particular group of learners

will need to be able to perform in the second language would imbue that program with a higher degree of credibility. Ideally, a needs analysis would incorporate multiple sources of information and measures (González-Lloret, 2014; Long, 2015) in an effort to define a *defensible curriculum* (Brown, 1995). Regardless of the sequence in which a language program gathers information (see Graves, 2000 for an example of a needs analysis cycle), the "use of *multiple sources* typically provides more detailed information and has the additional advantage of allowing cross-checking for information validation, and ideally, triangulation of findings" (Long, 2015, p. 118). Since administrators, teachers, and students have different, yet complementary roles in task-based language programs, gathering information from various sources helps to ensure that multiple stakeholder perspectives are represented in the curricular design—or test development—process.

2.1.2.1 University-level Writing Tasks in the United States and Abroad. One particularly useful source for identifying target tasks is professional literature that documents language use in the target context. For the target context of university-level, written academic communication, at least three large-scale projects have sought to identify the nature of writing tasks assigned by lecturers and professors in different degree programs. In the first study, which was sponsored by the Educational Testing Service (ETS), Hale et al. (1996) set out to gather actual samples of university-level writing tasks in order to contribute to the ongoing development of the Test of English as a Foreign Language (TOEFL). The researchers collected writing tasks assigned by participating faculty and instructors at eight North American universities, and they acknowledged having selected these universities in part due to the large number of international students enrolled at each campus. The sample set was narrowed further by focusing on disciplines that attracted the most international students and on courses within

those disciplines that recorded the highest enrollment. Of the 162 sets of class materials collected by the researchers, Hale et al. used a sample of 110 sets to develop a classification scheme resulting in six categories: locus of writing (either in-class or out-of-class), length of product, genre, cognitive demands, rhetorical task, and pattern of exposition. Equipped with this classification system, five of the authors proceeded to classify all of the writing assignments. High levels of inter-rater agreement were reported for the locus of writing and the expected length of the product—dimensions whose levels "had clearly defined boundaries, which were readily understood by the judges" (p. 43). However, for the remaining categories, the authors described their agreement rates merely as 'fair.' Whereas all tasks were classified according to the first three categories (physical location of writing, length, and genre), only those assignments that the researchers labeled as belonging to the 'essay genre' were classified further (i.e., according to cognitive demands, rhetorical task, and pattern of exposition). A principal finding was that short, in-class writing tasks were significantly more common in mathematics, physical science, and engineering courses than in courses in the social sciences and humanities. Furthermore, and as might be expected, short writing tasks were more frequently encountered as in-class rather than out-of-class assignments, a finding that Hale et al. attributed to the appearance of short tasks on in-class exams. Importantly, Hale et al. concluded their report with a call for additional research examining students' written responses to the writing assignments collected for the study as well as instructors' evaluations of those products.

Not long after Hale et al.'s (1996) findings were published in an ETS report, Melzer (2014) set out on an equally ambitious project. Rather than requesting materials from university faculty, the researcher bypassed the risk of a low response rate by taking advantage of the Internet. Between the years 1999 and 2007, Melzer accessed a plethora of course websites,

eventually gathering 2,101 undergraduate writing assignments from various disciplines housed within 100 American universities. Of those 2,101 writing assignments, the majority (66%) required an informative response in which students needed to demonstrate their content knowledge. In addition, more than one-fifth of the assignments consisted of short-answer or essay exams, and the intended audience for most of the writing tasks (some 64% of cases) was solely the course instructor. This discovery, along with a lack of poetic writing, was consistent with earlier survey research, yet Melzer documented the addition of journaling assignments, or 'exploratory' writing. Specifically, Melzer noted that "exploratory journals and their computer age equivalent, the electronic discussion board, [were] a common phenomenon" (pp. 25-26).

Outside of the North American context, Castelló et al. (2012) set out to investigate the writing practices in place at four Spanish universities. The researchers sent a survey with four Likert-scale questions and one open-ended item to participating professors at universities in Madrid and Barcelona. In total, 106 professors, primarily from the social sciences, responded to the survey. Out of a list of specific pedagogical activities, the professors indicated that academic writing and reading scientific texts were the activities most favorable for student learning. At the same time, the respondents reported that the skills for which their university-level students were least prepared were academic writing and reading scientific texts. On the flip side, the professors acknowledged that most students were competent in note-taking and sitting for exams. The most frequently assigned writing tasks included reports, essays, and open-ended exam questions, in descending order, with tasks such as PowerPoint presentations, summaries, blogs, and forums listed as "other texts" assigned in the respondents' courses. Although the language in which students were expected to perform these writing tasks was not specified, Spanish and Catalan are assumed to be the languages used most frequently, followed by English. Based on the findings of

Castelló et al. (2012), Hale et al. (1996), and Melzer (2014), the range of writing tasks assigned across North American and Spanish universities demonstrated striking similarities. Each project reported on the high-frequency of short-answer, or in-class writing assignments constructed in response to exam questions, and the prevalence of essays or compositions, the nature of which will be examined in subsequent sections of this chapter. In spite of the acknowledged pervasiveness of exams and "essays," the two most recent studies, Castelló et al (2012) and Melzer (2014), identified the emergence of digital writing tasks, including electronic discussion boards, blogs, and forums. Throughout their reports, the authors of all three studies recurred frequently to a term that warrants further elucidation ('genre'), and as such, it will serve as the subject of an upcoming section.

2.2 Genre and Genre-based Writing Tasks

Over the past several decades, researchers interested in first and second language writing development have proposed different foci for the teaching of writing. Of these theories, the process approach, with its clearly defined stages for planning, drafting, and revising, has had the greatest impact on research agendas and on the teaching of L2 writing (Hyland, 2003, 2016). Proponents of the process model relegated language concerns to the end of the composing process—the 'editing' or 'revising' stages—and they viewed writing primarily as an individual, cognitive activity. Thus, just as TBLT responded to an identified gap in the actualization of communicative language teaching theories, genre-based approaches to teaching writing responded to the failure of the process approach to account for the social nature of written texts and to address the role of language in a more comprehensive way (Hyland, 2003). In line with van Lier's (2004) assertion that "language cannot be 'boiled down' to grammar or meaning only…or separated from the totality of ways of communicating and making sense of the world

we use" (p. 24), genre-based approaches to the teaching of L2 writing recognized that individual linguistic choices responded to particular social contexts and communicative purposes.

2.2.1 Three Theoretical Perspectives on Genre

In theoretical discussions around the concept of *genre*, scholars often reference three schools of thought, which are not entirely independent of one another, but which define the term in unique ways. In addition to nuances surrounding their definitions of genre, the three theoretical perspectives characterized genre elements in a distinct manner and responded initially to different groups of language learners. The labels assigned to each approach include Systemic Functional Linguistics, New Rhetoric, and English for Specific Purposes (Hyland, 2004). Since the research for this dissertation more closely aligns with work conducted under the English for Specific Purposes (ESP) tradition, a brief discussion of the other two schools of thought is followed by a more protracted consideration of the theoretical framework behind ESP.

Common to each theoretical orientation is an understanding of genre as social practice (Tardy, 2012), and appropriately, the oft-cited definition from within Systemic Functional Linguistics (SFL)—the Sydney School—explains genres as "staged, goal-oriented social processes" (Martin, 2002, p. 56). For Martin (1999) the notion of genre served as a vehicle for broadening Michael Halliday's functional theory of language to include the global purpose of individual written texts. In other words, genre subsumed the construct of *register*, or the lexicogrammatical choices made by a language user according to the variables of field, tenor and mode (Byrnes, 2012). Specifically, *field* referred to the social activity, *tenor* to the relationships among individuals involved in an interaction, and *mode* to the manner in which the message was communicated (i.e., verbal, written, etc.). Thus, as Martin (2002) explained, "register is a pattern of linguistic choices, and genre a pattern of register choices" (p. 57). For SFL theorists then,

genres refer to sets of texts with similar discursive patterns, wherein combinations of linguistic elements are a central focus.

Within the Sydney School, genre is synonymous with rhetorical pattern or rhetorical mode (Hyland, 2004). Examples of rhetorical modes frequently encountered in second language writing textbooks and curricula include description, explanation, and persuasion. English as a second language learners are not the only students who have been asked to manipulate these genres, as their like are often included in national directives for K-12 education. In Hinkel's (2015) discussion of the features of writing valued in American schools, the author pointed out that the Common Core State Standards promote "three types of texts and writing skills: to persuade, to explain, and to convey real or imagined experience" (p. 22). Byrnes (2012) has maintained that within SFL, references to individual genres such as recounts or explanations serve to capture flexible textual characteristics that are often blended or inter-mixed.

In a similar vein, Hyland (2004) explained that for some SFL theorists, a further distinction is made between elemental genres such as description and narration, and *macrogenres*, whose authors utilize multiple elemental genres to arrive at a coherent whole. For example, the macrogenre of a lab report may combine several elemental genres, including procedure, report, and description. Although such labels for text types can be found across the curriculum, the scope of Melzer's (2014) research led him to conclude that "it is impossible to speak of 'arguing' or 'describing' or 'explaining' in general terms, since each discipline—and often different teachers in the same discipline—have a slightly different way of defining each of these strategies" (p. 64). In order to avoid the ambiguity associated with the mention of particular text-types, along with a sincere desire to minimize the amount of discipline-specific jargon

appearing in this dissertation, treatment of the SFL approach to genre will now give way to a consideration of rhetorical theory.

Discussions of the New Rhetoric (NR) approach to genre generally acknowledge research on first language rhetoric as a precursor to its theoretical framework. As part of the New Rhetoric, Miller (1984) is often credited with initiating a line of research that elevated the importance of social context. For Miller, "a rhetorically sound definition of genre must be centered not on the substance or the form of discourse but on the action it is used to accomplish" (p. 151). For example, the essence of the genre of a legal brief rests not in its structure, but rather in the manner by which the written legal argument allows the claimant to participate in the broader community. One notable feature of this perspective is that it tends to eschew taxonomic efforts, for in the New Rhetoric, genre analysis is not equivalent to classification. Another tenet of this school of thought, which has been regarded in positive and negative ways, rests in rhetoricians' tendency to focus on the dynamic nature of genres (Hyland, 2004; Tardy, 2012). On one hand, this perspective promotes an understanding of genres as non-static entities that cannot be boiled down to a fixed template. On the other hand, this reticence to acknowledge certain stable properties of different genres has resulted in criticisms of the New Rhetoric as a theory incompatible with pedagogic aims. Hyland (2004) stated that in contrast to other theoretical perspectives of genre, "NR research has generally been less interested in describing the linguistic similarities of texts for teaching purposes" (p. 36). Hyland also suggested that NR viewed the classroom as an inauthentic context, and as a result, proponents of NR have assigned less value to classroom genres in favor of texts encountered in real-world communities of practice (Hyon, 1996). However, research conducted by Tardy (2012) in accordance with Rhetorical Genre Theory seems to counter this claim as the project focused on genres of importance in a tertiary

academic context. Based on extensive ethnographic research of four international graduate students at a U.S. university, including the collection of the students' written work, interviews, and observations of class discussions, Tardy concluded that the writers relied on their experience with similar genres when confronting a writing task in an unfamiliar genre.

Although NR has also been criticized for focusing more on the ways in which so-called "expert users" manipulate genres (Hyland, 2004), researchers associated with the English for Specific Purposes (ESP) tradition have acknowledged the contributions of New Rhetoric studies in establishing ESP's theoretical foundation (Swales, 1990). Swales (1990) explicitly credited NR for influencing his understanding of the social aspect of genre. He also applauded the contributions of SFL to the study of genre, specifically referencing efforts to disentangle the notion of genre from that of register and to pinpoint the goal-oriented nature of distinct communicative events. The approach to genre analysis initiated by Swales laid out a working definition of the construct as communicative events (whether oral or written) that could be linked on the basis of purpose—as opposed to form—and that together established certain constraining conventions. Swales acknowledged that the labels assigned by discourse communities to specific genres were not to be discredited seeing as the active members of a particular discourse community were the individuals most likely to have advanced knowledge of genre conventions. As a result, from this theoretical perspective, genre labels such as "PhD dissertation," "conference abstract," and "letter of recommendation" came to the forefront, as opposed to "narrative," or "recount" in the SFL tradition.

Building on Swales' (1990) definition, Hyland (2004) defined *genres* as "the purposive social actions routinely used and recognized by community members to achieve a particular purpose, written for a particular audience and employed in a particular context" (p. 45).

Appropriately, this definition of genre shares certain similarities with Skehan's (1998) description of a true task in that both concepts emphasize the connection between an underlying communicative purpose and the means by which that message is conveyed. This link between task and genre serves as a foundation for outlining the synergy achieved by applying TBLT principles to the development of effective L2 writing tasks. Before arriving at that explanation, however, the next section will describe briefly an example of an overused academic writing activity that falls short of gaining recognition as a task or as a genre.

2.2.2 The Five-Paragraph Essay

Individuals who have some degree of familiarity with secondary and tertiary education in the United States and in Spain would likely find the term *essay* commonplace. Instructors of English in the United States continue to identify essays as a distinct genre (Ortmeier-Hooper, 2019), and secondary students in the autonomous community of Murcia must produce a successful argumentative, descriptive, or opinion essay in English to obtain admission to the University of Murcia (Universidad de Murcia, 2021). Hale et al. (1996) also reported on the 'essay genre,' examples of which the researchers classified according to the assignment's cognitive demands, rhetorical task, and pattern of exposition. Interestingly, Hale et al. concluded that all of the writing tasks assigned in the essay category involved exposition, and the three most common patterns of exposition included cause-effect/problem-solution, classification/enumeration, and comparison/contrast. In view of the above findings, it may come as no surprise that formulas for composing cause-effect, problem-solution, and comparison-contrast essays appear in many contemporary English language writing textbooks (Aquino-Cutcher et al., 2016; Beaumont, 2012; Ward, 2012). Furthermore, the scoring rubrics that correspond to the productive portion of the English language exam at the University of Murcia

also delineate the expected components of a five-paragraph argumentative, descriptive, or opinion essay. The rating scale for the argumentative essay, for example, outlines the requirements for the introductory paragraph and topic sentence, along with the expectations for relevant supporting paragraphs and a concluding paragraph that restates the topic presented in the introductory paragraph in "a new, more insightful manner" (Universidad de Murcia, 2021).

Essentially, an essay is defined by its form rather than its communicative purpose. As a result, the familiar five-paragraph structure consisting of an introduction, three body paragraphs, and a conclusion falls short of the three definitions of genre presented in the preceding section. A bulleted list, a form that can be adapted to different situations, would be closer in nature to the five-paragraph essay than a lab report or statement of purpose (Tardy, 2019). According to Caplan (2019), the form has been overextended, and "today's L2 writing textbooks have largely lost this sense of genre, audience, and purpose in favor of simplistic attention to structure" (p. 10). Following years of study, Johns (2019) arrived at a similar conclusion, noting that the essay moniker has been used to describe almost any type of writing assignment. Not only does the five-paragraph essay fail to satisfy modern definitions of genre, but also it struggles to live up to several tenets of Skehan's (1998) theorization of task. Instead of placing meaning at the center of task design, the five-paragraph essay emphasizes form. In addition, a sense of audience or communicative purpose has been removed from the equation, and the essay may have little relationship to the types of genres valued in actual discourse communities. In sum, whereas some genre-based, academic writing tasks will assume the form of an essay, the five-paragraph structure is neither a genre nor a true task. Accordingly, the following section seeks to elucidate the ways in which the concepts of genre and task can be combined to forward a dynamic and purposeful pedagogical framework.

2.2.3 Characteristics of Effective, Genre-based Writing Tasks

Similar to the ontogenesis of task-based language teaching, the development of genre-based pedagogies emerged from within the communicative tradition. More specifically, genre-based teaching has paid homage to Vygotsky's work on child development (Hyland, 2004), which emphasized the active role of the learner within a particular social environment. Vygotsky has been credited with formulating the notion of a learner's *Zone of Proximal Development* (ZPD), originally defined as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). Vygotsky's theory suggested that knowledge is acquired and transformed as a result of participation in meaningful interaction, an assumption that necessitates the presence of more than one individual. In other words, language learning—and by extension, second language writing development—rarely occurs in isolation, and genre-based approaches to the teaching of L2 writing have acknowledged the ways in which writers make linguistic choices based on an awareness of their intended audience (Hyland, 2016).

In TBLT's formative period, Swales (2009) identified a connection among tasks, the development of genre knowledge, and the relevance of that genre knowledge to successful participation in a particular discourse community. Yasuda (2012) then developed Swales' ideas further by suggesting that the amalgamation of genre and task would help "to operationalize a writing pedagogy that is focused on a range of social functions in written language" (p. 13). Byrnes (2014) also suggested a productive association among genre, meaning making, and task-inspired writing, highlighting the language learning potential of socially informed writing tasks. Synthesizing the work of all three scholars, I would propose that within genre-based models of

L2 writing pedagogy, researchers and practitioners can view *writing tasks* as a means through which second language learners in educational environments can participate in genre-based discourse communities, thereby providing contextualized opportunities for linguistic development. Thus, through the combination of genre-based approaches to teaching writing and task-based approaches to teaching language, practitioners may find a powerful means of promoting second language writing development. With the intent of identifying certain pragmatic extensions of this pedagogical framework, the section continues by outlining five features of effective, genre-based writing tasks. Specifically, research has suggested that second language writing tasks be explicit, genuine, recurrent, social, and varied.

2.2.3.1 Explicit. Samuda and Bygate (2008) asserted that decades of research in Applied Linguistics have demonstrated the necessity of accompanying language tasks with appropriate support. The researchers stated specifically that on its own, a task would be insufficient for garnering a level of engagement that would produce meaningful learner-learner interaction. In other words, simply providing a pair of learners with a series of pictures and asking them to write a story would be a less than adequate means of encouraging the development of the students' productive language skills. Research spanning multiple fields has contributed to practitioners' awareness of the role of communicating clearly defined objectives and instructions in promoting on-task behavior (Cangelosi, 2000; Wiggins & McTighe, 2005). Furthermore, the examination of model texts (Hyland, 2003, 2004) and the provision of assessment criteria could help learners succeed on individual writing assignments (Crusan, 2010; Ferris & Hedgcock, 2014). In fact, one of the primary advantages of combining task-based approaches and genre-based approaches to teaching writing rests in applying the explicitness of genre-based pedagogies to the creation and execution of true tasks. Essentially, genre-based teaching aims to make explicit for learners the

language patterns and structures that render a text recognizable as pertaining to a particular genre (Hyland, 2004). Thus, by presenting explicit goals for a given writing task and clarifying the steps by which learners could reach those targets, teachers would maximize learner involvement in working toward desired pedagogic aims. In a case-study documenting one language learner's progression through a genre-based writing task, Byrnes (2011) explained that the instructions for all genre-based writing in the German Department at the researcher's university adhered to a standard format that specified the task genre, the content "in terms of obligatory and optional genre moves," and its "linguistic realization at the discourse, sentence, and lexicogrammatical levels" (p. 138). The degree of explicitness in task instructional materials combined with the focal learner's investment in the composing process may well have contributed to the student's successful appropriation of the summary genre, which involved the use of reporting verbs, a consistent register, and nominalization. In sum, effective writing tasks ought to include the provision of clear instructions, the examination of model texts, and the communication of explicit descriptions of the genre under focus.

2.2.3.2 Genuine. Seeing as a central concern for many TBLT researchers lies in the relatability of pedagogic tasks to situations in which learners might use the target language outside of class, references to the 'real-world' and 'authenticity' abound. However, discussions concerning the breadth of tasks to which the 'authentic' label applies are often fraught with discord. Researchers and practitioners have many reasons for engaging in such discussions, not least of which is the search for a clear definition to guide the selection of appropriate writing tasks. In the midst of an ambiguous understanding of 'authentic,' doubts may arise concerning the nature of the criteria that ought to factor into pedagogical decision-making. For example, if an instructor hoped to align his or her curriculum with Robinson's (2011) interpretation of 'real-

world' target tasks—activities learners must accomplish "in order to be successful in various domains of lifetime endeavor outside the language classroom" (p. 11), he or she might question the authenticity of any classroom-based genres, such as syllabi or note-taking. On the other side of the debate, Casanave (2017) has suggested that discussions of authenticity ought to consider "whether the school and testing contexts themselves are their own authentic worlds" (p. 236). Harklau (2002) argued that for most of the world, a classroom is just as natural a learning environment as the 'wild' outside the classroom. Peyton (2000), for example, pointed to the authenticity of teacher-student written communication via dialogue journals as an example of a classroom-based genre that afforded students authentic reading and writing experiences. Given the theoretical subtext surrounding the term 'authentic,' the synonym, *genuine*, might be a more appropriate adjective, whose conceptualization would be best informed by local needs analyses. A mixed- or multi-method approach to gathering needs data would uncover the genres valued by literacy communities the learners hoped to join (Ferris & Hedgcock, 2014). Thus, a genuine writing task would be context-sensitive and relevant to the genre-based developmental needs of the target population.

2.2.3.3 Recurrent. To promote the development of second language students' academic writing, genre-based writing tasks need to be repeated. The repetition of tasks has been thought to prime students for language learning, seeing as the attentional resources necessary to manage the task's goals and procedures on the first occasion could be redirected toward engaging additional cognitive processes during subsequent iterations (Ziegler, 2016). Hedgcock and Ferris (2009) maintained that engaging in regular, even daily decision-making processes "about what to write (ideas), in what order to present those ideas (rhetoric), and what linguistic or extralinguistic tools to utilize" is likely to play a beneficial role in learners' linguistic development (p. 215).

Whereas a great deal of research in TBLT has compared learner performance across different tasks or different task conditions, relatively little interest has been demonstrated in examining the ways in which learners might improve their performance on the same task (Bygate, 2018). In spite of an associated lack of research on the repetition of writing tasks, pedagogically speaking, recurring tasks are needed to assist learners in gaining mastery over various academic genres (Tardy, 2012). Furthermore, repetitions of genre-based writing tasks are a useful vehicle for observing longitudinal growth in second language writing ability. As Crusan (2010) aptly noted, the writing tasks teachers design should allow for the comparison of early writing samples to writing samples within the same genre produced later in the term. Ortega (2009) indicated that augmenting the field's understanding of the ways in which learners' written texts change over time should be of primary interest to applied linguists, and recurring genre-based writing tasks could serve as a mechanism for strengthening that knowledge, and in turn, for informing future educational action.

In particular, the field of TBLT has paid scant attention to the role of task repetition in promoting second language writing development. Although scholarship in the fields of TBLT and L2 writing would benefit from longitudinal investigations of the language learning potential of repeated written output practice (Manchón, 2011b), to my knowledge, the work of Nitta and Baba (2014, 2018) and Sánchez López (2018) are the only recent examples to explore this line of inquiry. For their part, Nitta and Baba (2014) engaged 46 Japanese university students in writing for 10 minutes on a familiar topic once a week over the course of 28 weeks. No linguistic correction was provided, and at the end of the study, participant writing was analyzed in terms of speed (as measured by text length), syntactic complexity (as measured by average sentence length), and lexical sophistication (as measured by word frequency values from the CELEX

corpus and the Measure of Textual Lexical Diversity). In addition, the researchers examined group-level and individual normalized trajectories. A multivariate analysis of variance that compared the quantitative measures of students' compositions in the first and final week of the school year showed significant increases in students' sentence length and lexical sophistication. Ultimately, the researchers concluded that while it was difficult to find a task that could provide a point of entry for a group of mixed-ability students, "the iterative opportunities of task engagement established a condition where the students engaged in meaningful production while increasingly paying attention to formal aspects in their own ways" (Nitta & Baba, 2014, p. 127). Although the writing task seemed disconnected from a recognizable genre, the work of Nitta and Baba (2014) informed that of Sánchez López (2018) who found that the written products of low proficiency learners in his study benefitted significantly in terms of accuracy from the opportunity to repeat the assigned task. Taken together, these studies are the few illustrations available of empirical research into the language learning potential of the recurrent production of written texts.

2.2.3.4 Social. In spite of the predilection for oral tasks in TBLT research, the written word is still an important means by which people interact both within and across cultures. Writing is a means of communication for various social networking tools (Reinhardt, 2019), including Twitter and WhatsApp, as well as an effective tool for exchanging ideas in academic and professional contexts. Hyland (2016) noted that writing is rarely an isolated activity, but rather an inherently dialogic action undergone with a potential or specific reader in mind. Sociality serves as a common thread linking diverse definitions of genre, and one aim of genre-based approaches to teaching L2 writing is to highlight the ways in which writers make linguistic choices based on an awareness of their intended audience. Since part of a writer's work is to

balance audience knowledge and expectations alongside a particular message (Hyland, 2011), an effective writing task should reference a specific readership, thereby connecting the activity and the student's text to a wider discourse community. Weigle (2002) maintained that a distinguishing feature of expert writer texts as opposed to novice writers' written work is the capacity "to anticipate the audience and shape a message appropriately in the absence of a conversation partner" (p. 18). Thus, writing instructors can help students to consider various reader perspectives, and in turn, to become more skillful writers, by specifying a real or imagined recipient for each new writing task (Crusan, 2010; Hyland, 2003).

One avenue for capitalizing on the social nature of written texts and on the affordances of Web 2.0 technologies has been to engage language learners in online collaborative writing tasks. To investigate the language learning potential of such collaborative work, Elola and Oskoz (2010) recruited eight university students studying Spanish as a foreign language to write two argumentative essays: one individually and one collaboratively. Interestingly, the researchers discovered that the students found it harder to engage in the iterative process of writing while collaborating with a partner. In addition, the researchers did not observe any significant differences in the accuracy or syntactic complexity of the individually- and the collaboratively-produced essays. Although Oskoz and Elola (2014) have attributed the benefits of collaborative writing to "the ways co-writers were able to interact with the text and negotiate global aspects of content development" (p. 123), adhering to an essay format with the teacher as the only audience may strip collaborative compositions of their social relevance. Collaborative writing tasks represent an area worthy of further investigation, as learner-learner interaction surrounding the content, structure, and style of a particular text may be a fruitful avenue for language learning. At the same time, Manchón (2011b) has noted that participants in collaborative writing studies

appear to use an inordinate amount of on-task time for negotiating the division of responsibilities and the overall structure of the text. Although jointly constructed texts may well provide language learners with additional linguistic experience, learner collaboration while composing is not a sufficient condition for labeling a particular writing task as "social." In fact, I have witnessed learners take responsibility for separate paragraphs or even for different parts of the writing process in an effort to avoid sustained learner-learner interaction through a collaborative writing task. Writing, on its own, is an intense linguistic activity that acquires social meaning when connected to genre-based approaches to teaching second language writing.

2.2.3.5 Varied. Scholars have suggested that encounters with diverse genres are important for exposing students to a variety of discourse structures and advancing their knowledge of different genre conventions (Tardy, 2012). According to Byrnes and Manchón (2014), "writing tasks should be made to encompass the spectrum from informal, short, simple writing tasks to more advanced, cognitively challenging tasks" (p. 8). In spite of these well-founded recommendations, the five-paragraph essay continues to dominate in L2 writing classrooms (Johns, 2018; Tardy, 2018) and research (Elola & Oskoz, 2010; Oskoz & Elola, 2014; Ruiz-Funes, 2014). A few years ago, Leki (2011) distributed a survey intended to gauge the genre knowledge of 84 international graduate and undergraduate students enrolled at an American university. Whereas almost all of the participants (94%) indicated familiarity with writing emails in English, for the undergraduates, the genres of much of their required coursework were unfamiliar. In addition, Leki reported that "the students seemed to equate essay writing with academic writing in English generally" (p. 95), a finding that should encourage teachers and program developers to consider exposing learners to multiple written genres. Although most of the students in Leki's study were open to tackling new writing tasks, if

learners were introduced to a variety of genres in preparation for college writing, Leki's findings would not necessarily be echoed in future investigations of university students' genre knowledge. Without ignoring the function of the five-paragraph essay, an effective curricular sequence could still incorporate writing tasks that engaged learners in composing within different genres.

Importantly, the researchers identified in this section have contributed to the advancement of the sub-disciplines of TBLT and L2 writing, and it is my hope that this consideration of scholarly-based recommendations for the design of writing tasks will continue to inform second language educational praxis and task-based research. Irrespective of the method chosen to analyze L2 written products, the role of the task itself, including its ability to engender a participant's best performance, should not be underestimated.

2.3 The Digital Turn

One vehicle for exposing learners to a variety of genres is through the incorporation of digital technologies in the second language classroom. Multiple researchers have highlighted the facilitative, reciprocal relationship between computer-assisted language learning and TBLT (González-Lloret & Ortega, 2014; Thomas & Reinders, 2010; Ziegler, 2016). The model of technology-mediated TBLT described in González-Lloret and Ortega (2014) encompasses the full integration of technology and tasks throughout the entire cycle of curricular design. Research utilizing this framework has focused primarily on computer-mediated communication (CMC), with investigations covering both text-based and oral modes of language production. Examples of studies exploring audio and video tasks include Winke's (2014) examination of video-based oral assessment tasks and Yanguas' (2010) research into differences in learner negotiation of meaning between audio-based CMC tasks and audio-video CMC tasks. The pursuit of viable

technology-mediated tasks has also led to investigations on the affordances of virtual worlds and gaming for fostering L2 development (Canto et al., 2014; Reeder, 2010; Sykes, 2014). In research that has investigated the role of text-based communication in technology-mediated TBLT, attention has been given to problem-solving tasks via text chat (Alwi et al., 2012), email exchanges (Alcón-Soler, 2018), and collaborative writing tasks (Elola & Oskoz, 2010). The nature of written communication through emerging digital genres is only beginning to be explored, even though scholars have acknowledged that the ability to correspond effectively through the global digital network is increasingly dependent on good writing skills (Godwin-Jones, 2018; Hyland, 2003).

2.3.1 Synchronous versus asynchronous computer-mediated communication

A characteristic often used to distinguish diverse online environments is the timing of user contributions. Synchronous communication refers to simultaneous user-user interaction in chat rooms or between players in an online game. On the other hand, asynchronous communication implies that there may be some delay in the exchange of written discourse. In today's society, individuals can select from an array of synchronous and asynchronous modes of communication according to their intended purpose, and the characteristics of such exchanges may differ depending on the length of the interval between publication and reception of a particular message. In a study exploring qualitative and quantitative differences between ESL students' written discourse in synchronous chats and their writing in asynchronous posts to a course discussion board, Sotillo (2000) found that the students produced more complex and more accurate prose in the asynchronous environment. The researcher also observed that the language of the text-chats contained various attributes of spoken discourse, a finding consistent with later research by Sauro (2012). On the other hand, Stockwell (2010) discovered that students were

more grammatically accurate during a task involving synchronous computer-mediated communication (SCMC) as opposed to asynchronous computer-mediated communication (ACMC). In this case, the researcher attributed the results to learner avoidance of more complex sentences in the SCMC task and to increased risk-taking in the ACMC environment. Bloch (2008) has suggested that asynchronous environments allow language learners to compose their texts at a more leisurely pace, potentially helping "students with weaker language skills to respond in ways that are just as complex as those of students with stronger language skills" (p. 131). Hyland (2003) arrived at a similar conclusion, and given the focus of this dissertation, the remaining sections of this chapter examine research on academic written communication in asynchronous online spaces.

Vandergriff (2016) reported that "many genres that predate the Internet have been adapted to online environments" (e.g., *e*-mail, advertisements, news articles, etc.); however, "many new genres have also emerged in online spaces" (p. 69). Three of these new genres include blogs, fandoms, and discussion boards, each of which has been harnessed in the creation of effective, asynchronous writing tasks. Blogs—essentially online journals upon which other Internet users may stumble and peruse (Blake, 2008)—provide an avenue for expanding the audience of language learners' written production and a space for ruminating over ideas. Fandoms, on the other hand, connect individuals who share an interest in a particular fictional series, music group, or pop-culture icon. Sauro (2014) explained that fandoms also "foster the sharing of responses to the source material, including the production of novel fan-generated content" (p. 239). Learners with prior connections to or a sincere interest in a particular series may find their fanfiction contributions praised by other readers, which in turn, could build the self-confidence and motivation of those same writers (Black, 2006; Vandergriff, 2016). Finally,

online class discussion boards, though open to a smaller audience, provide a forum for the written exchange of ideas and for the provision of peer feedback. Of these three emerging genres, online discussion boards and blogs are more widely used in tertiary education, and researchers have reported that learner engagement in such genre-based, online writing tasks may serve as a catalyst for promoting second language writing development (Byrnes, 2014; Chen et al., 2015).

2.3.2 Online Discussion Boards

In reality, only a few researchers have examined the role of online discussion boards in university curricula, but one of the groups to do so viewed students' online posts as a vehicle for fostering their written communication skills. For their research, Andon et al. (2018) conducted two studies. The first study involved 35 graduate students in a one-year degree program in Applied Linguistics for experienced ESL teachers. As part of the course in question, the instructors incorporated two writing tasks into each unit of study: one in which the students responded to pre-reading reflection questions, and one in which they responded to an academic article by relating the concepts covered to their professional experience. In total, the researchers collected 150 extracts of discussion board posts, and they developed a coding scheme based on themes that emerged from the data. The researchers concluded that the students successfully displayed evidence of critical thinking in the discussion forum; however, Andon et al. (2018) reported a perceptual mismatch between the expectations of the course tutors for student writing and the formality the students felt the context warranted. Whereas the tutors regarded the discussion board posts as a chance for students to practice their academic writing (similar to a register that would be used in a formal course paper), the students adopted a less formal, more personal style for their writing. Andon et al. commented that the students' interactional style

appeared “closer to the conventions of social networking sites” (p. 247). After the conclusion of the semester-long course, the researchers set out to investigate the ways in which students had engaged with feedback provided by the course tutors. Eight of the original 35 students volunteered to participate in semi-structured interviews geared toward extracting their opinions of the comments the tutors had made on their posts throughout the course. The researchers found that the students did not always respond to the tutors’ constructive comments in the manner in which they would have expected. Very rarely did students respond to the tutors directly, and in cases where the tutors made direct comments concerning grammar or style issues, the students understood them as suggestions for contexts in which they would use more formal academic prose.

2.3.3 Educational Blogs

Educational or academic blogs share a number of characteristics with online discussion boards. For instance, both digital tools provide an audience beyond the course instructor for individually crafted texts, thereby supporting the formation of a wider discourse community. Blogs and discussion boards may also promote student literacy practices, as learners are able to read and to comment on the posts of other authors (Bloch, 2008; Ducate & Lomicka, 2008; Godwin-Jones, 2006; Hindley & Clughen, 2018). In addition, both tools support the longitudinal collection of students’ written products. In particular, blogs—the simplified and more frequently used identifier for the original term, *weblogs*—store an author’s texts in reverse chronological order, allowing learners and teachers to observe development in student writing. The primary difference between online discussion boards and blogs lies in the extension of potential readership. Whereas the online discussion boards common to many educational contexts restrict the audience to the instructor and other participants in the course, various software applications

available for blogging offer students and instructors the possibility of opening the audience to anyone with access to a Web browser. As a result, blogs have been seen as a vehicle for promoting the online publication of original written text (Arslan & Şahin-Kızıl, 2010; Bloch, 2007) and for bypassing gatekeepers traditionally responsible for selecting the texts that will be made available to the wider public, such as publishing houses and media corporations (Bloch, 2007; Bloch, 2008). A second, though less obvious, difference between discussion boards and blogs is that whereas multiple authors contribute to the content of a discussion board, with the exception of reader comments, blogs are generally authored by a single individual, thereby allowing for greater authorial control (Chen, 2015).

One of the most widely cited pieces on the use of blogs in second language studies is a case study documenting an ESL student's utilization of a course blog (Bloch, 2007). The participant was enrolled in a lower-level composition course at a four-year U.S. university, and the instructors of the course intended the blogs as a means of fostering students' academic writing skills. In spite of the anecdotal quality to the finding that the blog had enabled this learner to develop his critical thinking skills, the author raised some important points in the concluding paragraphs. First, potentially as a result of instructing students to pay minimal attention to grammar, Bloch (2007) did not uncover evidence that the series of blog posts had enabled the learner to gain greater control over syntax. Second, the author acknowledged a need for a thorough examination of the blog post genre and of the ways in which communication via this electronic medium might contribute to learners' development of academic writing skills. .

Following publication of Bloch's (2007) case study, a number of scholars have employed similar ethnographic and action-research techniques to investigate the use of blogs in university-level second language classrooms (Arslan and Şahin-Kızıl, 2010; Chen, 2015; Chen et al., 2015;

Ducate & Lomicka, 2008; Elgort, 2017; Pinkman, 2005; Sun, 2010; Sun & Chang, 2012; Vurdien, 2013). Of these studies, Ducate and Lomicka (2008) focused on the learning of French and German at an American university, and Elgort (2017) concentrated on comparing the writing of L1 and L2 English speakers in an ESL context. The rest of the studies came from EFL contexts in Japan (Pinkman, 2005), Spain (Vurdien, 2013), Taiwan (Chen, 2015; Chen et al., 2015; Sun, 2010; Sun & Chang, 2012), and Turkey (Arslan and Şahin-Kızıl, 2010). Ducate and Lomicka's (2008) project involved a year-long blogging project they had implemented in a third-semester German course and a fourth-semester French course at a U.S. university. During the first semester, the students were involved in following a blogger of their choice, and in the second semester, the students transitioned into blog writers, posting to their blogs once a week and then commenting on two of their classmates' posts. Student questionnaires and a post-course focus group revealed various perceived benefits of the blogging project, such as growing a community of learners, being able to express feelings, and sensing an improvement in their writing.

The other blog-focused study carried out in the United States was Elgort's (2017) research with first and second language English speakers in an Applied Linguistics course. At the end of the course, the researcher selected two blog posts and two essays written by each student and subsequently calculated a variety of lexical and syntactic sophistication indices, which served as the dependent variables in a multivariate analysis of variance. Post-hoc analyses revealed statistically significant differences between the independent variables, text type and language background, on several Coh-Metrix indices. Whereas the essay texts demonstrated a higher degree of lexical sophistication and syntactic complexity, in the blog posts, students used more personal pronouns, and there was more semantic overlap across paragraphs. In additional

to the linguistic differences between texts, two raters scored both the essays and the blog posts according to a scale reflective of Coh-Metrix's levels of analysis. Interestingly, although texts written by the L1 English participants received similar scores regardless of text type, blog posts written by the L2 English students were scored significantly lower than their essays. Elgort surmised that the "L2 writers may have found moving away from the familiar impersonal academic writing style to a relatively unfamiliar style of education blogging challenging" (p. 65). As a result, Elgort suggested that ESL students could benefit from opportunities to practice educational blogging, which would allow them to rehearse the communication of complex ideas to a wider audience.

In EFL contexts, a number of studies have piloted blogging tasks in in-tact classrooms. Sun and Chang (2012) incorporated blogs as a means of encouraging participants to reflect on the development of their academic writing skills in English. Students were allowed to post in English or Chinese, with Chinese being the language used most frequently (in 67% of cases). Based on an analysis of the topics addressed in learner posts, Sun and Chang concluded that the course blogs promoted social interaction and helped the students to grow as writers and authors. Chen et al. (2015) determined that closed, more structured blogging tasks, as opposed to open-ended prompts, encouraged participants to produce more "idea units," and Chen (2015) reported that the aspects of blogging most valued by the EFL students were notes the teacher left concerning language and the subsequent opportunity to self-edit. Similar to the method adopted in Elgort (2017), Sun (2010) analyzed participants' first three and final three blog posts, out of a total of approximately 30 posts per participant written over the course of a semester. Sun (2010) found that students used simpler syntactic structures in their final posts, as measured by total clauses, total subordinate clauses, and length of T-unit; however, raters scored the final texts

higher on a five-category rubric covering grammar, vocabulary, mechanics, fluency, and organization. Finally, Arslan and Şahin-Kızıl (2010) compared two groups of Turkish university students, with one group (*N*=23) receiving process-oriented writing instruction and the other group (*N*=27) contributing to course blogs in addition to covering the same material. Based on the scores assigned to student pre- and post-test essays, the treatment group significantly out-performed the non-treatment group, leading the authors to conclude that the utilization of blogs had a positive impact on the development of learners' academic writing.

Of all the studies covered in this section, Arslan and Şahin-Kızıl (2010) were the only researchers to include a control group. At the same time, Arslan and Şahin-Kızıl (2010) acknowledged that in addition to blogging, the experimental group also had access to electronic versions of course materials, resources to which students in the control group did not have access. Golonka et al. (2014) and Sun (2010) have bemoaned the scarcity of experimental studies on the impact of blogging on language development. Even though research on the use of blogs in second and foreign language contexts has contributed to the field's understanding of the genre, there are no empirically-based guidelines for composing effective academic blog posts. Notwithstanding this gap in the literature, educational blogging has become a required component of many university-level English courses. In Sun (2010), 20% of the students' final grade was based on fulfillment of a required number of blog entries and comments on other learners' posts. Similarly, in Pinkman (2005), 20% of the course grade comprised students' participation in blogging. The criteria for grading consisted of meeting the expected 150 words per week and incorporating language that had been used in class. Although other studies supply fewer details around the grading of student blogs posts, Chen (2015), Chen et al. (2015), Ducate and Lomicka (2008), Elgort (2017), and Sun and Chang (2012) all mentioned that the blogging

assignments were a required component of the language course. Furthermore, a variety of assessment formats have been employed to judge successful handling of this genre. Whereas Sun (2010) and Sun and Chang (2012) gave students full credit as long as they had written a post, Chen et al. (2015) intimated that learners' posts were graded based on the amount of dedication they had demonstrated in completing the task. The rubrics mentioned in Arslan and Şahin-Kızıl (2010) and Elgort (2017)—Jacobs et al.'s (1981) ESL composition profile and a rubric aligning with Coh-Metrix's levels of analysis, respectively—were only used after the fact to measure changes in student writing. The only two studies to include blog-specific scoring criteria were Pinkman (2005) and Ducate and Lomicka (2008), and to my knowledge, neither set of criteria has been employed in later research. Seeing as blog-based second language writing has become a factor in learners' academic success, it is even more important to investigate ways to assess student performance within this genre in a fair and valid manner.

2.4 Measuring Second Language Writing Development

To measure the effects of a pedagogical intervention such as a blog, researchers have often utilized indices of linguistic complexity (Elgort, 2017; Sun, 2010). According to Bulté and Housen (2014), in second language research, "complexity has been proposed as a valid and basic descriptor of L2 performance, as an indicator of proficiency and as an index of language development and progress" (p. 43). Through their study of the written texts of 45 adult ESL learners at the beginning and at the end of a four-month academic English program, Bulté and Housen sought to identify reliable measures of lexical and syntactic complexity that were sensitive to changes in L2 writing quality. Specifically, the researchers examined the relationship between quantitative measures of lexical and syntactic complexity and rater-assigned scores of participant essays. The results of their analyses led Bulté and Housen to conclude that

"complexity measures can capture changes in L2 writing ability and quality over time, including over relatively short periods of time such as the ones typically afforded by academic L2 (writing) proficiency courses" (p. 55). Both in Bulté and Housen's research and in other studies connected to Applied Linguistics and Corpus Linguistics research, the notion of linguistic complexity has been sub-divided into lexical and syntactic complexity, concepts that have been refined even further over the past several decades.

Literature in Applied Linguistics has supplied multiple formulas for calculating syntactic complexity. Four traditional measures of syntactic complexity included mean length of T-unit, mean length of clause, clauses per T-unit (C/TU), and the ratio of dependent clauses to total number of clauses in a text (DC/C) (Wolfe-Quintero et al., 1998). The two measures of subordination, C/TU and DC/C reappeared in Ortega's (2003) meta-analysis of syntactic complexity measures, and Lu (2011) found that the number of dependent clauses per total number of clauses reliably discriminated among participants' school levels. At the same time, Lu (2011) discovered a non-linear path of development for DC/C in that values increased between Level 1 and Level 2 and then decreased from Levels 2 to 4. Even though research in the area of syntactic complexity has continued to investigate new indices, in their longitudinal examination of second language writing, Kyle et al. (2021) determined that DC/C was "the clearest and most accurate indicator of longitudinal growth" (p. 18).

Lexical diversity and lexical sophistication are often subsumed under the concept of *lexical richness*, which has been defined as "the level of development of a learner's lexicon" (Cobb and Horst, 2015, p. 189). According to Cobb and Horst (2015), the most basic measure of a text's lexical diversity is the type-token ratio (TTR), which amounts to dividing the number of different words in a text by the total number of words in that same text. Since the type-token

ratio is sensitive to text length, meaning that words are more likely to be repeated as the number of total words increases (Jarvis, 2013; McCarthy & Jarvis, 2010), researchers have sought to identify a ratio that could take into account text length. Using a corpora consisting of 4,542 argumentative essays written by second language learners of English, Zenker and Kyle (2021) investigated nine indices of lexical diversity, including traditional measures such as TTR and more recent indices attempting to minimize the effects of text-length. Their analysis found that of the nine indices, a measure developed by Covington and McFall (2010), the moving-average type-token ratio (MATTR), was the most stable index, showing minimal correlation with the length of participant texts. As indicated by the measure's name, this index computes an average TTR for a text, first by setting a "window" of say 50 words and then by calculating the TTR for words 1 through 50, for words 2 through 52, and so on until the end of the text. Next, the mean of these moving TTR values is computed in order to provide a single value for the index.

Another measure of lexical richness, lexical sophistication, concerns a learner's use of more advanced, or sophisticated words. In contrast to lexical diversity, which centers on lexical repetition, indices of lexical sophistication are often based on ratios of the number of *sophisticated* words within a particular text. Different researchers have employed distinct definitions of sophisticated words depending on the discipline to which they belong (Kim et al., 2018). One area of lexical sophistication has focused on measures of word frequency, with researchers reporting that more proficient users of a second language are likely to incorporate less frequently encountered words in their written production (Crossley & McNamara, 2012). In addition to word frequency, researchers have investigated the concept of lexical decision time as a measure of lexical sophistication (Balota et al., 2007). In Balota et al. (2007), L1 English participants were asked to determine whether a series of letters presented on a computer screen

amounted to a word or a non-word, and the mean length of time participants took to identify a particular word became the value associated with that word. Within corpus research, words associated with longer reaction times "are considered more sophisticated than words that elicit shorter response times" (Kim et al., 2018).

More recent measures of lexical sophistication have centered on contextual distinctiveness, which concerns the degree to which a particular word is free to appear in different contexts. The corpus-based index, McD (for McDonald), provides measures of contextual distinctiveness, which were calculated using the British National Corpus. Words that appear in fewer contexts, such as *amok* in "run *amok*," are assigned a higher score for contextual distinctiveness, and words such as "clean," which can co-occur with many different words, receive lower scores (McDonald & Shillcock, 2001). The behavior-based index, USF (for the University of South Florida), also provides values related to probabilities of co-occurrence; however, this index was created on the basis of the number of words that participants were able to produce in relation to a particular cue (Nelson et al., 2001). In contrast to McD, for USF, a higher value for a particular word would indicate lower contextual distinctiveness, and a lower value would correspond to greater contextual distinctiveness. Both measures of contextual distinctiveness, McD and USF, appeared in Kyle and Eguchi's (2021) five-component regression model accounting for 23% of the variance in rater scores of argumentative essays. In addition to McD and USF, which provide information about lexical collocation behavior, two measures of the strength of association between words serving a particular grammatical function were included in the model. These measures, verb-direct object dependency bigrams and noun-adjective dependency bigrams, are calculated using the magazine and newspaper subsections of the Corpus of Contemporary American English (Davies, 2010).

In sum, measures of linguistic complexity play an important role in Applied Linguistics research. Not only are they useful for measuring changes in learners' productive capabilities, but also they can provide valuable information concerning the impact of a particular pedagogical intervention. However, much research in corpus linguistics utilizes the traditional essay as the basis for comparing learner production pre- and post-treatment or as the source material for identifying relationships between rater-assigned scores and measures of linguistic complexity. Therefore, this study attempts to apply well-supported indices of lexical and syntactic complexity to the investigation of longitudinal written development in the emerging genre of the educational, or academic blog post.

CHAPTER 3: ASSESSMENT OF SECOND LANGUAGE WRITING

Deeply connected to second language writing pedagogy is the assessment of writing, a task that some scholars have identified as one of the most important undertakings within a teacher's sphere of responsibilities (Crusan, 2010). Ferris and Hedgcock (2014) corroborated that "assessment is an essential *teaching* task" (p. 197, emphasis in original), and Casanave (2017) explained that many of the decisions made by first and second language writing teachers "revolve around assessment of students' writing" (p. 223). Given the centrality of assessment in the teaching of L2 writing, it is surprising that a number of graduate programs in language teaching do not require students to complete a course in language assessment (Weigle, 2007). Crusan et al. (2016) also found that nearly 80% of the writing instructors they surveyed (N=702) reported feeling inadequately prepared to design rubrics for assessment purposes. Thus, a discussion of the relevance, creation, and use of sound assessment instruments is both necessary and timely. The process of evaluating writing, especially when performed on the basis of carefully designed analytic rubrics, can contribute to a validity argument in favor of the proposed use of such assessment mechanisms and promote positive washback in the form of instructional and curricular adjustments.

3.1 Feedback

One form of teacher response to student writing involves *feedback*, which Wiggins (1998) defined as "information about how a person did in light of what he or she attempted—intent versus effect, actual versus ideal performance" (p. 46). Prior to the widespread adoption of the process approach to writing (Hyland, 2003), second language writing instructors would include a few remarks *only* on the final piece. According to Andrade and Evans (2013), those remarks were used more to substantiate the score the instructor had assigned rather than to

support the development of the student writer. For example, Ferris (2014) surveyed 129 tertiary-level writing instructors in Northern California to investigate their feedback practices. The researcher also interviewed 23 of the survey respondents, asking follow-up questions designed to illuminate issues in response practices that may not have come through on the electronic survey. For Ferris, a striking result of the survey and interview data was that many participants still restricted their feedback on student work to the final product. Despite these instructors having reported that they engaged students in peer review workshops, in many cases, "written feedback was only given on final, graded essays" (Ferris, 2014, p. 13). In other words, whereas two-thirds of the interview participants articulated beliefs surrounding the potential of teacher feedback for promoting student progress in writing, Ferris concluded that there was still work to be done in helping teachers to engage in feedback practices that would enable students to take greater advantage of that feedback. Rubrics, which were used extensively among the surveyed participants, could be one way for teachers to realize their instructional goals.

Inherent in the notion of teacher feedback is an *assessment*, or judgement, of the quality of a student's written work compared to a particular standard, which may or may not be articulated clearly. Messick (1996) affirmed that assessment "always involves, even if only tacitly, intervening processes of judgment, comparison, or inference" (p. 244). Explained differently, assessment in language education includes generating inferences about our students' abilities and then making decisions based on those inferences (Bachman & Palmer, 2010; Bailey, 1998; Weigle, 2002). Inferences are necessary because even though some scholars have made a distinction between 'direct' and 'indirect' tests of writing (i.e., actual writing samples vs. multiple-choice or discrete-point items), the term 'direct' is somewhat of a misnomer "since any test is at best an indirect indicator of an underlying ability" (Weigle, 2002, p. 59). Nevertheless, a

teacher's assessment of student writing can still provide useful information on learning and on learners (Norris, 2006). This information may be collected at the classroom or program level for diagnostic, progress, and achievement purposes (Brown, 1998). Further distinctions are frequently made between formative and summative assessment activities (Graves, 2000), the latter often being associated with achievement tests administered at the end of a course to ascertain student learning (Brown, 2005). Formative assessment, on the other hand, "focuses on letting students know where they have been, where they are, where they need to be, and how they can get there" (Matsuda, 2012, p. 155). Irrespective of the label a teacher assigns to a particular assessment activity, a key takeaway is that the information obtained as a result of the assessment of student writing can be useful to *both* teachers and learners.

3.2 Rubrics

Teachers have many options for providing feedback on and assessing language learners' writing, including written comments in the margins or in endnotes, verbal comments delivered face-to-face, online, or through a recorded message, and utilization of rubrics. Scoring rubrics, defined as "a set of categories, criteria for assessment, and the gradients for presenting and evaluating learning" (Brown, 2012b, p. 1), often accompany writing tasks that require students to compose some sort of constructed or extended response. The nature of the response will vary (e.g., a cover letter or a summary), but rubrics can be used to assess students' ability to write effectively within a particular genre. Entire books detail the role of rubrics in general education classrooms (Arter & McTighe, 2001; Stevens & Levi, 2005), and rubrics factor prominently in academic texts concerned with the teaching of second language (L2) composition (Ferris & Hedgcock, 2014; Hyland, 2003), language assessment (Brown, 2012a), and the assessment of L2 writing (Crusan, 2010; Weigle, 2002). In spite of their prominence in many educational contexts,

teachers sometimes "find it difficult to initiate the grading process, often because they fear making judgments that may be biased, unjustified, and possibly damaging to student writers" (Ferris & Hedgcock, 2014, p. 227). In addition, Casanave (2017) explained that one dilemma teachers face in assessing their learners' writing is a dearth of reliable, suitable instruments with which to approach the evaluation of student texts. Particularly in cases of student writing in genres beyond the five-paragraph essay, the scarcity of assessment mechanisms is more perceptible (Crusan & Ruecker, 2019). Recommendations in this area range from adapting pre-existing rubrics (Crusan, 2010) to conducting an extensive genre analysis (Hyland, 2004). While the former recommendation often functions as a "quick-fix," the latter suggestion—although sound and theory-based—may be out of reach for new teachers working to manage a plethora of responsibilities, including planning lessons, complying with administrative requirements, and providing meaningful feedback on student work. According to Popham (1997), a rubric that would be helpful for both teachers and learners would include three to five evaluative criteria representing key attributes of the writing skill under examination. Whereas a rubric focusing on less than three features might unwillingly promote a reductive definition of academic writing, human raters may struggle to distinguish among a relatively large number of attributes. In the evaluation of English as a second language writing, these attributes have typically included content, organization, vocabulary, language use, and mechanics (Jacobs et al., 1981; Trace, Janssen, & Meier, 2017; Winke & Lim, 2015).

3.2.1 Holistic Versus Analytic Rubrics

Potentially the most widely discussed issue related to rating scale design is the distinction between holistic and analytic rubrics. Holistic rubrics require raters to assign a single score to a particular piece of writing; however, "a major shortcoming of the holistic rubric is that it does

not take into account the possibility that different aspects of the performance may vary in quality" (Davis & Kondo-Brown, 2012, p. 34). For example, when a student's written product demonstrates a strong grasp of content but lacks the advanced vocabulary expected for that genre, a rater is forced to choose a score that will be an untrue indicator of at least one attribute of said writing performance. Another shortcoming of holistic rubrics is that the single scores resulting from their use have been found to correlate with measures of handwriting legibility and text length, characteristics generally assumed external to an ecologically-valid conception of writing ability (Weigle, 2002). Although holistic rubrics are often preferred for large-scale assessments of writing, as they may expedite the rating process (Crusan, 2010), analytic rubrics, which enable raters to assign scores for distinct features of the written product, are likely more useful for language learners in that they provide detailed feedback on various dimensions of a student's performance (Brown, 2012a; Ferris & Hedgcock, 2014; Shohamy, 1992). Feedback provided by means of an analytic rubric can aid teachers in diagnosing the strengths and weaknesses of students' written production, and it can provide students with a more nuanced view of genre-specific expectations. Finally, whereas a holistic scale that contains only a few score points (e.g., zero to five on the iBT/Next Generation Test of English as a Foreign Language Independent Writing Prompt) may not provide repeat test takers with the type of improvement evidence they seek, analytic rubrics could show test takers areas in which their writing has improved, even if those improvements do not correspond to the crossing of a particular threshold.

Barkaoui (2010) conducted a study aimed specifically at examining the variability in rating processes in terms of rater experience and the type of rating scale—holistic or analytic. The researcher used think-aloud protocols to investigate the rating processes of 25 raters (11

novice and 14 experienced evaluators of L2 writing) selected randomly from a larger pool. A counterbalanced design ensured that half of the raters began with the holistic rubric and the other half began with the analytic rubric. In total, each rater scored 48 essays written by adult English language learners in response to an independent writing prompt. For the 48 essays, each rater scored 12 essays silently using the analytic scale, 12 essays while thinking aloud using the analytic scale, 12 essays silently using the holistic scale, and 12 essays while thinking aloud using the holistic scale. Barkaoui coded the think-aloud protocols according to rater decision-making behavior. Subsequently, the frequencies of each action were subjected to Wilcoxon Signed-Ranks tests that identified statistically significant differences between the two rating scales. The article did not specify the number of Wilcoxon Signed-Ranks tests run on the data, thereby casting doubt over the number of differences found significant at $p < .05$; however, the author reported that the holistic scale elicited more interpretation strategies during rating, while the analytic scale prompted more judgment strategies. In addition, Barkaoui explained that when using the analytic scoring rubric, both groups of raters (experienced and novice) referred to the scale and articulated and justified scores more frequently. These results led the researcher to conclude that with analytic rubrics, raters paid more attention to rating scale criteria and were also more internally consistent in their ratings. Barkaoui also noted that in general, the raters preferred analytic scoring, as they were not forced to label compositions with only one value.

3.2.2 Task-dependent Versus Task-independent Rubrics

Scholars and practitioners have also made distinctions between task-dependent and task-independent rubrics. Whereas task-dependent rubrics include criteria referencing the degree to which a test-taker has accomplished successfully a particular task, “task-independent rubrics focus on more general linguistic abilities that are thought to play a role in performance across a

domain of tasks or form a part of language ability generally" (Davis & Kondo-Brown, 2012, p. 37). One benefit of task-dependent rubrics is the extent to which scale descriptors can refer directly to specialized vocabulary in the prompt or to specific rhetorical moves. Such specificity may, in turn, provide more scaffolding for learners as they work to accomplish the task effectively. At the same time, the more closely a scale is linked to a particular task, the less suitable that rubric may be for evaluating student performance on a different writing task (Upshur & Turner, 1995). The versatility of a scoring rubric would be of particular concern to scholars who conceptualize second language writing ability as a single construct, independent of task conditions and context of use. From this perspective, a task-dependent rubric would lack generalizability, and in turn, validity, as any conclusions drawn about a writer's ability in that scenario might not apply to his or her performance in another situation or in another genre. On the other hand, a task-independent rubric, though appropriate from a unitary perspective of writing ability, might fail to account for the nuances observed in genre-sensitive texts.

As part of a comprehensive research project that led to the creation of a battery of task-based assessments, Norris et al. (2002a) employed three university ESL teachers to rate examinees on two scales: a holistic, task-dependent scale, and a holistic, task-independent scale. The researchers discovered that both sets of ratings "resulted in predicted systematic performance differences among four levels of English language users" (p. 410). In other words, Norris et al.'s results suggest that task-dependent and task-independent rubrics are equally effective in distinguishing among levels of writing ability. Consequently, researchers and teachers can adopt the rating scale that aligns most closely with their theoretical understanding of second language writing ability.

3.2.3 Criticisms of Rubrics

Rubrics are not the only method available to teachers for providing feedback on student writing, nor are they without their detractors. Scholars from a variety of disciplines have criticized the quality of rubrics in use in many institutional contexts (Popham, 1997, Wilson, 2007). Popham (1997), for example, stated that "the vast majority of rubrics are instructionally fraudulent" (p. 73), as the evaluative criteria fail to describe the skill being measured, are too broadly defined, or are overly detailed. Wilson (2007) reported that generic rubrics did not always account for nuances in students' written work, and other scholars have suggested that the use of rubrics reinforces dominant hierarchies and stifles creativity (Ashby-King et al., 2021). The first criticism can be addressed through increasing the assessment literacy of language practitioners, and the second criticism can be countered by using actual student writing samples as the basis for rubric creation. Arising from theory connected to critical pedagogy, proponents of a third set of criticisms have suggested that helping learners to identify characteristics of widely used written genres only strengthens the existing power structures (Hyland, 2004). In response to this critique, Hyland (2004) has argued that explicit awareness of the features of various gatekeeping genres serves as a necessary prerequisite to critical analysis of those discursive modes. Like genres, rubrics impose certain constraints; however, they do allow learners to express their individuality in the way they meet those genre expectations. In fact, well-designed analytic rubrics render the constraints apparent, thereby freeing learners to make concerted choices within a textual or electronic space that is still recognizable to their intended readership. Ultimately, "the types of verbal/descriptive feedback teachers can provide with rubrics-based assessments are so much more useful than simply and curtly presenting students with a numerical score" (Brown, 2012a, p. 31). The well-trained language teacher is likely to see

the assessment of student writing and the creation of analytic rubrics as essential components of his or her professional expertise.

3.3 Rubric Creation

The Conference on College Composition and Communication (2009a) recommends that instructors provide learners with rating scales that outline the criteria upon which student work will be assessed. In the words of Ferris and Hedgcock (2014), "we cannot avoid judgment, but we can make it transparent and useful for our students" (p. 228). Analytic rubrics communicate teacher expectations for a particular task, and the development of rating scale criteria is a vehicle through which writing instructors and researchers can articulate their understanding of the construct under examination. At the same time, Al-Hoorie and Vitta (2019) and Weigle (2016) have remarked that the process of developing rating scales is rarely addressed in literature, and in some cases, the rubric used to evaluate student texts is not provided, thereby making it impossible to replicate the study. In research articles that do address the manner in which a particular rubric was designed, two traditions dominate: one in which rating scale criteria are developed by knowledgeable experts (East, 2009; Kuiken & Vedder, 2017) and one in which descriptors are crafted on the basis of sample texts (Fulcher, 1996; Turner & Upshur, 2002; Upshur & Turner, 1995). Neither perspective has escaped criticism, with expert-centered designs being faulted for imprecise or irrelevant descriptors and with data-based scales receiving censure for being a-theoretical. In the case of performance assessments, Turner and Upshur (2002) have argued that empirically-derived rubrics are both relevant and theory-based in that "they represent a particular instance of a more global language proficiency theory" (p. 53). For these reasons, this research team along with other scholars in the field (McNamara, 1996) have consistently recommended that criteria included in rating scales be based on empirically supported

descriptions of target language use, rather than on theorists' expectations for a range of written products.

A commonly cited technique for developing rating scales on the basis of actual student texts is the empirically derived, binary-choice boundary-definition (EBB) procedure outlined in Upshur and Turner (1995) and explored further in Turner and Upshur (2002). The first step in this design involves specifying the number of levels of achievement among which writing samples will be distributed. Next, a team of individuals sorts a series of texts into different levels. The group then considers the dissimilarities among texts located in distinct performance levels, ultimately agreeing on a set of binary questions that could be used to sort additional samples. In spite of this seemingly straight-forward procedure, Turner and Upshur (2002) discovered that rubrics created on the basis of different sample texts had a significant effect on the rating of subsequent texts, and even in cases where development teams were provided with the same set of sample texts, there was little correspondence between respective scale descriptors. These reservations around the EBB process, coupled with the requirement that a group of raters come to a consensus on the binary questions to be utilized in the scale's development, suggest a need for further investigations of procedures through which stakeholders can create user-friendly rubrics on the basis of sample texts.

To my knowledge, the only large-scale project employing a data-based approach to the design of a writing rubric is the team of Diederich et al. (1961) in partnership with the Educational Testing Service (ETS). For their study, the researchers collected 300 essays written by liberal arts freshmen enrolled at three prominent Northeastern universities. Then, 53 faculty readers representing several fields of research independently sorted the essays into nine piles according to merit. The readers were also instructed to comment on all 300 papers, with the only

guideline for both steps being to "use your own judgment as to what constitutes 'writing ability'" (Diederich et al., 1961, p. 11). To identify the various constructs according to which the readers based their rankings of student essays, Diedrich et al. conducted a factor analysis on the pile numbers attached to each text. Next the researchers identified the three readers who had the highest loadings on a particular factor, and an assistant combed through every notation made by those individuals in order to pinpoint specific 'categories of comments' that reflected the most frequent observations of those three readers. In spite of the acknowledged limitation of solely calculating comment percentages (i.e., tabulating without prejudice longer written comments and marks indicating spelling or punctuation errors), the researchers came to label their factors ideas (including relevance, clarity, quantity, and development), form (organization), flavor (style, interest, and sincerity), mechanics (punctuation, spelling, and grammatical errors), and wording (choice and arrangement of words). Ultimately, the work of Diedrich et al. was recognized as the origin of the traditional, five-category writing rubric (Crusan, 2010).

3.4 Raters

Whereas investigations centered on the development of rubric criteria appear rather infrequently, a great deal of research in language testing has examined the behavior of the individuals responsible for applying scoring criteria to individual texts. These individuals, the *raters*, may or may not have been involved in the creation of a particular rubric, yet they must use the information on the scale to assign scores to examinees' written products. Of primary interest to researchers is whether or not human raters are consistent—both internally and in relation to other raters—in scoring student texts. The consistent application of rating criteria across diverse sample texts, or even across different tasks, is often seen as a necessary, if not sufficient, condition for advocating for a rubric's use in a particular context (Wind & Peterson,

2018). Interestingly, in many studies, the rubric itself escapes criticism, and observations are directed toward the human—and 'less reliable'—component of the rating process. In other words, frequently the point of interest is not the quality of the rating scales, instruments that Lumley (2002) has defined as "inevitably of limited validity, because of their inability to describe texts adequately" (p. 268), but rather the raters' ability to reconcile scale descriptors with features of the texts under evaluation.

An extension of the discussion on rater consistency is an examination of rater bias toward particular categories, tasks, or test-takers. To investigate potential rater bias in assessing examinees' written performance on a department-level placement test, Kondo-Brown (2002) recruited three experienced teachers of Japanese to rate the Japanese compositions of 234 university students. The 234 essays were scored by each of the three teacher-raters using a modified version of the ESL Composition Profile (Jacobs et al., 1981). Following the assignment of scores according to the adapted analytic rubric, Kondo-Brown conducted a detailed investigation of rater bias using a many-faceted Rasch analysis. The results of the analyses showed that the raters differed in terms of overall severity, and that each rater had a unique bias pattern. For instance, one rater scored vocabulary more harshly than the other rubric categories, while another rater was harshest on content and a third rater on mechanics. Another interesting pattern was that "a much higher percentage of significantly biased interactions was found for the candidates with extremely high or low abilities" (Kondo-Brown, 2002, p. 24).

3.4.1 Rater Training

Training raters prior to the scoring of written products is a familiar practice in the field of Applied Linguistics. In fact, most studies examining the process by which raters make scoring decisions conclude with recommendations for continued rater training surrounding the

application of rubric criteria to individual writing samples (Kang et al., 2019; Lumley, 2002). At the same time, research that has investigated the impact of rater training has delivered mixed results on its effectiveness. For example, Weigle (1998), analyzed the scores assigned by new teaching assistants as well as experienced raters to compositions written by ESL students at UCLA. The researcher had the newer raters and the more experienced raters score different subsets of 15 essays prior to the treatment. The treatment consisted of a rater training session of approximately 90 minutes, and one to three weeks after the session, the raters scored different subsets of 16 essays, four of which were also included in the original subsets. A subsequent FACETS analysis of the data revealed that on the pre-training scoring, the more experienced raters were more consistent, which was not especially surprising. In reviewing the post-training data, the researcher found that overall, the raters were sharper in distinguishing among participant ability levels; however, major differences in rater severity remained. Weigle suggested that the results supported the notion that while training may not help raters to score more in line with one another, it may help them to be more self-consistent.

3.5 Rubric Validation

Validity—potentially the most important issue in measurement (Furr & Bacharach, 2014)—is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA/APA/NCME, 2014, p. 11). In other words, validity involves judging the appropriateness of an inference made on the basis of a learner's test score and of subsequently using that inference to support a decision concerning the individual. Validity is not inherent to a particular assessment instrument, such as the holistic rubric used to score essays produced in response to the independent writing prompt on the Test of English as a Foreign Language (TOEFL), but rather it concerns both the interpretations made about a student's

writing ability based on his or her score *and* the ways in which a language program or institution uses that score. Brown (2005) explained that "teachers certainly want to base their admissions, placement, achievement, and diagnostic decisions on tests that are actually testing what they claim to measure" (p. 220). At the same time, the degree to which a student's score on a test of writing serves as an accurate indicator of his or her L2 writing ability is open to debate. As a result, the onus is on the test or rubric developer as well as on the user to gather support in defense of interpretations and decisions made on the basis of those assessment results.

Validation is an ongoing process (Messick, 1995) that begins with explicit statements concerning "the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA/APA/NCME, 2014, p. 11). As a logical next step, stakeholders should gather empirical evidence in support of those propositions. Per unitary conceptualizations of validity (AERA/APA/NCME, 2014; Messick, 1995), the empirical support serves as evidence for construction of a validity argument, not as a demonstration of different *types* of validity (i.e., content, criterion, and construct validity). It is tempting to assume that a self-contained writing sample would, in and of itself, provide the necessary validity evidence to justify making inferences about the author's second language writing abilities. However, support for intended interpretations and uses of a particular assessment generally requires multiple sources of validity evidence, including evidence based on test content, evidence based on response processes, evidence based on relations to other variables, evidence based on internal structure, and evidence based on consequences of testing (AERA/APA/NCME, 2014). A minimum threshold regarding the amount of evidence needed to make a strong validity argument does not exist, and "the quality and quantity of evidence sufficient to reach this judgment may differ for test uses depending on the stakes involved in the testing" (AERA/APA/NCME, 2014,

p. 13). Nevertheless, a careful consideration of each type of validity evidence may assist in demystifying the test validation process. The discussion that follows adopts the five types of validity evidence as a framework for validating the use of rubric-based measures of L2 writing.

3.5.1 Evidence Based on Test Content

Both in classical and in modern test theory, the items included on a test are assumed to be a mere sample of the universe of items or tasks available for measuring a particular construct. Therefore, one goal of the validation process should be "to establish an argument that the test is a representative sample of the content the test claims to measure" (Brown, 2005, p. 221). One avenue for substantiating test content is through drafting test specifications. These specifications would identify the construct(s) being assessed and, in the case of productive skills, the number of samples needed to assess adequately an examinee's degree of mastery of the construct (Alderson et al., 1995). In cases of content underrepresentation on a test of second language writing, the potential interpretation of student scores would be narrowed considerably, seeing as inferences related to second language writing ability would be restricted to the type of writing they were asked to produce (AERA/ APA/NCME, 2014). For example, if a test of L2 writing failed to include a variety of writing tasks or ignored common modes of written communication, the test developer(s) and user(s) would need to consider whether or not scores obtained on that assessment could be used to draw broader conclusions about the learners' underlying writing ability in different contexts. By extension, a student's score on a writing sample composed under timed conditions may not necessarily allow for inferences regarding his or her ability to write a successful, out-of-class term paper.

Conducting a needs analysis is one method for uncovering evidence that a set of writing tasks adequately represents the performance domain. Seeing as "the ability to function

competently in a range of written genres is often a central concern for ESL learners" (Hyland, 2004, p. 43), a needs analysis conducted in an ESL or EFL context ought to focus on identifying the types of texts students will need to produce in their target context. A needs analysis generally involves an assortment of data gathering techniques (Long, 2015), an example of which is recounted in Brown's (2005) comprehensive guide to language assessment. In the text, Brown describes a project undertaken as a graduate student at the University of California Los Angeles in which he set out to design a test measuring the English reading and listening abilities of non-native English speakers enrolled in the university's engineering program. Brown explained that the research team "soon discovered that nobody had any idea what the components of engineering-English reading ability might be" (p. 225). As a result, they met with engineering professors, reviewed the textbooks currently in use in a sophomore-level engineering class, and examined videotapes of recorded engineering lectures. Professional organizations governing educational assessment practices in the United States have supported the use of expert judges in assessing "the relative importance, criticality, and/or frequency of the various tasks" (AERA/APA/NCME, 2014, p. 14), and Brown's consultations with subject-matter experts provided a sound explanation for choices made about test content. Ultimately, Johns (2018) has asserted that it is the responsibility of English teachers in academic or university-based English language programs to collect authentic samples of university students' written work as part of the needs analysis process.

3.5.2 Evidence Based on Response Processes

Validity evidence based on test content is a necessary but not sufficient component of a comprehensive validation argument (Bachman, 1990; McNamara, 1996). As such, AERA/APA/NCME standards (2014) have suggested that stakeholders also gather validity

evidence surrounding individual responses to assessment tasks. A particularly useful approach to examining test-taker response processes involves adopting a Mixed Methods Research (MMR) design, which "simultaneously or sequentially integrates qualitative and quantitative points of view, data collection methods, forms of analysis, interpretation techniques, and modes of drawing conclusions" (Brown, 2014, p. 9). Qualitative methods such as interviewing test takers about their performance strategies and their approach to a particular writing task (Chapman, 2016) would provide information on examinee response processes. In addition, researchers could observe L2 writers while engaged in a writing assessment task to glean quantitative and qualitative data that could augment the formulation of a validity argument. The AERA/APA/NCME (2014) has also recommended analyzing "the development of a response to a writing task, through successive written drafts or electronically monitored revisions" (p. 15) to gather relevant validity evidence.

Evidence based on response processes also includes monitoring the ways in which raters apply scoring criteria. In a study investigating the behavior of raters entrusted with scoring 40 student essays, Winke and Lim (2015) found that of the five categories included in Jacob et al.'s (1981) ESL Composition Profile, the final category—mechanics—was the category to which raters attended least. While scoring, the researchers analyzed raters' eye movements over a computerized version of the rubric. Based on the eye tracking data, Winke and Lim determined that the raters referenced the mechanics category least, and while not necessarily linked causally, interrater reliability for the mechanics category was lower than the other four categories. Irrespective of the conclusions to be drawn from raters' inattention to mechanics, the study of Winke and Lim's study demonstrates the utility of exploring writer and rater response processes from multiple perspectives.

3.5.3 Evidence Based on Internal Structure

Evidence based on internal structure is the contemporary term for construct validity. Essentially, a construct is "the concept or characteristic that a test is designed to measure" (AERA/APA/NCME, 2014, p. 11). Defining the construct may be the most important consideration in designing a rating scale (Weigle, 2002), yet McNamara (1996) surmised that evidence based on internal structure has not been addressed sufficiently in the field of second language studies. Particularly in performance assessments, such as in the evaluation of student writing samples, several factors can contribute to construct-irrelevant variance, including factors within the writing task, factors related to the development of rating scale descriptors, and factors associated with the scoring procedures. As a result, the process of collecting evidence based on internal structure ideally will include reference to theoretical models of second language performance (e.g., Bachman's (1990) model of communicative language ability or McNamara's (1996) distinction between strong and weak language performance tests) and analyses of empirical data. Although "no test can ever completely capture the construct" (Messick, 1994, p. 20), at least two statistical methods exist for uncovering validity evidence based on an analytic rubric's internal structure: factor analysis and item response theory.

One method frequently used by test developers to evaluate the internal structure of a test is factor analysis (Furr & Bacharach, 2014). Factor analysis can illuminate the number of dimensions or *factors* present within a set of test items or tasks. In the case of multidimensional tests—tests designed to measure several components of a particular construct—factor analysis can also help to reveal connections among theorized factors. Thus, factor analysis is a useful statistical method for addressing fundamental questions regarding how accurately the criteria on an analytic rubric "capture the fully functioning complex skill" (Messick, 1994, p. 20). In

addition to factor analysis, item response theory (IRT) has also been employed to uncover validity evidence based on the internal structure of a writing assessment. IRT is a general term for a family of mathematical models based on the concept that the probability of an individual's expected performance on a particular item, task, or category is a mathematical function of the person's underlying latent ability and one or more item parameters (Bond & Fox, 2015). An important assumption of IRT models is that the data are unidimensional, meaning that only one latent trait is measured by a particular assessment (Hambleton et al., 1991). If a scree plot of the eigenvalues obtained from running a factor analysis shows that there is only one factor, then a one-parameter logistic model, such as the model developed by George Rasch, can be extremely useful for examining the internal structure of a scoring rubric. Within the Rasch model, the "odds" of an examinee receiving a particular score in a particular category from a particular rater are expressed as a logarithm or log, and the units on the resulting measurement scale are in turn called "log odd units" or *logits* (McNamara, 1996). An advantage of conducting an IRT analysis of data obtained through use of a rating scale is that estimates of test-taker ability, task difficulty, rater severity, and category difficulty can be placed on a common logit scale, thereby enabling comparisons among multiple facets within a specific research design. Furthermore, the output of such analyses typically includes fit statistics, which enable the identification of misfitting items and categories, and bias reports, which mark any biased interactions between raters and categories, raters and tasks, and raters and examinees.

In addition to providing data concerning rater variation and rater bias, an IRT analysis can be useful for revisiting the functioning of a rubric already in place. Janssen et al. (2015) adopted a mixed-methods approach in revising the Jacobs et al. (1981) analytic rubric, which had been adopted as part of the placement exam process in an English program for doctoral students.

Based on an examination of the category probability curves included in the output of a FACETS analysis, the researchers identified a number of redundant score levels within the rubric. Seeing as it was possible for raters to select from a range of scores within each proficiency band, the steps between score values were not as meaningful as they could have been. Ultimately, Janssen et al. (2015) determined that six points "maximized the number of meaningful levels within each category, while maintaining broad qualitative distinctions between different performance levels" (p. 58), and in turn the rubric was rescaled to include only six levels.

3.5.4 Evidence Based on Relations to Other Variables

In addition to the statistical methods available for gathering validity evidence based on the internal structure of a writing assessment, test developers and users should seek evidence concerning the extent to which rubric-based scores relate to other measures of writing ability. Essentially, the gathering of experimental and correlational evidence would determine whether or not the construct underlying the interpretation of rubric-based scores was associated in predictable ways with other variables (AERA/APA/NCME, 2014). Two terms used consistently in validation studies based on relations to other variables are *convergent* and *discriminant* evidence. Essentially, convergent evidence is found when there are positive correlations between the test and another test purported to measure the same construct. In contrast, correlations between two dissimilar measures of two different constructs serve as discriminant evidence.

3.5.4.1 Multitrait-Multimethod Matrices. One way to evaluate the quality of convergent and discriminant validity evidence is through construction of a multitrait-multimethod matrix (Furr & Bacharach, 2014). According to Bachman and Palmer (1983), an advantage of the multitrait-multimethod research design "is that it allows the researcher to distinguish the effect of measurement method from the effect of the trait being measured" (p.

156). In their 1983 study, Bachman and Palmer created a multitrait-multimethod matrix using three measures (oral interviews, translation tasks, and self-ratings) of oral proficiency and reading comprehension skills in English. The study involved 75 first language Mandarin Chinese speakers and two raters who scored the participants' performances on each of the six tasks (two traits times three methods). Based on the resulting multitrait-multimethod matrix, Bachman and Palmer concluded that the effect of method was much higher for the translation and self-rating measures of oral proficiency and reading comprehension. A subsequent confirmatory factor analysis revealed that the oral interview loaded highly on the speaking trait factor (0.819), whereas oral translation and oral self-ratings did not load highly, suggesting that the interview task was a more suitable means of gauging participants' oral communication skills.

3.5.4.2 Corpus Linguistics. Although multitrait-multimethod matrices necessitate complex research designs, relational evidence can also be found using correlations between test scores and measures of related constructs. For example, researchers can investigate the correlation between rubric-based scores of student writing samples and students' scores on standardized, even discrete-point assessments of writing ability. Recent research in corpus linguistics has investigated the predictive capabilities of measures of lexical and syntactic sophistication in students' written work (Kyle & Crossley, 2017; Kyle et al., 2018). In a validation study of a free text analysis tool, TAALES 2.0, Kyle et al. (2018) found that ten indices of lexical sophistication explained 58% of the variance in holistic scores assigned to 180 free writes written by academically-oriented English language learners. The same study also reported that 11 variables, including "indices related to phonological neighbors, lexical-decision times, word familiarity and frequency, and association strength" explained 32% of the variance in word-choice category scores on a sample of 716 narrative essays. In a separate study

investigating the predictive capacity of computational indices of syntactic sophistication in L2 writing, Kyle and Crossley (2017) determined that four indices related to the inclusion of frequent and less-frequent verb argument constructions predicted approximately 14% of the variation in the holistic scores assigned to 480 TOEFL independent essays. The results of these research programs signal the contributions such tools can make to an overall validity argument, including highlighting ways in which rating criteria might be enhanced, for example by specifying that essays reflective of higher score bands generally contain more strongly associated verb argument constructions.

3.5.5 Evidence Based on Consequences of Testing

A final piece of validity evidence can come from an investigation of the positive and negative (both intended and unintended) consequences of decisions made on rubric-based scores. Whereas some testing experts believe that the social repercussions of these decisions are impossible to anticipate in early discussions of validity (Reckase, 1998), Messick (1994) insisted that evidence based on the consequences of testing "should especially address the anticipated positive consequences of performance assessment for teaching and learning as well as potential adverse consequences bearing on issues of bias and fairness" (p. 22). Ideally, practitioners and researchers would engage in longitudinal research to investigate the impact of testing procedures. Among scholars in the language testing community, the consequences of testing are often discussed in terms of an instrument's washback. The origin of this designation along with scholarly recommendations for fostering positive responses to testing will be discussed in the remaining sections of this chapter.

3.5.5.1 Washback. In a case study presented as part of an edited volume on communicative language teaching, Swain (1984) outlined four principles of communicative

language *testing* that may still be useful to small- and large-scale test developers. The first principle, *start from somewhere*, encompasses the articulation of a theoretical framework, which in the case of an assessment of second language writing, might entail the creation of rating-scale criteria that reflect current theoretical understandings of the components of 'good' writing in a particular context. The second principle, *concentrate on content*, concerns the appropriateness of test stimulus materials and tasks for the target population, including learners' age, proficiency levels, interests, and goals (Bailey, 1998). The third principle, *bias for best*, implores the test developer to do "everything possible to elicit the learners' best performance" (Swain, 1984, p. 195), such as by allowing learners to use online dictionaries or thesauri during a writing assessment (Oh, 2020). Swain's fourth, and final, principle is to *promote positive washback*. Although the term is not utilized in general education or educational measurement literature (Hamp-Lyons, 1997), *washback* "refers to the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning" (Messick, 1996, p. 241). Explained differently, washback encompasses the positive—and negative—consequences of implementing a particular assessment tool.

Researchers and scholars have proffered several suggestions for promoting beneficial washback in language programs, the first of which is the use of more valid tests (Messick, 1996). In addition to exploring validity evidence that supports use of a particular scoring rubric, teachers and program administrators can encourage positive washback by incorporating language learning goals into rubric descriptors and by supplying detailed score reports (Bailey, 1996). When analytic rubrics are carefully designed, they should reference specific language learning objectives, which would serve, in turn, as the basis of detailed score reports. Messick (1996) and

Brown (1998) have also suggested that positive washback can be fostered by using criterion-referenced measures, a group of instruments that would include many classroom-based analytic rubrics. Brown (1998) recommended "testing those abilities you want to encourage" as well as "expanding the skills that are tested to include non-academic out-of-school tasks" (p. 17). Although three decades have passed since Swain (1984) coined the term, interest in washback on teaching and learning continues to grow (Tsagari & Cheng, 2017), with scholars investigating various types of washback, including washback on student learning, washback on learner autonomy, washback on instruction, and washback on curriculum design.

3.5.5.2 Washback on Student Learning. Weigle (2016) commented recently that a disproportionate amount of research in language testing has focused on scoring processes rather than on assessment for learning. Likewise, Tsagari and Cheng (2017) called for more longitudinal investigations of the ways in which assessments affect learners—the individuals most impacted by test scores. In addition, Lee and Coniam (2013) maintained that for assessment to impact student learning, students need to be aware of the criteria upon which their work will be evaluated and the steps available for improving their performance. Whereas a single holistic score may not provide learners with usable information for improving their written communication, a clear analytic rubric that accurately represents the construct under examination can increase learners' explicit genre awareness, ultimately facilitating longitudinal growth in this measurable skill. Analytic rubrics supply teachers and learners with information on the genre-specific attributes students have mastered and the features on which they should focus to further the caliber of future written production in that genre. Furthermore, when the same analytic rubric is used for diagnostic, progress, and achievement tests, students will receive a detailed record reflecting their progress toward individual learning objectives, which may inspire continued

effort on that task. Norris (2006) cautioned educators and program administrators involved in assessment not to overlook the value of information provided by sound assessment instruments for achieving specific course or program goals.

3.5.5.3 Washback on Learner Autonomy. In addition to aiding teachers in responding to student writing, analytic rubrics serve as a useful tool for learner self-assessment. Bailey (1998) affirmed that self-assessment "is consistent with a trend in language education in general—that of more emphasis on learner responsibility and learner-centered curricula" (p. 227). When introduced to the meta-language frequently included in criteria descriptors and trained in the application of those descriptors to the evaluation of texts, learners can assess their degree of mastery of an assignment's objectives. As Casanave (2017) explained, "assessment, importantly, eventually involves students' increasingly assessing their own work" (p. 258), and analytic rubrics provide learners with a means of doing so. A clear understanding of the expected rhetorical arrangement, communicative function, and register for a final product will enable leaners to monitor their performance during a particular writing task (Andrade & Evans, 2013). Andrade and Evans (2013) also surmised that these autonomous learners will establish practical goals, appraise their advancement toward skill acquisition, and modify their output to align with individual, task-specific, and course objectives. In sum, the systematic use of analytic rubrics can help students to take control of their learning, focusing attention on areas in which they want to see improvement.

3.5.5.4 Washback on Instruction. A number of outcome-driven pay initiatives in which a teacher's salary is linked to student performance on large-scale assessments assume a causal relationship between teaching and testing in that good teaching leads to higher student performance. However, Popham (2003) wrote that the other side of the equation is much less

understood, “that *how* a teacher tests—the way a teacher designs tests and applies test data—can profoundly affect *how well* that teacher teaches” (p. 1, emphasis in original). The practice of designing and implementing classroom-based assessments should enhance instructional activities (Turner, 1997), thereby benefitting all individuals involved in the educational system. The criteria and individual descriptors that constitute a sound analytic rubric can help learners to target areas of improvement and can aid teachers in tailoring their instruction based on student outcomes. Described by Popham (1997) as “instructional illuminators,” analytic rubrics also help teachers to address the issue of fair and accurate assessment of student writing. Specifically, “a scoring rubric provides the instructor with a standard by which to score papers consistently” (Weigle, 2002, p. 182), and according to Ferris (2018, personal communication), when writing assignments are assessed via an analytic rubric, students rarely, if ever, complain about their score. Finally, on a very practical level, rubrics can reduce the amount of time spent responding to student work (Davis & Kondo-Brown, 2012), possibly by focusing teachers’ efforts on supplying feedback relevant to the attributes of that particular skill or genre.

3.5.5.5 Washback on Curriculum Design. In addition to the aforementioned examples of beneficial washback, rubric-based assessment of productive language skills can assist teachers and program administrators in evaluating student needs and overall course effectiveness. Tests are a necessary element of Brown’s (1995) systematic approach to designing and maintaining language curricula because they can coalesce disparate curricular elements as well as imbue individual units with structure and meaning. Specifically, analytic rubrics can help L2 writing teachers to operationalize course objectives, ensuring that the desired learning outcomes are both observable and measurable. Crusan (2010) remarked that “rubrics can do for a curriculum what objectives do—they can help explain terms and clarify expectations” (p. 43). The development

of valid analytic rubrics involves several stages, which, similar to systematic curriculum development, inform one another in an iterative manner. As the Conference on College Composition and Communication (2009b) explained, "best assessment practice is informed by pedagogical and curricular goals, which are in turn formatively affected by the assessment" (Guiding Principles for Assessment, Section 1A).

In sum, teacher feedback and assessment play an important role in supporting second language writing development. Analytic rubrics in particular enable the provision of feedback on different aspects of the writing process or product. When utilized effectively, analytic rubrics make apparent certain constraints of a particular genre, thereby empowering second language writers to exercise their creativity within those boundaries or to challenge the constraints in order to achieve a specific outcome. Certainly, analytic rubrics are not the only component of a sound philosophical approach to the teaching of second language writing. Carefully crafted writing prompts (Kroll & Reid, 1994), tasks (Weigle, 2002), corrective feedback (Ferris, 2011), and grammar instruction (Larsen-Freeman, 2003) also form part of a teacher's repertoire of tools for promoting students' writing development. Nevertheless, the use of data-based, analytic scoring rubrics in local and global assessments of second language writing is likely to promote positive washback in the form of increased student learning, promotion of learner autonomy, and adjustments to instruction and curriculum design.

CHAPTER 4: METHOD

4.1 Methodological Framework

4.1.1 Mixed Methods Research

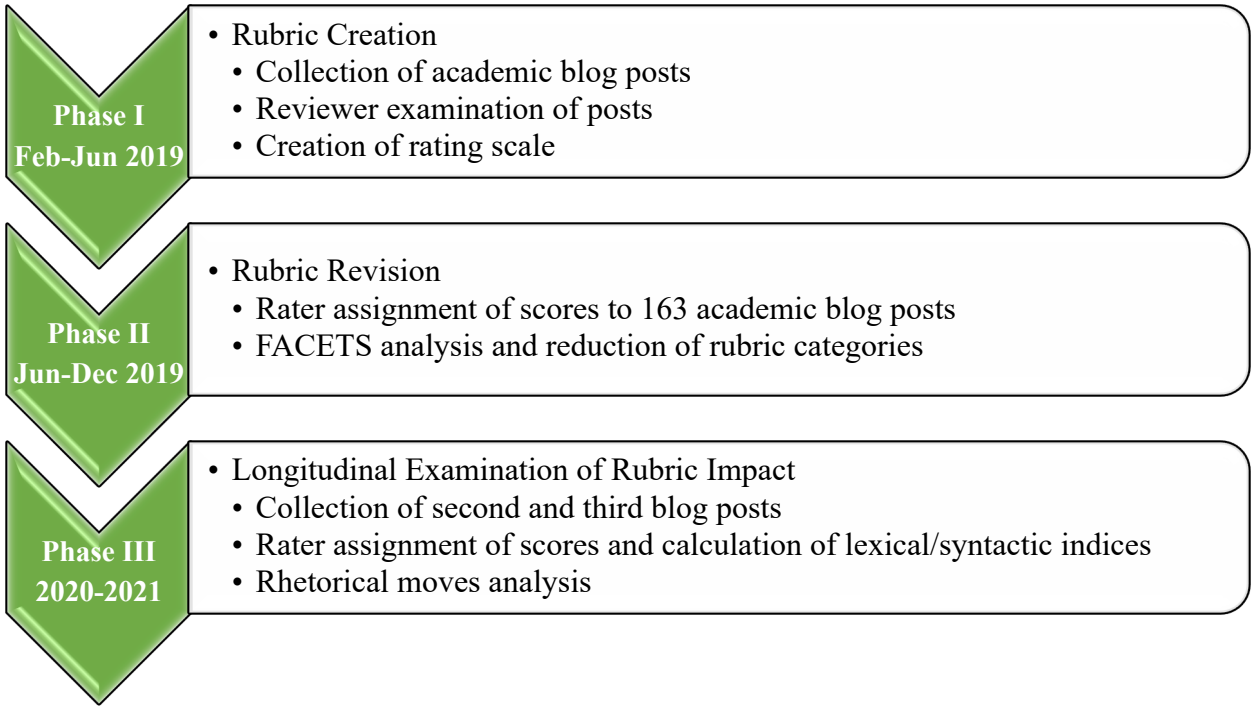
Seeing as the scope of this research necessitated multiple phases of data collection and analysis, a Mixed Methods Research (MMR) approach was adopted as the overarching methodological framework. Brown (2014) explained that Mixed Methods Research involves the simultaneous or sequential integration of "qualitative and quantitative points of view, data collection methods, forms of analysis, interpretation techniques, and modes of drawing conclusions" (p. 9). Beyond a simple mixing of qualitative and quantitative data, MMR encompasses a purposeful, systematic selection of research methods to capitalize on the strengths of multiple modes of inquiry. MMR is particularly suited for research designs that involve the use of multiple data points, including various data sources (e.g., participant writing samples and rater quality judgments) and distinct assessment instruments (Long, 2015). As a result, MMR has become an increasingly important design framework for doctoral dissertations (Patharakorn, 2018; Yasuda, 2012; Youn 2013), even generating discussions around different types of Mixed Methods designs (Creswell & Creswell, 2018). Despite the challenges of Mixed Methods Research, including extended periods of time for data collection, researcher familiarity with qualitative and quantitative forms of data analysis, and complex iterations of research procedures, MMR supports the gathering of the most relevant, useful information possible in order to answer questions of pragmatic import.

This research followed an exploratory, sequential mixed-methods design, in which qualitative and quantitative data contributed to the creation of a new assessment instrument. This instrument was tested in the second phase of the project and implemented in the study's third

phase to evaluate the success of using said assessment instrument as an intervention. Specifically, the first phase of this project involved developing a new rating scale that accurately described various levels of written performance on the emerging genre referred to here as the academic blog post. Within this first phase, sample texts were collected, and quantitative and qualitative data provided by expert reviewers of those texts were used to build an analytic rubric. In the second phase of the research, several raters piloted the rating scale by scoring 163 academic blog posts. This quantitative data enabled an examination of various psychometric properties of the assessment instrument, which in turn, led to the revision of the initial rating scale. In the final phase of the project, both qualitative and quantitative longitudinal data were collected, so as to evaluate the impact on second language writing development of the application of a genre-specific rubric. Figure 1 presents a visual model illustrating this exploratory, sequential mixed-methods design.

Figure 4.1

Diagram of Sequential Mixed-Methods Design



Not only does a mixed methods orientation enable researchers to address the needs of various stakeholders, but also the data obtained from such an approach may minimize the limitations of working with solely quantitative or qualitative data. Long (2015) maintained that "use of *multiple* sources typically provides more detailed information and has the additional advantage of allowing cross-checking for information validation, and ideally, triangulation of findings" (p. 118). In order to address the unique goals of the project as well as to strengthen the validity of key findings, three types of triangulation were selected: triangulation of method; triangulation of data sources; and triangulation of analysis. Employing various methods of data collection, including a review of relevant literature, questionnaires, sample texts, and ratings, helped to develop a more robust description of the genre in question and to document systematically changes in participants' written products over time. In addition, since students,

instructors, and test developers have different, yet complementary rolls in the assessment process, gathering information from various sources (e.g., second language writers, experienced reviewers, raters, and assessment-design experts) helped to ensure that the perspectives of various stakeholders were taken into account in the development process as well as to cross-verify findings. Finally, triangulation of data analysis in the longitudinal phase of the project, including the use of rater scores, the calculation of various lexical and syntactic sophistication indices, and the qualitative examination of three participants' blogs posts over time, strengthened the validity of the research endeavor. Table 4.1 summarizes the type of data collected in each phase of the project as well as the contribution made by each data source to the resolution of the research questions presented at the outset of the study. The remaining sections of this chapter detail the sequence in which qualitative and quantitative data were collected and analyzed, and the ways in which both types of data informed subsequent steps of the information gathering process.

Table 4.1

Data Types, Spacing, and Contribution to Research

Data	Type (QUANT or QUAL)	Timing (Phase)	Purpose
Needs Analysis	Qualitative	Phase I	To design an authentic, online writing task
Questionnaires	Qualitative and Quantitative	Phase I	To gather participant- and genre-relevant information
OOPT	Quantitative	Phase I	To assess participants' English level objectively
Post 1 (*N*=163)	Qualitative and Quantitative	Phase I	To create a data-based analytic rubric
Reviewer Analysis	Qualitative and Quantitative	Phase I	To identify rubric categories and descriptors
Rater Analysis	Quantitative	Phase II	To revise initial six-level rating scale
Post 2 (*N*=31)	Qualitative and Quantitative	Phase III	To document changes in written performance
Post 3 (*N*=31)	Qualitative and Quantitative	Phase III	To document changes in written performance
Rater Scores	Quantitative	Phase III	To analyze impact of rubric on written performance

4.1.2 Design-Based Research

A second methodological lens through which to view the present research can be found under the umbrella term design-based research (DBR). In spite of being a relatively recent addition to the methodological toolkit of Applied Linguistics researchers, DBR is a research methodology that utilizes iterative cycles of development, refinement, and implementation in real-world contexts (Wang & Hannafin, 2005). As a result, design-based research is often longitudinal in nature (Larsen-Freeman, 2013), and it frequently employs multiple or mixed methods to study the effect of a particular intervention (The Design-Based Research Collective, 2003). However, beyond merely perfecting an instructional tool or assessment instrument, DBR aims to contribute to current theories of language learning. As Reeves and McKenney (2013) explained, the two primary goals of DBR are to develop solutions to actual educational problems

and to refine theoretical principles "that can guide other researchers and practitioners focused on these same or closely related educational problems" (p. 10). Therefore, in addition to following an exploratory, sequential mixed-methods design, this project can be classified as design-based research in that it seeks to clarify the nature of an emerging genre in online and face-to-face tertiary education and to advance the use of genre- and context-specific rating scales in second language writing pedagogy.

4.2 Phase I: Rubric Creation

4.2.1 Participants

4.2.1.1 Second Language Writers. 148 English as a Foreign Language (EFL) university students volunteered to participate in the first phase of the project. The learners were undergraduate (*N*=121) and graduate (*N*=27) students enrolled in various degree programs, including English Philology, Bilingual Education, and International Relations, at the University of Murcia in Spain. The average age of the examinees was 21 years and 6 months, and the participant sample was predominately female (83%). All of the examinees reported Spanish as their first language, though one examinee listed both Spanish and Arabic as first languages. The English language level of the participants was assessed by means of the Oxford Online Placement Test (OOPT). The exam provided numerical scores out of a total possible score of 120, as well as the corresponding proficiency level on the Common European Framework of Reference for Languages (CEFR). Table 4.2 summarizes the demographic information collected on the examinees, and Table 4.3 includes descriptive statistics for the OOPT and the participants' age.

Table 4.2

Participant Demographic Information

Variable	*N*	Percentage of Total
Academic Status		
Undergraduate	121	81.7
Graduate	27	22.3
Gender		
Female	123	83.1
Male	25	16.9

Table 4.3

Descriptive Statistics for Age and Participant Performance on English Placement Test

Variable	*N*	Mean	*SD*	Min	Max
Age	147	21.52	5.33	18	53
OOPT	148	72.47	18.50	15	119

4.2.1.2 Reviewers. Six Individuals (four females and two males) with expertise in second language writing assessment were recruited to review the academic blog posts written by the 148 examinees. The reviewers held at least a master's degree in Second Language Studies, Teaching English to Speakers of Other Languages (TESOL), English Education, or Applied Linguistics, and they had all been exposed to the emerging genre of the academic blog post as a student or as an instructor of English. Both native and non-native speakers of English were included as reviewers in an effort to avoid what Fulcher (1988) observed as a tendency to use native-speaker competence as a "yardstick upon which non-native performance is to be measured" (p. 5). All of the reviewers were expert users of English, and each reviewer was proficient in at least one additional language, of which Hindi, Japanese, Russian, Spanish, and Vietnamese were represented.

4.2.2 Materials

4.2.2.1 Consent Form. As required by the Institutional Review Board at the University of Hawaii, all participants acknowledged their willingness to participate by signing an informed consent form. In addition to explaining the purpose of the research and the participants' contributions, the consent form outlined the measures in place to keep personally-identifiable information confidential. For example, after individual consent forms had been scanned and saved on a password protected computer, all physical copies were destroyed. In addition, data obtained during the study was linked only to randomly assigned ID numbers. Consent forms were available in both English (Appendix A) and Spanish (Appendix B).

4.2.2.2 Background Information Form. In addition to the consent form, examinees completed a background information form (Appendix C). The questionnaire asked participants about their educational background, their current field of study, and their training in linguistics or language analysis. In line with the questionnaire used in Rebuschat et al. (2013), the form also asked the EFL students to describe previous language learning experiences, including native language(s), and second or third languages. In the case of second or subsequent languages, participants were asked to provide additional information related to the context of instruction or use, the total hours of study, the length of any stays abroad, and their self-reported proficiency in the language. Particularly relevant to this research, the questionnaire also sought to gauge students' familiarity with blogs by asking them to list any bloggers they were following and to describe the content of any self-maintained blogs. All items on the questionnaire appeared in English and Spanish, and participants had the option of completing the form in either language.

4.2.2.3 Oxford Online Placement Test. To obtain a measure of examinees' English language ability, students completed a computer-adaptive, online version of the Oxford

Placement Test. The Oxford Online Placement Test (OOPT) used the Common European Framework of Reference for languages (CEFR) as a guide for developing test items that could discriminate among CEFR's six levels of proficiency. The proficiency levels range from A1 (basic user) to C2 (proficient user), and the OOPT provides test takers with a score from 0 to 120, which would align with one of the six proficiency levels (see Table 4.4 for OOPT-CEFR alignment data). The OOPT contains two sections, the first of which measures test takers' grammatical and pragmatic knowledge (approximately 30 questions). The second section assesses test takers' English listening ability (approximately 15 questions). Although the OOPT does not engage test takers' productive skills beyond the occasional one-word fill-in-the-blank item, the online version offers a number of advantages over the pen-and-paper placement exam. First, the computer-adaptive nature of the OOPT reduces the amount of time students need to be engaged with the test. Since test takers' answers to early items enable the program to select more ability-appropriate items for ensuing sections, students spend only 30 to 60 minutes—as opposed to multiple hours—taking the test. Second, the database available through Oxford English Testing (www.oxfordenglishtesting.com) allows administrators to keep test data organized and secure in a password-protected platform. Third, test administrators have the ability to control the accent(s) used in the listening section and the release of individual scores upon completion of the placement test. For this study, a mix of British and American accents was selected due to the Spanish education system's preference for British English and to the students' familiarity with American pop culture. In addition, scores were released to participants at the end of the test, as an incentive for taking it. Since many Spanish university students are required to obtain certification of a B1 or B2 level of English, releasing the scores could give participants a sense of their progress toward that goal.

Table 4.4

OOPT Score Alignment with CEFR Level

Level	Score Range
A1	1-20
A2	21-40
B1	41-60
B2	61-80
C1	81-100
C2	>100

4.2.2.4 Writing prompt. Several researchers have suggested ways in which writing prompts affect learners' written products (Chapman, 2016; Macaro, 2014; Melzer, 2014; Schoonen et al., 2009). In an analysis of writing prompts assigned in writing across the curriculum courses, Melzer (2014) concluded that writing tasks should specify a real-world rhetorical situation and genre, aiming ideally for an audience beyond the classroom. Thus, Melzer's recommendation—supported by research on genre-based approaches to teaching writing—led to the inclusion of a context-setting paragraph prior to the prompt's rhetorical cues (Appendix D). This paragraph placed writers in an English language course at a U.S. university, and it identified the genre of the expected response: an academic blog post. Acknowledging the possibility that students were unfamiliar with the genre, clues concerning the probable audience of a blog post (e.g., the course instructor, classmates, and other Internet users) were specified. Furthermore, to simulate as closely as possible the environment in which a university student would realize this particular task, the context-setting paragraph included information concerning the appropriate use of online linguistic tools. By not restricting participants' access to online dictionaries or thesauri and by providing extended time for the composition of the blog post (i.e., an hour and a half versus the frequently encountered thirty minutes), the task acquired additional ecological validity.

After situating the writing task within a plausible rhetorical situation, Chapman's (2016) investigation of the ways in which writing prompts influence test-takers' written products was considered. Ultimately, Chapman identified four distinguishing characteristics of independent writing prompts: prompt *domain* or topic area (e.g., educational, occupational, public, or personal); *response mode* (narrative or argumentative); the number of *rhetorical cues* (i.e., instructions or questions to which the writer must respond); and the *focus* (open or focused). Chapman also discovered that prompts addressing an academic subject matter led to higher use of academic vocabulary on the part of the test taker, and that the response mode implicated in the prompt (either argumentative or narrative) affected the textual features of the participants' essays. Specifically, "prompts that elicited argumentative responses were characterized by syntactically complex writing" (p. 141). Information gathered during post-study participant interviews indicated that test takers favored prompts with more rhetorical cues and prompts that allowed various points of entry (i.e., open prompts), enabling them to write about topics with which they were familiar. In light of Chapman's findings and in order to establish prompt equivalence during the longitudinal phase of the study, the first writing prompt (Appendix D) targeted the educational domain (language learning through technology) and encouraged an open, argumentative response. The prompt also included five rhetorical cues as well as visuals to give participants something to write about and to encourage consideration of multiple points of view.

4.2.3 Procedure

4.2.3.1 Needs Analysis. Since a needs analysis is often the first stage of any task-based language teaching project (González-Lloret, 2014), previous research on university level writing tasks was consulted (Burstein et al., 2016; Hale et al., 1996; Melzer, 2014). The two most recent

reviews (Burstein et al., 2016; Melzer, 2014) identified the inclusion of several online writing tasks. Melzer (2014), for example, discovered that a number of university instructors incorporated blogs as a type of exploratory writing, and Burstein et al.'s (2016) survey results pointed similarly to the presence of electronic modes of written communication (e.g., email and blogs). The emergence of blogs in both studies inspired a subsequent search for empirical information concerning the nature of this new academic genre. One of the studies accessed in this search (Elgort, 2017) suggested that second language learners may be less familiar with the discourse features of educational blogs, which supported the selection of the academic blog post for this research. This background combined with Chapman's extensive consideration of writing prompt characteristics guided the design of the first writing prompt (Appendix D).

4.2.3.2 Data Collection. After an appropriate genre and writing prompt had been designed, the next phase of the research involved gathering relevant writing samples. To recruit participants for the study, I visited several classrooms at the University of Murcia to explain the purpose of the research and the benefits of participating (i.e., practice writing in English and receipt of a gift card to a local place of business). The EFL university students who chose to participate in the project signed a consent form, completed a background questionnaire, took the Oxford Online Placement Test, and posted their response to the writing prompt in a secure learning management system (www.kidblog.org). Kidblog was a unique platform in that it allowed students to select a colorful header for their blog entry as well as to decide whether or not they wanted other members of the study to be able to read their post. Participants had up to 90 minutes to write their post, and they were permitted access to online linguistic tools such as dictionaries and spell check, features generally available to university students when preparing an out-of-class written assignment (Oh, 2020). Although a handful of students needed assistance

in publishing their post, as the "Publish" button was a light-grey color and therefore not always perceptible, all participants managed to navigate the platform successfully. Data collection took place between February and June of 2019.

4.2.3.3 Review of Participant Blog Posts. At the end of the data collection period, participant blog posts were transferred to PDF files to ease distribution of the texts to six reviewers. Blog posts were labeled by participant number, and any titles assigned to the blogs remained in the documents. In total, 148 texts stripped of any personally-identifying information were uploaded to a Google Drive Folder to which an individual reviewer had access. A copy of the writing prompt and an empty spreadsheet, titled "Reviewer Marking Sheet," were also uploaded to the shared folder. Next, an email was sent to each reviewer with instructions on how to proceed with the review (Appendix E). The instructions asked the reviewers to familiarize themselves with the writing prompt and the reviewer marking sheet before beginning their evaluation of the blog posts. Since there was no way to control the order in which the reviewers would read the texts, the instructions stated explicitly that they were free to proceed with the task in a way that made sense to them. Ultimately, the reviewers were tasked with examining the participants' blog posts and separating them into six electronic folders in order of merit [rather than nine piles as in Diederich et al., 1961]. The least successful blog posts were to be placed in a folder labeled with the number one, and the most successful entries were to be placed in a folder identified by the number six. Six levels were deemed appropriate seeing as Janssen et al. (2015) had determined that human raters were generally able to distinguish between five and seven category levels and that six points "maximized the number of meaningful levels within each category" (p. 58) for their English as a second language (ESL) composition rubric. Reviewers were free to move posts from one folder to another if, after reading additional blog posts, they

wanted to reassess their initial classification. In addition, the reviewers recorded the precise folder number to which they had relegated each post and the salient characteristics of each text on a Google spreadsheet. Reviewers were encouraged to use short phrases and to avoid vague descriptors such as “adequate” and “acceptable” in their comments. The only strict guidelines provided were to use every folder and to ensure that no folder contained less than four entries. Although each reviewer completed the task within the allotted five hours, the turn-around time for the reviews varied from ten days to five weeks.

4.2.4 Data Analysis

4.2.4.1 Quantitative Analysis. The first analysis of participant blog posts and reviewer data was quantitative in nature. Specifically, descriptive statistics were used to examine variation in post length as well as to calculate measures of central tendency for the folder numbers to which reviewers had assigned individual blog posts. Next, the software program SPSS was used to calculate a set of descriptive statistics specific to factor analysis: the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett’s test of sphericity. These measures indicated that the quantitative data provided by the reviewers was suitable for factor analysis. As a result, an exploratory factor analysis was run to identify the number of factors, or constructs, that could have accounted for variation in reviewers’ decisions to assign a particular post to one folder over another.

4.2.4.2 Qualitative Analysis. After a factor analysis, a qualitative analysis of the short phrases the reviewers had attributed to each post was conducted. First, all 148 texts were ranked according to average folder score. Second, seeing as measures of central tendency indicated that the distribution of folder scores was relatively normal, an intent was made to split the posts into six levels of performance, with a desired allocation of 4, 20, 50, 50, 20, and 4 posts to levels one

through six, respectively. The reviewer comments associated with each post were compiled by participant and by level in a Microsoft Word document. Next, I considered carefully each set of comments, and as recurrent themes emerged from the data, I retraced my examination, color-coding descriptors that referred to the same textual feature. Ultimately, the titles for each of the rubric categories as well as the respective level descriptors were pulled directly from reviewer comments. In this way, the resulting rating scale, including the criteria that corresponded to different scoring bands, had an empirical basis in performance data.

4.3 Phase II: Rubric Revision

4.3.1 Participants

4.3.1.1 Examinees. A total of 163 individual, academic blog posts formed the data base for the second phase of the project. In addition to the 148 texts collected for Phase I, another 15 individuals responded to the prompt on language learning in technology. The additional 15 participants followed the procedure outlined in Phase I, adjusting the makeup of the overall group only slightly. Since the extra 15 individuals were majoring in English Philology at the University of Murcia, the total number of undergraduate students increased from *N*=121 to *N*=136, with the number of graduate students remaining the same, *N*=27. The average age of the 163 examinees dropped from 21.52 to 21.47 years, and the sample remained predominately female. The first language of the examinees was Spanish, with one participant citing Spanish and Arabic as first languages and another participant listing Spanish and Catalan as first languages. Table 4.5 summarizes the demographic information collected on the examinees for Phase II of the research, and Table 4.6 includes descriptive statistics for the participants' age and their scores on the Oxford Online Placement Test (OOPT). Of note is that with the additional participants, the mean score on the OOPT rose only 0.05 points from 72.47 to 72.52.

Table 4.5

Participant Demographic Information Phase II

Variable	*N*	Percentage of Total
Academic Status		
Undergraduate	136	83.4
Graduate	27	16.6
Gender		
Female	136	83.4
Male	27	16.6

Table 4.6

Descriptive Statistics for Age and Participant Performance on OOPT Phase II

Variable	*N*	Mean	*SD*	Min	Max
Age	147	21.47	5.11	18	53
OOPT	148	72.47	18.33	15	119

4.3.1.2 Raters. Six individuals new to the study (four females and two males) were tasked with assessing the 163 examinee posts. Similar to the background of the six reviewers, each rater had extensive teaching and research experience in the field of Applied Linguistics, with one rater holding a doctoral degree in Second Language Studies, one rater holding a doctoral degree in English Philology, and the other four individuals holding master's degrees in English Philology, TESOL, or Second Language Studies. Half of the raters had received their higher education in Spain, with the other half completing graduate school in the United States, and all of the raters were engaged in pedagogical and/or curricular projects involving English language learners at the time of the study.

4.3.2 Materials

4.3.2.1 Academic Blog Posts. The academic blog posts written by the 163 participants served as the primary data source for this phase of the study. While composing their responses to the rhetorical cues surrounding the role of technology and language learning, several participants

sought information regarding the expected number of words for the post. This information was purposely kept out of the instructions leading up to the prompt, since the ideal length for an academic blog post had not been suggested in previous literature. Whereas some prior researchers indicated that they had provided participants with a suggested word count for course discussion board or blog posts, the construction of this writing task left the question of length open to participants' interpretation of the audience and context. Potentially as a result of this decision, the descriptive statistics related to word count indicated an extremely large range between the word count of the shortest (108) to the longest text (2504). Closer inspection of the frequency data, however, reveal that only one post was over 1,000 words: a text with 2,504 words. Although this outlier contributed to high values for kurtosis and skewness, the post remained part of the data set. The average number of words per post was 350 with a Standard Deviation of 222.

4.3.2.2 Initial Rubric. The comprehensive work of the reviewers in the first phase of the study contributed to the formation of a rating scale that consisted of five categories and six levels (Appendix F). The categories were labeled, "Task Fulfillment & Relevancy," "Content," "Organization & Balance," "Genre Specific Features," and "Language Use." Along with these five categories, unique data-derived descriptors were selected to coincide with a numerical scale from 1 to 6, with the higher values of 5 and 6 indicating written work of greater quality. With a minimum score of 1 and a maximum score of 6 for each category, total scores would range from 5 to 30 points. Finally, instructions included at the top of the rubric specified that the scale was meant to be used as an instructional tool as well as for self- and/or instructor-assessment of an academic blog or discussion board post written in response to a particular set of rhetorical cues.

4.3.3 Procedure

All 163 posts were uploaded as PDF files to rater-specific folders on Google Drive. Raters were also provided access to the prompt, the six-level analytic rubric, and a spreadsheet on which they were asked to record participant scores. Via written instructions and oral communication, raters were instructed to score the posts in accordance with the descriptors provided on the analytic rubric. Specifically, raters were requested to select scores from 1 to 6 for each of the five categories, keeping in mind that they did not need to assign the same score to every category for a particular post. No additional training was provided for the raters in an effort to mimic the reality of many actual rating contexts, with the added benefit of eliminating consideration of any rater training effects in the interpretation of results. Raters took from two to four weeks to complete the task, and although they had the opportunity to record written comments on their scoring sheets, none of the raters chose to do so. There were no texts that the raters were unable to score and no missing data points.

4.3.4 Data Analysis

The goal of the analysis in the second phase of the study was to gather validity evidence based on the internal structure of the rubric. In other words, it was important to examine whether or not the rating scale could measure successfully the construct it had been designed to measure, namely, the ability to compose an effective academic blog post in line with five data- and expert-determined categories. Although Messick (1994) pointed to the inability of tests to encapsulate a construct perfectly, an effective method for uncovering validity evidence based on a rating scale's internal structure is Item Response Theory (IRT). An extension of Classical Test Theory, the foundation of Item Response Theory is a mathematical function that relates examinees' probability of answering a particular test item in a particular way according to their competence

(on a continuum from low to high) on a particular trait (Ostini & Nering, 2006). In its most basic sense, IRT models describe the likelihood that an examinee of ability level, θ, will answer a particular item of difficulty, b, correctly.

There are several assumptions of IRT models, including monotonicity, unidimensionality, local independence, and invariance. For a monotonic function, the slope is always increasing or always decreasing, and with respect to IRT models, the probability of answering an item correctly should increase as a test taker's ability increases. Another key assumption of IRT models is that only a single attribute, or one construct be measured at a time (Bond & Fox, 2015). This concept, known as unidimensionality, is essential if we want estimates of examinee ability and item difficulty to be meaningful, and in turn, to produce an assessment instrument strong enough to provide useful evaluations of examinee performance. A common method used to show the unidimensionality of a rating scale is a Principle Component Analysis (PCA). If a scree plot of the eigenvalues obtained from running this analysis demonstrates that there is only one factor, then a one-parameter logistic model could be a logical next-step in evaluating the internal structure of the rating scale. Generally, if the percent of variance explained by the first eigenvalue exceeds 50%, we can consider the data unidimensional; however, it is also suggested that the ratio of the first and second eigenvalues exceed a value of five. In addition to monotonicity and unidimensionality, IRT models assume local independence of items (categories in this study). In other words, examinee performance within one category of the rubric should be relatively independent of that individual's performance in another category.

An IRT model used frequently in the field of Applied Linguistics is the one-parameter logistic model developed by George Rasch. A theoretical basis for using the Rasch model in this research is that the research questions are concerned primarily with examining whether the

categories on a data-based analytic rubric are capable of estimating a single latent trait: English writing ability on a genre-specific online task. Two parameter models take into account item difficulty as well as item discrimination, which is not of interest in this study. In other words, I did not set out to identify whether or not certain rubric categories were better at discriminating among participants with a given ability level. The three-parameter model adds "guessing" to the equation, which is also not relevant to a performance task. As a result, as long as the model fit statistics support the above theoretical considerations, the mathematical model to be used in the analysis is shown below in Figure 4.2

Figure 4.2

Rasch Model

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Note. θ = examinee ability; b_i= item difficulty

Since the basis for rater scores in this phase of the project was a rubric containing six levels, the data were polytomous in nature, meaning that an examinee would not just fail or succeed on a given category, but rather the examinee could be given credit for progress toward mastery of a particular attribute of academic blog-writing ability. Thus, the Partial Credit Rasch Model (Masters, 1982) was employed for a quantitative analysis of rater scores. According to Bond and Fox (2015), the Partial Credit Model (PCM) "exemplifies the nature of the bridge between purely qualitative theory, methods, and ordinal data on the one hand and the construction of interval-level Rasch measures on the other" (p. 142). Furthermore, since all six raters provided scores on every rubric category for all 163 academic blog posts, the design was considered fully crossed. In this study, the many-faceted Rasch analysis was conducted using the software program FACETS, version 3.8.2 (Linacre, 2019). The output from FACETS placed

estimates of participants' task-specific writing ability, rater severity, and category difficulty on a common logit scale, enabling comparisons among the three facets. The output also included bias reports signaling potentially biased interactions between raters and categories and raters and individual participants. Finally, FACETS provided category-specific information concerning whether or not the raters made use of all six levels on the rating scale. These category probability curves led to the identification of redundant scoring levels within the rubric, and seeing as the steps between score values were not as meaningful as had been predicted, the rubric was rescaled before utilization in Phase III of the research.

4.4 Phase III: Longitudinal Examination of Rubric Impact

4.4.1 Participants

4.4.1.1 Authors. Of the 163 participants who composed an academic blog post in response to the first prompt, 31 individuals returned to write a second and then a third post in response to two different prompts. For this phase of the research, the participants were divided randomly into two groups: one group had access to the revised rubric before and during the composition of their second and third posts, whereas the other group wrote their second and third texts without reference to the rubric. The rubric group consisted of 15 individuals (4 males and 11 females), and there were 16 participants (3 males and 13 females) in the non-rubric group. To confirm that there was no statistically significant difference in learners' English ability level between the rubric group and the non-rubric group, an independent samples *t*-test, with alpha set at 0.05, was run on the participants' OOPT scores. All assumptions for using the parametric *t*-test, including a nominal independent variable with two levels and an interval-like dependent variable approximating a normal distribution (Turner, 2014), were met. The result of the *t*-test indicated that there was no statistically significant difference in the mean scores of the two

groups (see Table 4.7 for descriptive statistics describing the behavior of the OOPT scores of the rubric and non-rubric group). At the same time, the lack of a statistically significant difference between the groups did not mean that the two groups were identical. In fact, the results will show that on the first (pre-treatment) post, the rubric group received an average score that was less than the average score earned by the non-rubric group. Finally, all participants listed Spanish as their first language and English as a second language, and all but four participants claimed working knowledge of at least one additional language, of which French, German, and Italian were the most common languages in descending order.

Table 4.7

Descriptive Statistics Concerning Participant Performance on OOPT

Group	*N*	Mean	*SD*
Rubric	15	79.13	17.43
Non-rubric	16	78.25	20.69

4.4.1.2 Raters. Given the efforts taken in Phase II to collect validity evidence in support of a revised rubric, a group of three raters was deemed sufficient for Phase III of the research project. The three raters self-identified as female, and they held advanced degrees in fields closely associated with Applied Linguistics. With recent experience in higher education, they were familiar with the genre of focus in this research. Furthermore, as young professionals with careers centered on English language education, these raters were well-versed in the assessment of second language writing. Each of the raters signed a consent form that outlined the measures taken to keep their information as well as participant information confidential.

4.4.2 Materials

4.4.2.1 Prompts. The second and third prompts were designed according to the criteria around which the original prompt on language learning and technology was constructed.

Specifically, each prompt addressed a topic in the realm of education. Prompt 2 (Appendix G) asked participants to discuss the use of homework in K-12 education and Prompt 3 (Appendix H) engaged participants in writing about the place of religion in public and/or private education in Spain. Since the participants came from academic fields associated with the learning and teaching of English as a Foreign Language, selection of topics from within the educational domain sought to ensure that participants would enter the task with a similar level of background knowledge. Both prompts included four rhetorical cues, and they encouraged open, argumentative responses. The context-setting paragraphs for each prompt were identical to the context described in the initial task, and visual cues that corresponded to the topics of the respective prompts rounded out the documents.

4.4.2.2 Revised Rubric. For the longitudinal portion of the study, the raters used the revised, four-level rubric completed in Phase II. The revised rubric contained the same categories as the first rating scale, though the number of levels had been reduced from six to four, and in turn, descriptors corresponding to each of the rubric cells had been adjusted. Details outlining the steps involved in revising the analytic rubric are discussed in Chapter 5.

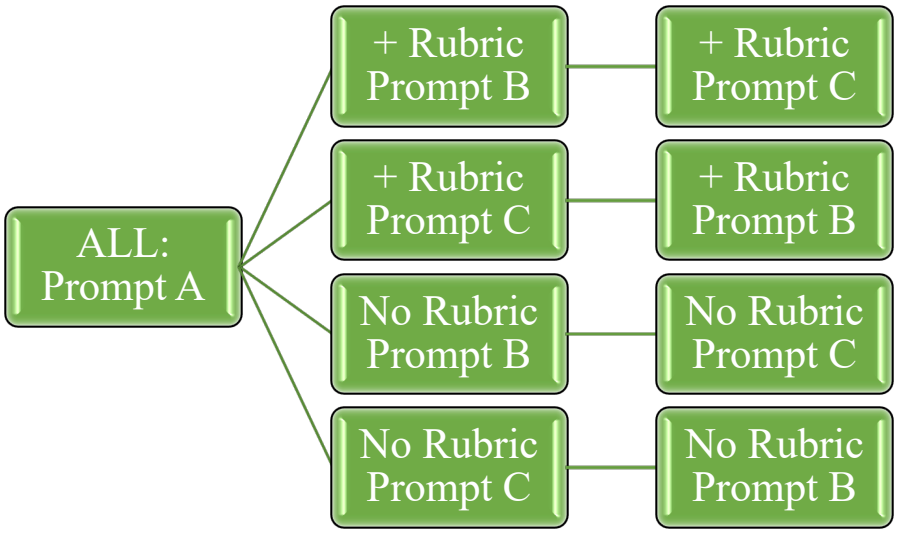
4.4.3 Procedure

4.4.3.1 Collection of Writing Samples. To ensure a counter-balanced design, half of the participants in the rubric group and half of the participants in the non-rubric group received Prompt 2 for their second post, and the other half responded to Prompt 3. For the third post, those participants who had responded to Prompt 2 received Prompt 3, and vice versa. Figure 2 demonstrates graphically the crossed design. I met one-on-one with participants in each group for the composition of their second and third posts. When meeting with individuals in the non-rubric group, I asked them to read their initial post and then to identify verbally any aspects of

their previous written work that they might like to improve for their second post. After this step, I read aloud the rhetorical cues for the new prompt, and the participants proceeded to compose their posts online within the allotted time frame (90 minutes). During composition of the post, participants retained access to the prompt (in English and in Spanish) and to various online linguistic tools (e.g., bilingual and monolingual dictionaries, thesauri, etc.).

Figure 4.3

Visual Depiction of Crossed Design



When meeting with participants assigned to the rubric group, they were asked first to examine the rubric that had been developed in Phase II. I pointed out the five categories included on the rubric, and together, we reviewed the descriptors that corresponded to the four levels of the rating scale. Participants also had the opportunity to self-assess their performance on the initial post, and based on their assessment, they were asked to identify verbally particular attributes of their previous written work that they might like to adjust in subsequent posts. After this step, I read aloud the rhetorical cues for the new prompt, and the participants proceeded to compose their posts online within the allotted time frame (90 minutes). During composition of

the post, participants retained access to the prompt (in English and in Spanish), to the rubric, and to various online linguistic tools (e.g., bilingual and monolingual dictionaries, thesauri, etc.).

Although many of the meetings described above took place in-person, mid-way through collection of participants' second academic blog post, in-person meetings proved impossible due to the emergence of COVID-19 in Europe. As a result, the data collection procedures were migrated to an online environment powered by Zoom, a well-known video conferencing software. Thankfully, all of the individuals who participated in the longitudinal phase of the study had had prior experience with Zoom or had learned to adapt quickly to the virtual environment given the necessities of the time. Through Zoom, the participants and I were able to share screens to review the prompt's rhetorical cues or to ensure the successful use of the blogging platform (kidblog.org). The enduring presence of COVID-19 throughout 2020 also necessitated virtual collection of the third writing sample, and procedures for data collection mimicked the steps followed for collecting participants' second posts. The time elapsed between consecutive posts ranged from six months to one year, and at the conclusion of data collection, participants in the non-rubric group were provided with the revised rating scale.

4.4.3.2 Ratings. With the culmination of data collection, the three blog posts written by each of the 31 longitudinal participants were transferred to separate Microsoft Word documents. These 93 texts were then stripped of any personally-identifying information, and they were assigned random numbers from 1 to 93. The raters were not advised as to the order in which the posts were written, so as not to influence them unduly to assign higher scores to texts written at later stages of the research. In addition to the posts, the three writing prompts, the revised rating scale, and a spreadsheet, labeled "Rater Marking Sheet," were uploaded to password-protected folders on Google Drive. Raters were familiarized with the three tasks and the rubric via lengthy

electronic correspondence and a set of ordered instructions. The instructions asked the raters to assign five scores (one for every category) to each post, with a maximum score of 4 per category and a maximum total point value of 20. Raters were also advised to avoid recording half-scores, or values such as 1.5, 2/3, and 3+. Only one of the four raters approached me with questions during the rating process, and each rater completed the evaluation of academic blog posts independently. Time for task completion ranged from five to ten hours per rater, and each rater was provided a modest stiped for their contribution. Just as the identity of the second language writers was withheld from the raters, the raters' scores on individual posts were not revealed to the student writers. Similar to the raters in Phase II, the Phase III raters signed consent forms acknowledging their willingness to participate.

4.4.4 Data Analysis

The third phase of the research involved the analysis of longitudinal written data collected over the course of nearly two calendar years to document the ways in which repeated interaction with a genre-specific rubric influenced participant performance on an online, genre-based task over time. To strengthen the credibility of any findings concerning longitudinal changes in written performance, I utilized three modes of inquiry—two quantitative methods and one qualitative method—to analyze the data. Specifically, I utilized rater scores of participant texts, nine lexical and syntactic indices of writing quality, and documentation of individual rhetorical moves to provide a thick description of any longitudinal development in second language writing performance across the rubric and non-rubric groups.

4.4.4.1 Rater Scores. Essentially, the purpose of the rater scores in this phase of the project was twofold: to examine the internal structure of the revised rubric and to identify whether or not there was a statistically significant difference in participant written performance

across the rubric and non-rubric groups. A two-way Analysis of Variance (ANOVA) was run on the mean values of participant scores on each of the three posts to determine whether or not there existed a statistically significant difference in scores assigned to texts written by participants in the rubric group versus the non-rubric group over time.

4.4.4.2 Lexical and Syntactic Analyses. The purpose of this analysis was to examine the lexical and syntactic development of participants' written work over the course of two years. The corpus used for this analysis included 93 academic blog posts written by 31 English as a Foreign Language learners enrolled at the University of Murcia in Southern Spain. All texts in the corpus were converted to text files, and any spelling or grammatical anomalies were left intact. Specific indices corresponding to lexical diversity, lexical sophistication, syntactic complexity, and syntactic sophistication were selected for their saliency in previous research within this area.

The moving-average type-token ratio (MATTR) was selected as a measure of lexical diversity based on findings presented in Zenker and Kyle (2021). MATTR was calculated using the Tool for the Automatic Analysis of Lexical Diversity (TAALED 1.4.1), which is capable of calculating a range of traditional and more robust indices of lexical diversity (Kyle et al., 2020). Given the variety of indices available for measuring lexical sophistication, a representative sample of six indices was selected. The first index was a measure of word frequency, which tallies the instances of use of a particular word and its grammatical inflections (Cobb & Horst, 2015). Specifically, content word frequency was measured in relation to samples of written English collected by the British National Corpus. In addition to content word frequency, lexical decision time, a popular measure of lexical sophistication in psycholinguistics, was selected. Next, two measures of contextual distinctiveness, McD and USF, were included along with two dependency bigram strength of association indices, noun-adjective and verb-direct object

dependency bigrams. These measures were selected based on the results presented in Kyle and Eguchi (2021), which identified a five-component regression model accounting for 23% of the variance in rater scores of participants' argumentative essays. The dependency bigram strength of association indices were calculated via python scripts, and McD, USF, lexical decision time, and content word frequency were calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES), version 2.0 for Mac (Kyle et al., 2018).

The final two indices selected for this analysis provide a measure of syntactic complexity and a measure of syntactic sophistication. Due to the recurrent nature of DC/C as a measure of syntactic complexity and given its established role in predicting longitudinal growth in second language writing, DC/C was included in the calculations. To complement DC/C, a measure of syntactic sophistication, main verb frequency (MVF), was selected due to its significant contribution to Kyle et al.'s (2021) model of written longitudinal development. Both DC/C and MVF were calculated with the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC), version 1.3.8 for Mac (Kyle, 2016).

In conclusion, in an effort to select judiciously from the wide range of linguistic variables available for calculation, I chose nine indices corresponding to measures of lexical diversity, lexical sophistication, syntactic complexity, and syntactic sophistication based on their saliency in previous literature. These indices included MATTR, LDT, CWF, McD, USF, noun-adjective bigrams, verb-direct object bigrams, DC/C, and MVF. Line graphs depicting the change in the mean score for each variable were analyzed before identifying potentially significant interactions between a particular variable and time. Where a visual inspection of the line graph warranted a further look, a one-way repeated measures ANOVA was used to see if there was a significant change, as measured by the variable in question, in participant written

development over time. Finally, seeing as this analysis was also meant to enrich the data provided by rater scores of participant performance in the "language use" category of the rubric, a correlation matrix was produced to look for patterns among the linguistic variables. A subsequent logistic regression analysis determined the degree to which participant scores on the above-specified indices could predict rater scores in the language use category.

4.4.4.3 Qualitative Analysis. For the qualitative analysis of academic blog posts, a form of genre analysis known as rhetorical moves analysis was employed. Rhetorical moves analysis is an investigative tool useful for uncovering key features of familiar or unfamiliar genres. Although several techniques exist for conducting a genre analysis, a rhetorical moves analysis aims specifically to discover preconceived expectations surrounding required and optional *moves*. Rhetorical moves describe the action being accomplished by a particular portion of text, and in this sense, moves are often labeled with action verbs in the "-ing" or progressive form (Jacobson et al., 2021). In addition to identifying moves that are common across sample texts, a rhetorical moves analysis can bring to light various optional moves, which provide opportunities for writers to express their individuality and creativity. In other words, the result of a rhetorical moves analysis is not a prescriptive formula for composing an article or paper, but rather a text-driven approach for describing the ways in which different authors accomplish their goals within the bounds of a recognizable genre.

Seeing as Jacobson et al. (2021) recommended collecting five to 10 samples of the genre in focus, I randomly selected six academic blog posts from six unique longitudinal participants (two from the non-rubric group and four from the rubric group). Two of the texts were written in response to the first prompt on technology and language learning, two of the texts were written in response to the prompt on religion in Spain, and two of the texts were written in response to

the prompt on the usefulness of homework. Whereas rhetorical moves analyses have generally been conducted on the basis of successful samples within a given genre (e.g., accepted conference proposals, published academic articles, etc.), non-standard, or non-target-like examples can be appropriate for critiquing the genre and for revealing the functions performed by certain obligatory or more common moves.

The length of a move can vary from a clause to an entire paragraph (Jacobson et al., 2021), and each move may consist of multiple *steps*, or text fragments, which articulate in detail the means by which a move is achieved (Moreno & Swales, 2018). For example, to design a coding scheme that described the various moves and steps present in the discussion sections of a pre-selected corpora of research articles in English, Moreno and Swales (2018) determined that the move titled, "drawing implications," consisted of three steps. These steps included making recommendations for future research or practice, suggesting the applicability of results or usability of outcomes, and hypothesizing for future research. In addition, Moreno and Swales recommended beginning a rhetorical moves analysis at the step level, before proceeding to the level of the move. Ultimately, their research led to the redefinition of a step as a portion of text that contains "new propositional meaning from which a specific communicative function can be inferred at a low level of generalization by a competent reader of the genre and is perceived as essential to advance the text in the direction expected to achieve the purpose(s) of the (part-)genre" (p. 49). Following this definition, prepositional phrases, parts of quotations, and elaborations would be not considered steps as they do not move a text forward. However, *announcing* functions, or signposts that guide a reader into a subsequent section of the text, and *elaborating* functions, which continue an idea through the use of examples and clarifications,

were included in their final model. In other words, announcing and elaborating functions are identified during the analysis, even though they are not technically considered steps.

In light of the above recommendations, after selecting the texts for analysis, a spreadsheet was created in which the six samples were set as the column delimiters and individual steps were referenced by rows (Appendix N). Next, I proceeded through each academic blog post, identifying the steps that distinct fragments of texts were assumed to accomplish. Only verbs in their "-ing" form were used to describe individual steps, and whenever possible, I utilized the exact wording provided by Moreno and Swales (2018) in their coding scheme for communicative functions in research article discussions. For example, steps common to both projects included "reporting background information without citations," "exemplifying what has been stated in a previous proposition," and "clarifying what has been stated in a previous proposition." On the other hand, it was necessary to identify several steps exclusive to the genre of the academic blog post, such as "addressing potential readership," "introducing main topic," and "restating position on topic." Upon completion of the step analysis, each step was labeled as obligatory, common, or optional. For a step to be labeled as obligatory, the action had to have been identified in at least five of the six texts. A step understood as common was a function found in three or four texts, and an optional step was a rhetorical action identified in only one or two of the six sample texts. Finally, all of the obligatory and common steps were organized sequentially in a table. From this vantage point, the progression of overarching rhetorical moves became apparent, and in turn, descriptions of each move were added to the table.

In sum, a Mixed Methods Research design served as the methodological framework for the project, which was divided into three phases. The first phase covered the development of an analytic rubric for an online, genre-based writing task. During this phase, six experts reviewed

148 academic blog posts, and the qualitative and quantitative data that they provided served as the basis for creating an initial rating scale. The second phase of the project documented the revision of the initial rubric, which was based on the analysis of rater assigned scores to 163 academic blog posts. Finally, the third phase of the research reported on the longitudinal examination of the rubric's impact on participants' second language writing development. This final phase included the collection of participants' second and third blog posts, the analysis of raters' rubric-based scores of the longitudinal posts, the calculation of several linguistic indices, and the rhetorical moves analysis of six sample texts. This detailed account of the participants, the materials, the procedures, and the analyses relevant to each of the three phases underlies the presentation of results in the following two chapters.

CHAPTER 5: RESULTS PHASES I & II

This chapter presents the results from the first two phases of the study. It begins with an examination of various descriptive statistics, including the length of participant blog posts and the ranking of those texts per reviewer judgments. Next, the chapter explains the manner in which quantitative data obtained through this process, along with qualitative reviewer comments, led to the development of rating scale criteria. After presenting the steps involved in the creation of the rubric, rater scores of 163 academic blog posts are examined through a many-facets Rasch analysis. The results of this analysis respond directly to the first and second primary research questions, which aimed to identify and to substantiate the features that should be included in an analytic rubric assessing learner performance on an online, genre-based writing task. The analysis also provides necessary validity evidence in support of the use of a revised rubric in Phase III of the research.

5.1 Descriptive Statistics

While writing their blog posts, several participants asked about an appropriate number of words for their compositions. This information was kept out of the instructions leading up to the prompt, since the ideal length for an academic blog post had not yet been determined by previous literature. Rather than provide participants with a suggested word count for their posts, the writing task left the question of length open to participants' interpretation of the audience and context. As a result, it is not surprising that the descriptive statistics included in Table 5.1 indicate an extremely large range between the word count of the shortest text (108) to the longest text (2504). Closer inspection of the frequency data, however, reveal that only one post was over 1,000 words: a text with 2,504 words. Since this outlier contributed to the high values for kurtosis (60.04) and skewness (6.47), descriptive statistics for the average word count were run a

second time without Participant 21, producing the results shown in Table 5.2. Although Participant 21 remained part of the data set, the second set of descriptive statistics reveal that this individual was in fact a "word count outlier," as the new kurtosis value was within an acceptable range for normally distributed data. At the same time, with or without Participant 21, the descriptive statistics reveal a positive skew to the data, meaning that the distribution curve had a longer tail to the right, or to the positive direction.

Table 5.1

*Descriptive Statistics for Blog Post Word Counts (*N*=148)*

Mean	*SD*	Min	Max	Skewness	Kurtosis
350.41	222.37	108	2504	6.47	60.04

Table 5.2

*Descriptive Statistics for Blog Post Word Counts Without Participant 21 (*N*=147)*

Mean	*SD*	Min	Max	Skewness	Kurtosis
335.76	133.44	108	992	1.31	0.40

From a quantitative perspective, the next point of interest would be the reviewers' estimation of the texts. A simple tally of the number of posts that each reviewer assigned to Folders 1 through 6 revealed individual biases in text evaluation. As seen in Table 5.3, the first reviewer classified a comparatively large number of posts as successful, placing 34 texts in Folder 6. On the other hand, Reviewer 2 assigned more posts to the lower three folders, with only 27 total posts (a mere 18%) deemed appropriate for inclusion in Folders 5 or 6. This result would suggest that the mental criteria according to which Reviewer 2 sorted the posts were rather strict. Another noteworthy finding is that Reviewer 5 appeared to have forgotten that a minimum of four texts were to be placed in each of the folders, as per written reviewer instructions. Along with assigning only three texts to Folder 1 and three texts to Folder 6, this

reviewer included the most posts in Folders 3 and 4, indicating that this expert may have been averse to assigning extreme scores to student texts.

Table 5.3

Reviewer Distribution of Texts by Folder Number

Reviewer	Folder 1	Folder 2	Folder 3	Folder 4	Folder 5	Folder 6	Total
1	9	23	38	20	24	34	148
2	16	41	41	23	19	8	148
3	7	10	48	54	20	9	148
4	9	24	44	46	20	5	148
5	3	22	51	51	18	3	148
6	8	21	40	38	30	11	148

As can be deduced from the information provided in Table 5.3, no two reviewers were in complete agreement on the relative quality of all 148 posts. Therefore, to gather more information concerning the degree to which the reviewers agreed on the distribution of texts—also known as interrater reliability—intraclass correlation coefficients (ICCs) were calculated in SPSS. According to McGraw and Wong (1996), ICCs are "a measure of the proportion of a variance (variously defined) that is attributable to objects of measurement, often called targets" (p. 30). For this research, I used a two-way random effects model to calculate interrater reliability because each post was rated by the same set of independent raters who were drawn from a population of possible raters. SPSS allows users to select between consistency and absolute agreement measures for the two-way random effects model, and since systematic differences in reviewers' judgments of participant texts were considered relevant to the research, the intraclass correlation coefficients using absolute agreement were selected for this study. The average measures ICCs revealed an interrater reliability of 0.85. Of note is the much lower value obtained for single measures (0.48), which suggests that if only one individual had reviewed the blog posts, reliability would have been compromised.

Table 5.4

Reviewer Intraclass Correlation Coefficients Using an Absolute Agreement Definition

	Intraclass	95% Confidence Interval		*F* Test with True Value 0			
	Correlation	Lower Bound	Upper Bound	Value	*df*1	*df*2	Sig
Single Measures	0.48	0.41	0.56	7.00	147	735	0.00
Average Measures	0.85	0.80	0.88	7.00	147	735	0.00

Note. Two-way random effects model where people effects and measures effects are random.

Although interrater reliability was high, the next point of interest for the present analysis was to determine the number of constructs, or factors, that could account for the main sources of variation among reviewers. For this task, an exploratory factor analysis was conducted using the principal component analysis (PCA) factor extraction method. The first step involved running a set of factor analysis descriptive statistics, including the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity (Table 5.5). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy shows whether or not the data set is suitable for factor analysis, with KMO values ranging from zero to one. The closer the value is to 1.00, the greater the factorability of the data set. The suggested cutoff for determining whether or not a data set is factorable is 0.60, and the KMO value for the reviewer data was both above 0.60 and close to 1.00 (0.88), and Bartlett's Test of Sphericity reported that the null hypothesis was to be rejected. These measures indicated that the data were indeed good candidates for factor analysis.

Table 5.5

KMO and Bartlett's Tests

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.88
Bartlett's Test of Sphericity	Approx. Chi-Square	377.91
	df	15
	Sig.	0.00

The next step, the principal component analysis, provided output in the form of a scree plot (Figure 5.1), a table explaining the variance for which different factors could account (Table

5.6), and a rotated component matrix (Table 5.7). Brown (2009) outlined several rules for determining the number of factors in a particular data set. The first estimate, Kaiser's stopping rule, would recommend that researchers consider factors with eigenvalues over 1.00. In this case, only one factor had an eigenvalue higher than 1.00. The second guideline centers on an examination of the scree plot, which would recommend stopping the analysis at the point where a drastic change in slope was first noticed. As is visible in Figure 5.1, only one factor is located firmly "on the mountain," or before encountering the scree, or debris at the bottom of the mountain. Nevertheless, the scree test is not always an exact measure, and the interpretation of the plot involves human judgment (Tabachnick & Fidell, 2013). Thus, a third consideration for determining the total number of factors is the proportion of variance explained by adding successive components. Whereas a six-factor solution would account for 100% of the variance, a one-component solution would explain only 60% of the variance among reviewers. However, by adding a second factor, another 12% of the variance can be explained. As a result, a two-factor solution was chosen for this data set. Since the ratio between the eigenvalue of Factor 1 (3.62) and the eigenvalue of Factor 2 (0.71) was greater than four, an orthogonal rotation—Varimax—was requested. Loadings of variables on the two factors indicated that Reviewers 3, 4, and 6 loaded with the first factor, and Reviewers 1, 2, and 5 loaded with the second factor. These results suggest that the variation in the reviewers' sorting behavior loaded differently. In other words, the variation in the sorting behavior of Reviewers 3, 4, and 6 was distinct from the variation observed in the sorting behavior of Reviewers 1, 2, and 5.

Figure 5.1

Scree Plot from Principal Component Analysis of Reviewer Folder Assignments

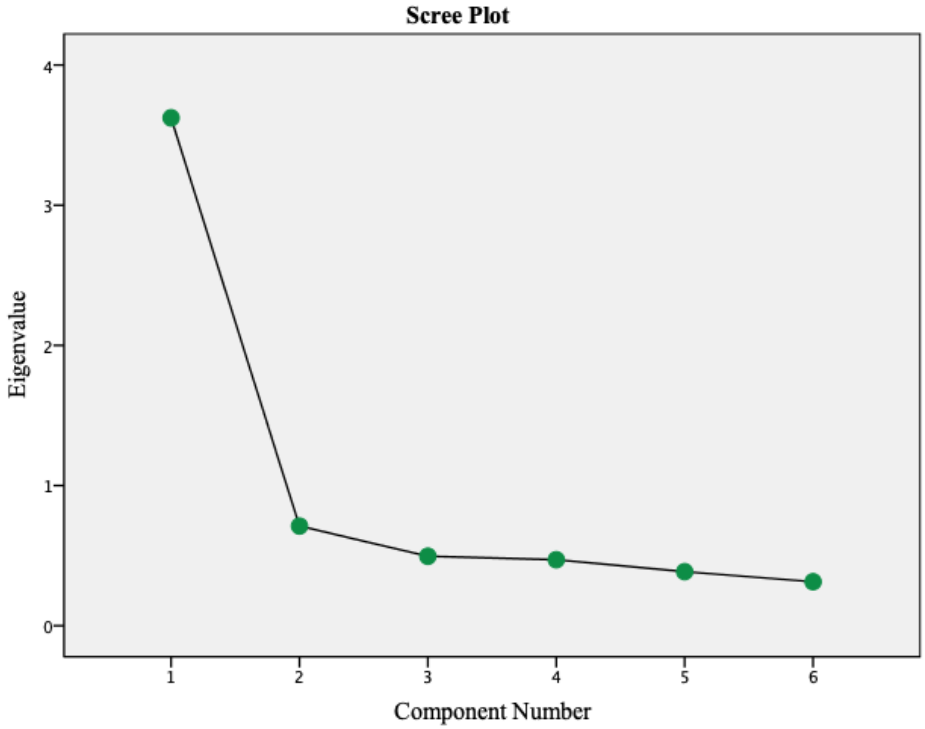


Table 5.6

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.62	60.39	60.39	3.62	60.39	60.39
2	0.71	11.85	72.24			
3	0.50	8.26	80.50			
4	0.47	7.85	88.35			
5	0.39	6.43	94.77			
6	0.31	5.23	100.00			

Note. Extraction Method: Principal Component Analysis.

Table 5.7

Rotated Component Matrix (Varimax)

Rater	Component 1	Component 2	h^2
R1	0.12	**0.91**	0.84
R2	0.54	**0.61**	0.66
R3	**0.79**	0.24	0.68
R4	**0.69**	0.48	0.71
R5	0.52	**0.67**	0.72
R6	**0.84**	0.16	0.73

Note. Extraction Method: Principal Component Analysis.

5.2 Analysis of Reviewer Comments

To undertake a systematic analysis of reviewer comments, an average folder score was calculated for all 148 texts. The mean folder score was 3.52 with a standard deviation of 0.98 and normal values for skewness and kurtosis. Appendix J displays the frequency count of the average folder scores, and Figure 5.2 depicts the distribution of those scores in the form of a histogram overlaid with a normal curve. According to the histogram, the dispersal of average folder scores approximated a normal distribution. Consequently, if the posts were to be dispersed across six levels of performance in accordance with values suggested by a normal distribution, four posts would be allocated to Folders 1 and 6, 20 posts to Folders 2 and 5, and 50 posts to folders 3 and 4. However, so as not to assign participants with the same average score to different groups, 4, 15, 53, 50, 20, and 6 posts were allocated to levels one through six, respectively. In other words, 4 posts were considered representative of Level 1, 15 posts of Level 2, 53 posts of Level 3, 50 posts of Level 4, 20 posts of Level 5, and 6 posts of Level 6. Table 5.8 includes this information along with the corresponding score range for each level.

Figure 5.2

Histogram of Average Folder Scores

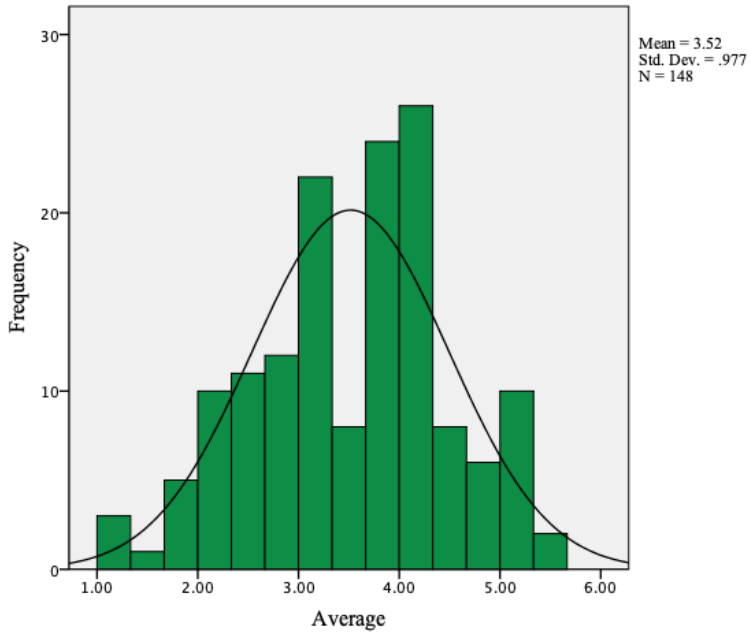


Table 5.8

Distribution of Posts per Level

Level	Posts	Score Range
1	4	1.00-1.50
2	15	1.51-2.33
3	53	2.34-3.50
4	50	3.51-4.33
5	20	4.34-5.00
6	6	5.01-6.00
Total	148	

The next stage involved an electronic redistribution of the 148 texts by level according to the average folder score. Then, reviewer comments associated with the texts in each level were grouped together. For example, four student posts constituted the lowest performance level—Level 1. Each of these posts received a written comment from each of the six reviewers. Thus, 24 reviewer comments (6 comments per text for 4 texts) became the reservoir from which

descriptive characteristics of posts at this level could be identified. The process was repeated for Level 2, although in this case, there were 90 comments (6 comments per text for 15 texts). For Levels 3, 4, 5, and 6, the process was identical. An academic blog post representative of each level is included in Appendix I.

Each group of comments was examined carefully, and descriptors and foci that materialized across reviewers on multiple occasions within a given level were highlighted. Based on the themes that emerged from the review of thousands of comments, descriptors that referred to similar discourse or textual features were color-coded. As an example, Figure 5.3 contains reviewers' comments on the blog post composed by Participant 23, who had an average folder score that fell within the boundaries of Level 4. Reviewers 1 and 5 pointed out that this participant had not responded to all of the rhetorical cues provided in the prompt (in blue). Reviewers 1, 3, and 6 remarked on features of the text that they recognized as belonging to the blog genre (in green), and Reviewers 1, 2, 5, and 6 mentioned characteristics of the participant's use of English, including vocabulary, idiomatic expressions, syntax, and non-target like forms (in purple).

Figure 5.3

Reviewer Comments on Text of Participant 23

R1: Addresses three of four questions with specific and relevant details / high level vocabulary / blog style writing / active voice

R2: Flows well; Interesting idioms; Ideas need more support

R3: Written as a blog post and poses a question to the readers. Personal opinion on LLT is not given, but an even-handed discussion of pros/cons offered, but without many specific examples

R4: A little contradictory

R5: Prompt was partially answered; many sentences were very confusing and difficult to parse

R6: Refers to a blog post and includes some questions which engage with an audience, but some language errors and the structure let it down a little.

Figure 5.4

Key to Color-coding of Reviewer Comments.

Task Fulfillment & Relevancy

Content

Organization & Balance

Genre Specific Features

Language Use

After coding every set of comments attached to the 148 blog posts, an overarching label was assigned to each of the five colors, or major themes, that emerged from the data (Figure 5.4). The five category labels included task fulfillment and relevancy, content, organization and balance, genre specific features, and language use. Once the categories had been identified, the entire set of reviewer comments was divided further according to category. In other words, the comments had already been sorted by ability level, and at this point of the process, the comments connected to Level 3, for example, were organized into subgroups linked to each of the five categories. These groupings represented each of the cells that would be formed by a five-by-six

analytic rubric (five categories by six levels). In this format, similarities among the reviewers' phrasing came to the forefront, and scale descriptors were pulled directly from the words the reviewers had used. To illustrate this process, Figure 5.5 contains reviewer comments from Level 1 that were associated with task fulfillment and relevancy. Most of the comments in Figure 5.5 report either that the material in the academic blog post was not relevant to the task or that the author had not addressed the prompt's rhetorical cues. In spite of the single comment stating, "answers one to two writing prompt questions," the comments reflective of the majority opinion were used as the source material for drafting individual descriptors. Thus, the criteria selected to reflect the lowest level of performance in task fulfillment and relevancy, stated simply "does not address the questions from the prompt" and "text has no relevance to the prompt." Figure 5.6 shows how the descriptors appeared in the actual rubric. Finally, Reckase (1998) recommended that titles not mislead their users, but rather reflect the content of the rubric as closely as possible in a short phase. Thus, the rubric was titled simply, "Analytic Rubric for an Academic Blog Post." A summary of the procedure utilized to create the data-based analytic rubric is located in Figure 5.7.

Figure 5.5

Reviewer Comments Task Fulfillment and Relevancy Level 1

No relevance to the writing prompt
Attempted to answer the writing prompt, but answers are not relevant
Not all points seem relevant
Answers one to two writing prompt questions
Introduction not relevant to point
Does not address the questions
Does not address the prompt
Most was not relevant to the prompt
Post doesn't address the questions raised in the prompt
Does not address all questions
Doesn't address the prompt well at all
Barely relevant final paragraph

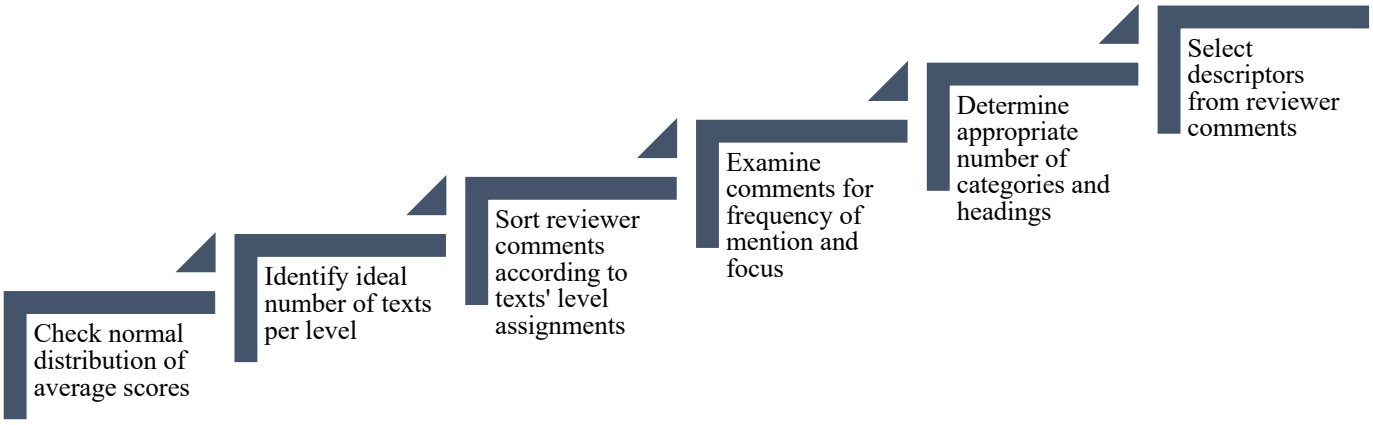
Figure 5.6

Analytic Rubric Descriptors for Task Fulfillment and Relevancy Level 1

	1
Task Fulfillment & Relevancy	Does not address the questions from the prompt Text has no relevance to the prompt

Figure 5.7

Steps Involved in Rubric Creation



5.3 Rubric Revision

Using the first iteration of the rubric, six raters individually scored the academic blog posts written by 163 individuals. These ratings provided a wealth of quantitative data appropriate for an examination of the internal structure of the rubric. Thus, this section reports descriptive statistics surrounding the distribution of rater scores and results obtained through many-facets Rasch measurement using FACETS 3.82 (Linacre, 2019). FACETS measurement reports for examinees, raters, and rubric categories are examined, followed by an analysis of problematic

score levels and instances of rater bias. Taken together, these analyses supplied information necessary for a revision of the initial six-level rubric, whose end result was an assessment instrument capable of supplying relevant, useful information for examinees and examiners alike.

5.3.1 Descriptive Statistics

Rater scores covering five rubric categories (task fulfillment and relevance, content, organization and balance, genre-specific features, and language use) for 163 academic blog posts were first inspected descriptively. The mean category scores assigned by each rater, along with the corresponding standard deviation values are presented in Table 5.9. A close inspection of the descriptive data revealed that the most highly rated category was task fulfillment and relevancy, for which the six raters assigned an average score of 4.30. On the flip side, raters assigned the lowest average scores to the category covering genre-specific features. The mean score for this category, 2.76, was the only mean score to fall below 3.00. Another remarkable feature of this category is that across raters, scores for genre-specific features had the highest average standard deviations, indicating that the raters assigned a wider range of scores for this category as opposed to the other four categories. Although rater-assigned scores are not directly linked to estimates of examinee ability under classical test theory, this result could also suggest that students' utilization of genre-specific features varied more widely than their performance in the other four categories. It is also noteworthy that while Raters 4 and 6 appear to be two of the more lenient raters, they differed in terms of the categories they scored most harshly. For example, the language use category recorded Rater 4's lowest average score but Rater 6's second highest. As a result, a many-faceted Rasch analysis was employed as a means of exploring certain nuances of rater behavior, including rater-category interactions.

Table 5.9

Descriptive Statistics of Rater Scores by Category (N=163)

Category	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6	
	M	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*	***M***	*SD*
Task fulfillment	**3.83**	1.34	**3.64**	0.87	**4.24**	1.03	**5.02**	1.21	**3.89**	0.97	**5.12**	0.86
Content	**3.28**	1.17	**3.26**	0.88	**3.72**	1.23	**4.53**	1.31	**2.46**	0.77	**4.64**	1.05
Organization and balance	**3.61**	1.06	**3.31**	0.83	**3.59**	1.22	**4.48**	1.40	**2.53**	0.83	**4.20**	1.23
Genre-specific features	**1.78**	0.90	**2.05**	0.90	**2.29**	0.96	**4.88**	1.18	**1.53**	0.67	**3.97**	1.11
Language use	**3.61**	1.05	**3.19**	0.85	**3.94**	1.10	**3.91**	1.32	**2.72**	0.93	**4.99**	0.72

5.3.2 FACETS Analysis

Prior to running a many-faceted Rasch analysis, I checked the dataset for unidimensionality. The unidimensionality of a dataset can be examined in various ways, and for the purposes of this research, I ran a Principle Component Analysis (PCA) on the category-specific scores the raters had assigned to the 163 participant texts. For a FACETS analysis to be meaningful, it is necessary to ensure that the primary factor behind the variation in scores among rubric categories account for more than 50% of the total variance. Although the PCA analysis revealed two factors, the first component explained 68.86% of the total variance, thereby meeting an acceptable threshold to consider the dataset unidimensional.

After checking for unidimensionality, rater scores were transferred to a command and data file (Appendix K) necessary for conducting many-facets Rasch measurement via FACETS 3.82 (Linacre, 2019). The first several lines of this data file outlined the specifications for the analysis, including the number of facets under consideration, the range of values expected for each category, and the model by which the program should look for interactions among facets. The three facets selected for this analysis—examinees, raters, and rubric categories—were

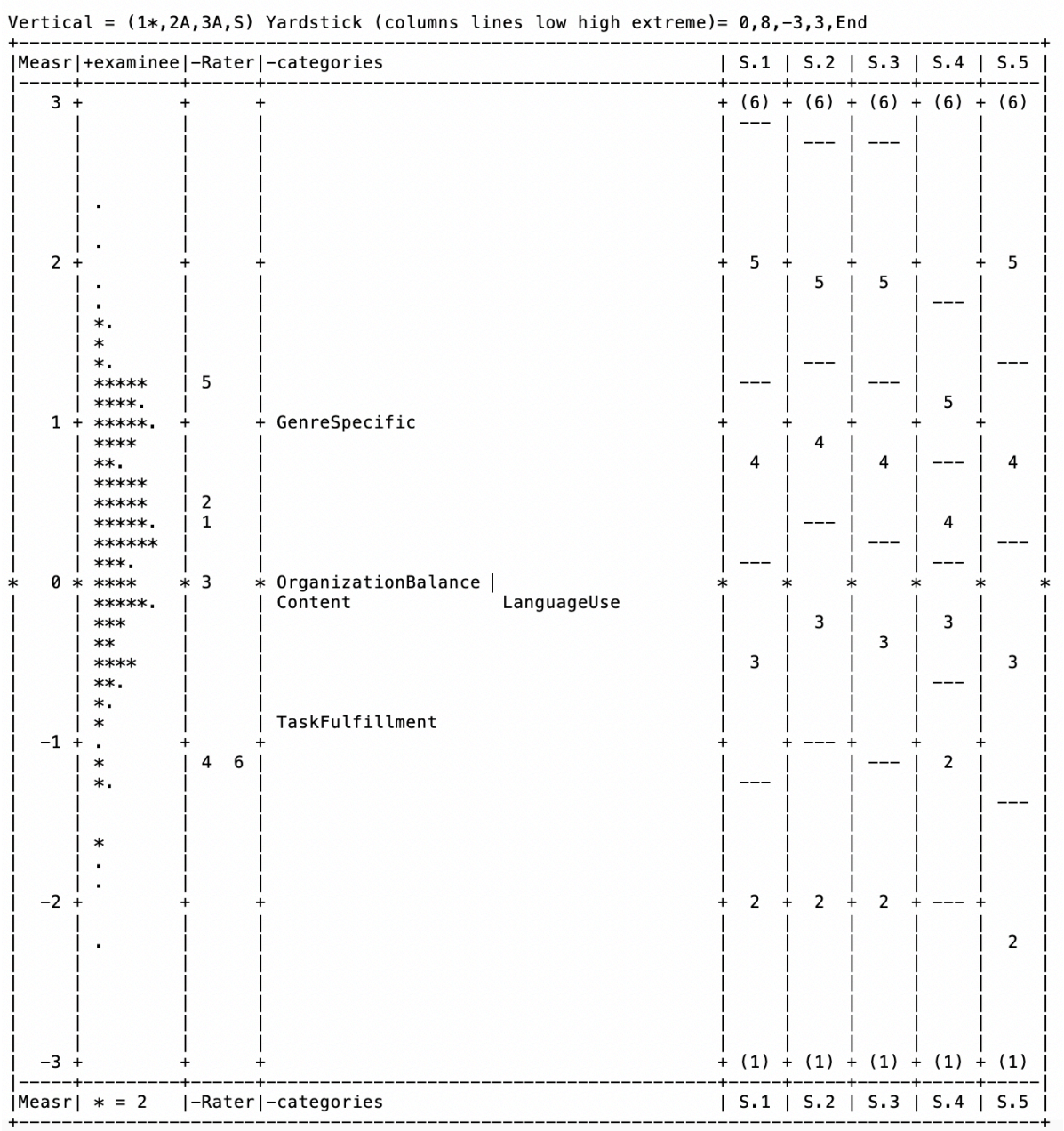
assumed to have impacted the total score assigned to a particular academic blog post by one of the six raters. In the analysis that follows, a graphic representation of all three facets is presented first, followed by a systematic examination of the examinee, rater, and category measurement reports, in that order.

One of the most useful tables resulting from a FACETS analysis of rater data is the Vertical Ruler presented in Figure 5.8. This graphic representation of examinee ability, rater severity, and category difficulty was made possible by the transformation of raw data into individual logit scores, thereby enabling comparisons among model facets against a true interval scale. Per convention, rater severity and category difficulty were centered at zero, and the examinee facet was allowed to float, meaning that participant performance on the task was allowed to vary widely, without expectations for a set mean score. The first column of the vertical ruler contained the logit values according to which estimates of participant ability, rater severity, and category difficulty were measured. The second column ranked examinees according to ability level, with the most able participants located near the top of the scale and the least able participants spread thinly near the bottom. Each dot in this column corresponded to one participant, and every asterisk represented two participants. The third column presented the results of the rater facet, and the fourth column covered the category facet. The specifications for this analysis equated higher logit scores with more severe raters and more difficult categories, whereas comparatively lenient raters and easier categories received lower logit scores. The remaining five columns provided a visual representation of the distance between steps, or levels, for each of the five categories on the rubric. S.1 corresponded to task fulfillment and relevancy, S.2 to content, S.3 to organization and balance, S.4 to genre-specific features, and S.5 to

language use. The significance of the information contained in these five columns is examined in depth in Section 5.3.2.3.

Figure 5.8

Vertical Ruler of FACETS Results for Participant Ability, Rater Severity, and Category Difficulty



Note. The asterisk (*) indicates two participants, and the dot (.) indicates one participant.

5.3.2.1 Examinee Measurement Report. In addition to placing examinee ability, rater severity, and category difficulty on a single measurement scale, a FACETS analysis provides individual measurement reports for each of the three model components. A closer look at the examinee measurement report (Appendix L) revealed several nuances concerning participants' performance on the first online writing task, including the range of ability levels, the degree of separation, and the estimates for examinee infit and outfit. First, the report ordered examinees from the least able to the most able individual. These estimates of participant ability ranged from -2.21 logits to 2.42 logits, with the majority of participants falling between -0.70 and 1.20 logits. Since the wide spread of ability levels (across 4.63 logits) on this task approximated a normal distribution, the separation estimate and the corresponding reliability score provided with the examinee measurement report indicated the approximate number of performance levels into which examinees could be grouped and the reliability of that stratification. Specifically, the reported separation index of 3.75 and the reliability coefficient of 0.93 indicated that participant performance on this online, genre-based task could be separated into approximately four, statistically distinguishable levels, as opposed to the six levels outlined in the analytic rubric.

The examinee measurement report also provided values concerning participant "fit" with the Rasch model. Whereas the outfit statistics provided information regarding the mean squares of unweighted residuals (the difference between the actual and the expected score for each category), infit statistics were weighted by variance within the model, and thus, more useful for analyses of Rasch output data. Seeing as rater assignment of category scores for each of the 163 texts was based on a preliminary, six-level analytic rubric, conservative prescriptions for an acceptable range of infit values (e.g., 0.8-1.2 or 0.5-1.7) were eschewed in favor of the calculation of 95% confidence intervals. This method uses the uses the average of the infit mean

squares along with the standard deviation to identify an acceptable range of fit. The standard deviation, 0.41, was doubled (0.41 + 0.41), and this value, 0.82, was subtracted from the mean of 1.01 and then added to 1.01 to determine the range. Therefore, acceptable values for infit mean squares should have fallen between 0.19 and 1.83, which would equate to within two standard deviations of the mean. An analysis of infit values revealed zero overfitting participants, and seven underfitting examinees. Whereas the Rasch model expects a certain degree of intra-participant variation across categories, an underfitting examinee is indicative of an erratic performance, where the weighted residuals are too high to meet the specifications of the Rasch model. A subsequent review of rater category scores for the texts produced by the underfitting examinees indeed revealed wide variations in rater scores for particular categories.

5.3.2.2 Rater Measurement Report. Whereas the vertical ruler in Figure 5.8 provided a visual representation of rater severity, the rater measurement report supplied additional details concerning the raters' behavior in relation to one another. Table 5.10 presents the logit values for rater severity in descending order, along with the standard error of measurement, and the fit statistics for each of the six raters. With a logit score of -1.14, Rater 6 emerged as the most lenient rater. On the other end of the scale, Rater 5 proved the most severe, with a logit score of 1.25. Interestingly, the gap between Rater 5 and the three moderate raters (Raters 1, 2, and 3), was rather wide, just as the gap between the two more lenient raters (Raters 4 and 6) and the moderate raters spanned more than one logit. Another indication of the large variation in rater severity can be found in the extremely high value for separation, 21.89, with a corresponding reliability measure of exactly 1.00. The fixed, chi-squared statistic was also significant ($X^2 = 2817.3$, $df = 5$, $p < 0.00$), thereby necessitating the rejection of the null hypothesis, or the assumption that the raters were equally severe. An additional point of note, which will be

addressed at length in the discussion, is that the most severe rater as well as the two most lenient raters happened to be non-native speakers of English.

Table 5.10

Rater Measurement Report for Six Raters

	Severity (logits)	Model *S.E.*	Infit MnSq	Outfit MnSq
Rater 5	1.25	0.04	0.79	0.76
Rater 2	0.55	0.04	0.69	0.71
Rater 1	0.43	0.04	1.23	1.26
Rater 3	0.02	0.04	1.02	1.01
Rater 4	-1.11	0.04	1.41	1.41
Rater 6	-1.14	0.04	0.72	0.78
M	0.00	0.04	0.98	0.99
SD	0.87	0.00	0.27	0.27

Note. Reliability of separation index = 1.00; fixed (all same) chi-squared: 2817.3, $p<.001$.

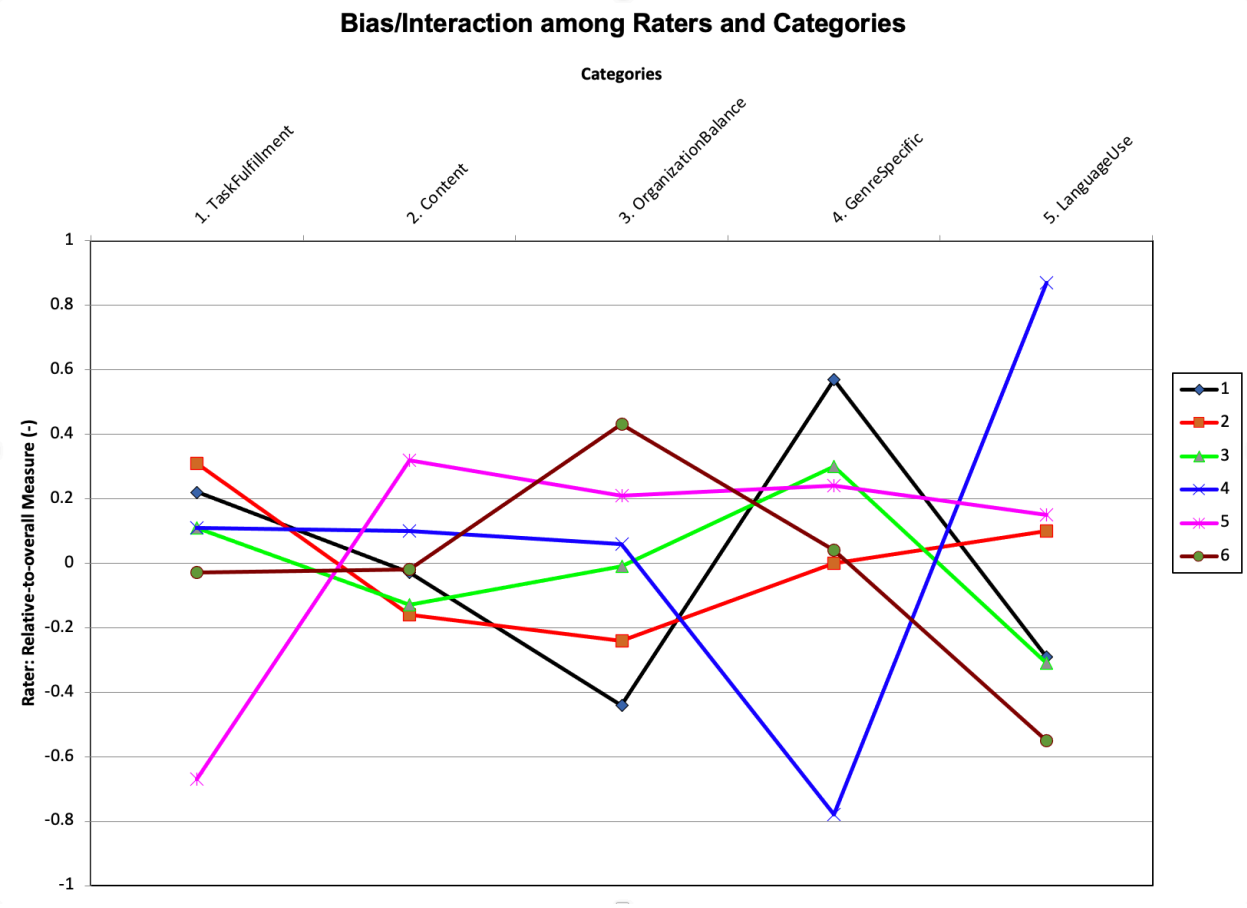
Weigle (1998) explained that within the many-faceted Rasch model developed by Linacre, "rater variation is seen as an inevitable part of the rating process and…is considered actually beneficial because it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and examinee ability on the same linear scale" (p.264). However, whereas inter-rater variation is expected, the hope is that raters will be internally consistent. In other words, although raters will vary in terms of the severity with which they apply rubric criteria, raters who score texts more consistently, whether by consistently assigning respectively higher or lower scores are desirable. Using the same method of interpreting the infit indices as per the examinee measurement report, it became clear that all raters fell within an acceptable range of intra-rater variation, meaning that the raters did not exhibit greater than expected fluctuations in their scoring behaviors.

In order to examine the degree to which raters were biased toward particular categories, a visual representation of raters' interaction with the five rubric categories is presented in Figure 5.9. The graph shows that the six raters treated some categories more evenly than other

categories, and that raters at either end of the leniency-severity continuum scored certain categories more leniently or more harshly than the Rasch model would have anticipated. For example, although Rater 5 was the harshest of the six raters, he was extremely lenient in assigning scores for task fulfillment and relevancy. On the other hand, Rater 4, one of the two most lenient raters overall, scored test-takers more harshly than expected on their language use. Furthermore, whereas Rater 4 was quite lenient in assigning scores for the inclusion of various genre-specific features, Rater 1 interpreted the descriptors for this category in a way that led to the assignment of unexpectedly low scores. Across categories, Rater 2 was the most consistent, which coincided with this rater having the lowest Infit Mean Square value of the group. Perhaps surprisingly, given the diversity of the group of raters recruited for this project, the content category produced the smallest range of bias values. In other words, the raters appeared to have been able to apply the leveled descriptors for this category more consistently than they did, as a group, for the other four categories.

Figure 5.9

Graph of Rater Bias Toward Individual Rubric Categories



5.3.2.3 Category Measurement Report. The category measurement report in Table 5.11 presents the five categories in order of descending difficulty. The high separation index value of 15.82 along with a reliability of 1.00 and a statistically significant chi-squared value indicated that the categories were different from one another. In other words, even though the five categories contributed to the measurement of a single construct, their difficulty values indicated that they measured reliably different aspects of examinee writing performance on the online, genre-based task. Genre-specific features emerged as the most difficult category, with a logit value of 0.98, and task fulfillment and relevancy arose as the category in which raters assigned

the highest scores. The remaining three categories, organization and balance, language use, and content, received logit scores near the mean; however, they still measured different aspects of second language writing performance. Finally, the calculation of 95% confidence intervals using the mean and standard deviation of the values for infit mean squares revealed no misfitting categories.

Table 5.11

Category Measurement Report for Five Categories

	Difficulty logits	Model *S.E.*	Infit MnSq	Outfit MnSq
Genre-specific features	0.98	0.03	1.01	1.03
Organization & balance	0.04	0.04	0.94	0.93
Language use	-0.09	0.04	1.17	1.19
Content	-0.11	0.04	0.72	0.72
Task fulfillment & relevancy	-0.82	0.04	1.14	1.11
Mean	0.00	0.04	1.00	0.99
Std. Deviation	0.58	0.00	0.16	0.16

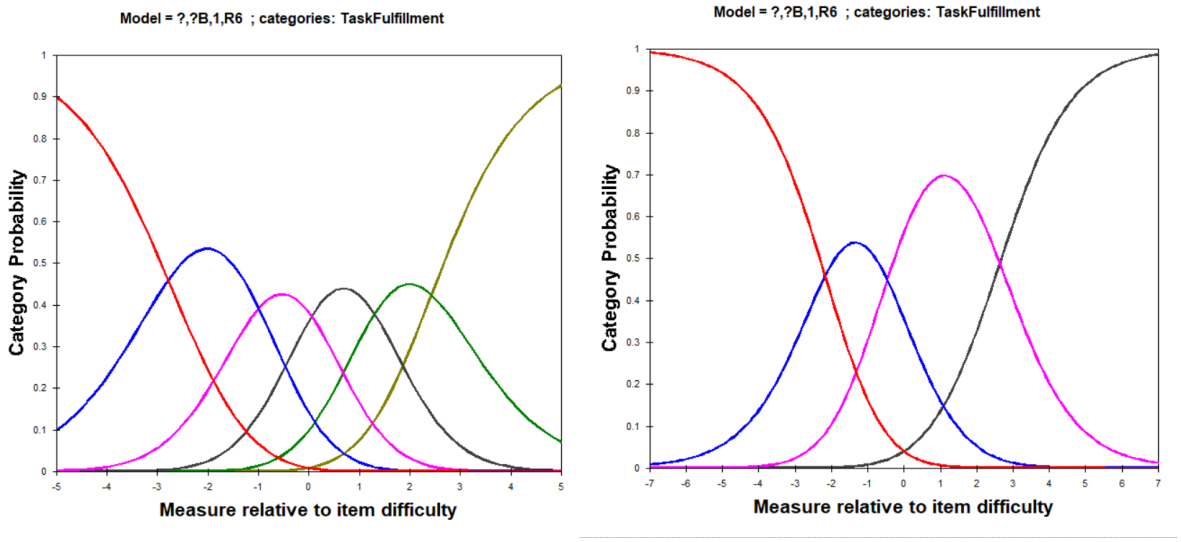
Note. Reliability of separation index = 1.00; fixed (all same) chi-squared: 1359.9, *df*: 4; $p < .001$.

For more information related to category function, FACETS enables an analysis of the degree to which each of the six score levels were utilized for rating participant performance within each category. Category probability curves from a high-functioning scale show clear boundaries between adjacent levels as well as associations between ability levels and observed scores. Thus, category probability curves for each of the rating criteria were examined, and the information gleaned from this analysis, along with the data covered earlier in this chapter, enabled a systematic revision of the original, six-level analytic rubric. In the presentation that follows, a six-level category probability curve is displayed next to a modified, four-level probability curve for each of the five rubric categories, beginning with task fulfillment and relevancy. These visual representations of scale functioning serve as a sign post for category-specific descriptions of the ways in which rubric cells were reduced and modified. In general, I

focused on the largest cases of overlap between adjacent levels. I then worked to remove redundancies within level descriptors, and I eliminated any construct-irrelevant or vague indicators.

Figure 5.10

Category Probability Curves for Task Fulfillment Before and After Level Reduction



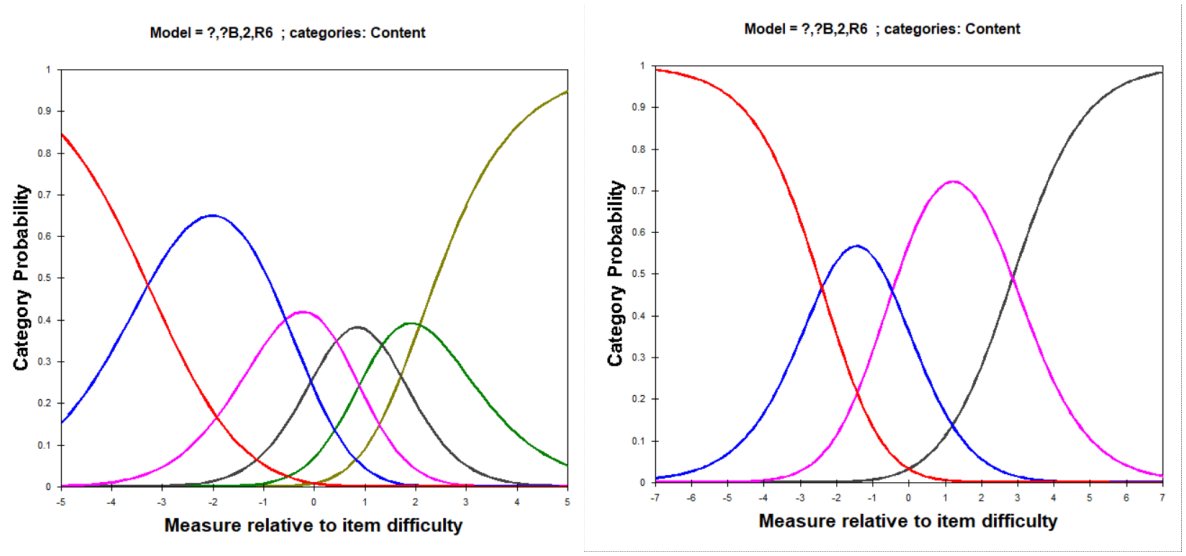
5.3.2.3.1 Task Fulfillment and Relevancy. In reference to the first of the five categories, task fulfillment and relevancy, the category probability curves proved the strongest. Nevertheless, to be consistent across categories, I combined scores of 3 or 4 into a single rating of "3," and I re-coded scores of 5 and 6 as a rating of "4." Although the reduction to four levels resulted in slightly lopsided probability curves, the distinction between adjacent levels was improved. A subsequent review of rubric descriptors revealed instances in which criteria were repeated across levels, and in turn, could be subsumed into a single rubric cell. For example, the phrases, "prompt addressed only minimally" (Level 2) and "prompt addressed in a cursory or glancing way" (Level 3), were extremely similar, not to mention of dubious merit in helping raters to distinguish between levels. As a result, levels two and three were combined into one

cell, with the new descriptor stating, “addresses only one or two prompt questions in a cursory or glancing way.” In addition, the descriptor, “addresses all questions from the prompt,” which originally applied to Level 5 and Level 6, became exclusive of the highest level within task fulfillment, now denoted by a score of 4.

Another opportunity for revision rested in descriptors expressing probabilities. Harsch and Martin (2012) made similar scale revisions in their validation study, as the raters who participated in the project expressed difficulty with interpreting rubric phrases that indicated possibilities, as opposed to descriptors outlining explicit features of sample texts. For example, a descriptor from Level 2, stating “paragraphs *may* not relate to the prompt,” and a descriptor from Level 3, indicating the text “*may* contain some irrelevant information,” were eliminated, with a new descriptor, “likely contains some irrelevant information,” taking their place in the combined cell. As with the probabilistic statements for Levels 2 and 3, the phrase, “*may* contain some irrelevant points and information” from the original Level 4 was dropped entirely, and the criterion, “a few details are not relevant or the connection between those details and the prompt is unclear,” was used as a descriptor of texts meeting the second-highest level for task fulfillment and relevancy on the revised rating scale.

Figure 5.11

Category Probability Curves for Content Before and After Level Reduction

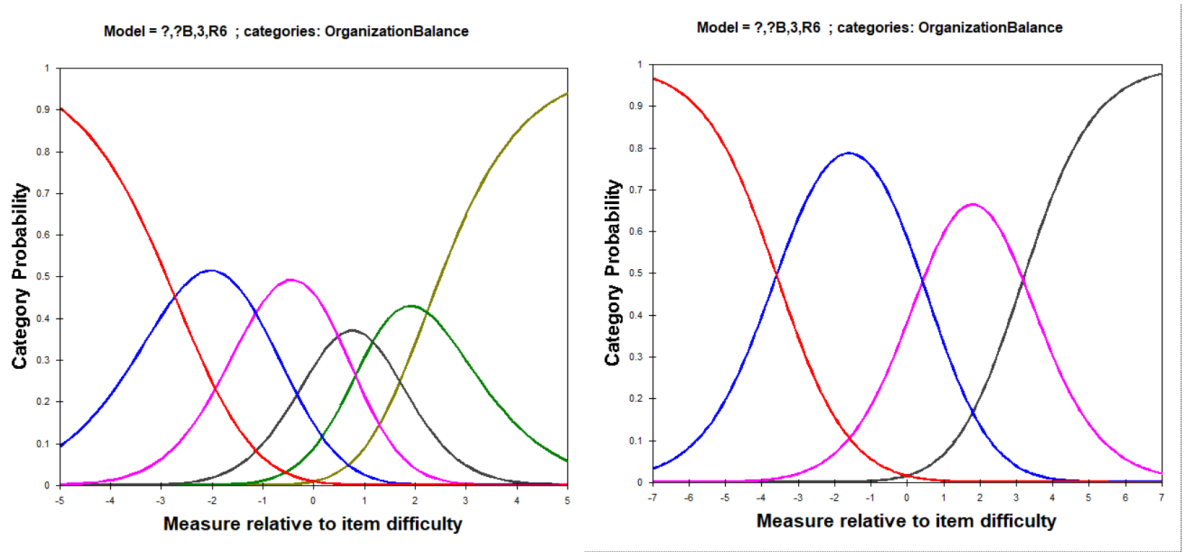


5.3.2.3.2 Content. Rater scores on content were somewhat distinguishable from one another across six levels. Nevertheless, the graph of category probabilities for each level demonstrated a high degree of overlap between consecutive scores, signifying that a performance slightly above the mean, for example, could well receive a score of 3, 4, 5, or even 6. To test the functionality of a reduced scale, rater scores of 4 were recoded as "3" and rater scores of 5 and 6 were assigned a score of "4," the cell corresponding to the highest level of performance on the new rubric. Interestingly, the new curves, though a bit cleaner, still demonstrate overlap, particularly for individuals with ability levels in the middle of the participant pool. In this case, rubric descriptors were reformatted, so as to distinguish more clearly between Levels 2 and 3 on the revised rubric. Whereas descriptors for the lowest and highest score levels remained the same, descriptors corresponding to the original Levels 2, 3, 4, and 5 were scrutinized and substantially adapted for the reduced rubric. Probabilistic statements, such as "*may* contain good ideas" (Level 2), "*may* be a short response" (Level 2), and "one or two ideas *may* be loosely or

partially supported" (Level 3) were deleted entirely. In addition, rather than ask raters to distinguish among "only a few vague or general examples" (Level 2), "provides some clear examples" (Level 3), "concrete examples given, but argument could be strengthened" (Level 4), and "many good points supported with personal examples" (Level 5), two new descriptors were drafted to reflect performances in Levels 2 and 3 on the revised rubric. The Level 2 descriptor for content now reads, "provides a few examples, though the ideas are not developed or explored fully," and the Level 3 descriptor states, "contains good points supported with concrete and/or personal examples, though argument could be strengthened with additional detail." With these revisions, raters will no longer be asked to demarcate the phrases, "a *few* vague or general examples" and "*some* clear examples" on the one hand, and "*concrete* examples" versus "good points with *personal* examples" on the other hand.

Figure 5.12

Category Probability Curves for Organization Before and After Level Reduction

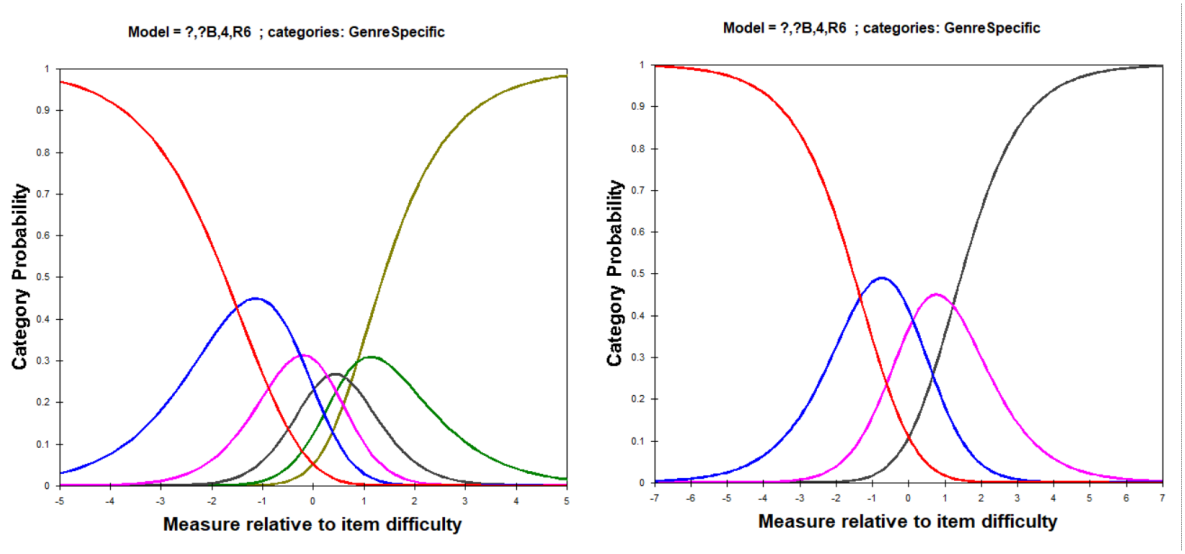


5.3.2.3.3 Organization and Balance. An examination of the original category probability curves for organization and balance revealed a great deal of crowding around Levels 3, 4, and 5.

As a result, I re-coded scores of 1 and 2 as "1," scores of 3 as "2," scores of 4 and 5 as "3," and scores of 6 as "4." Since very few texts had received a score of 1 in this category, descriptors originally listed in the first cell, including "lacks organization" and "a series of one-sentence paragraphs," were left out of the revised rubric. An additional justification for the omission of the Level 1 descriptors rested in the similarity of interpretation between descriptors in Level 1, such as "lacks organization," and descriptors in Level 2, for example "little discernible structure." Therefore, the statements that had originally described a Level 2 performance were moved to the cell defining a Level 1 performance on the revised rubric. Further revisions within this category centered on distinguishing between overall and within-paragraph organization across levels. This distinction, which was present in the descriptors for Levels 3 and 5 in the initial rubric, was addressed on the revised rubric in Levels 2, 3, and 4. Seeing as the original Level 4 had contributed little to separating participant texts according to clear ability steps, most of its descriptors were left out of the new rubric, with the exception being a note about abrupt conclusions, which was added to the Level 3 cell in the revised rubric.

Figure 5.13

Category Probability Curves for Genre Features Before and After Level Reduction

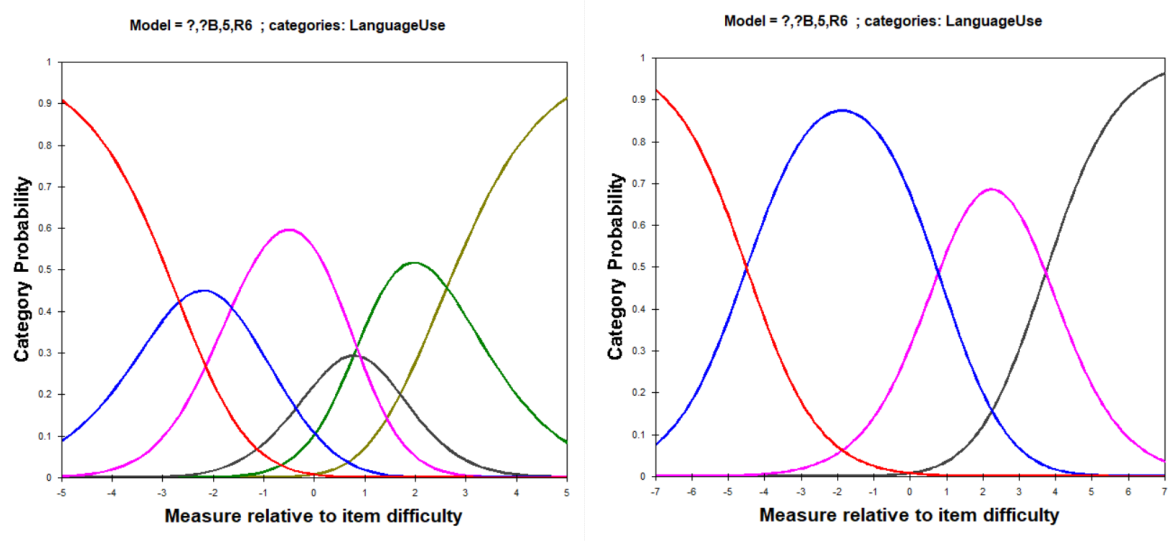


5.3.2.3.4 Genre-specific Features. From the graph of the category curves representing the six levels described on the original rubric, it is possible to see that scores of 4 and 5 were almost entirely subsumed by the surrounding levels. After reassigning a score of "3" to all ratings of 4 and converting scores of 5 and 6 to "4," the category curves achieved a higher level of distinction from one another. In terms of descriptors, Level 1 remained identical, describing a text void of any genre-specific characteristics, such as personal opinions, hyperlinks, or the acknowledgment of a particular readership. The remaining cells were revised to address more clearly—and from the outset—the inclusion of a specific number of features that the reviewers had identified as desirable in various forms of online writing. For example, rather than lead with dissimilar descriptors, such as "apart from use of personal experience…does not contain any genre-specific characteristics" and "acknowledges audience by directly addressing a group of readers or through inclusion of rhetorical questions," each of the four cells in the revised rubric stated, "makes use of [insert: *none, one, two,* or *three or more*] of the following:

acknowledgement of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience /opinion." In this way, the revised rubric should lead to greater rater reliability in assigning scores for the genre-specific features category. At the same time, and as has been pointed out at conferences and in private discussions, the descriptive criteria corresponding to the highest level of this category could pose a challenge to even the savviest of second language academic bloggers. Depending on the nature of a given text, it may not make sense to include all of the specified features in one piece, not to mention that apart from links to multimodal resources, the effective incorporation of visual or auditory modes of communication are not addressed.

Figure 5.14

Category Probability Curves for Language Use Before and After Level Reduction



5.3.2.3.5 Language Use. In this set of graphs, Level 4 was entirely subsumed by Levels 3 and 5, thus making the level superfluous. To re-test the categories, scores of 1 and 2 were coded as "1," scores of 3 were coded as "2," scores of 4 and 5 were coded as "3," and scores of 6 were converted to "4," the highest score on the new rubric. In line with the reduction in levels, the

descriptors for the highest and lowest levels of performance remained nearly identical in the revised rubric. Furthermore, most of the descriptors for Level 2 were transferred to the new rubric, with only one substitution. A descriptor from Level 2 on the original rubric had read, "uses some simple structures correctly, but overall lacks syntactic variation;" however, on the revised rubric, the second clause was updated with wording from the original Level 3, "uses some simple structures correctly, but struggles to control complex syntax." The new Level 3 cell contained descriptors found previously in Levels 4 and 5. In addition, a Level 5 descriptor related to the use of idiomatic expressions was dropped, since this facet was not addressed in other cells in the category. The Level 5 descriptor, "minimal difficulty with complex syntax," was also dropped as this phrase could have been used to describe written performances at the highest level. In its place, the original Level 4 descriptor, "some mixture of syntactic structures, though not error-free," was included in the Level 3 cell in the revised rubric.

5.3.3 Summary

This extensive examination of the initial rubric's internal structure led to a reduction of the number of scoring levels—from six to four—and revisions to the wording of individual descriptors. These adjustments to the rubric provide further validity evidence of the revised scale's representation of the emerging genre of the academic blog post. Two of the reviewers from Phase I were consulted for an assessment of the revised rubric (Appendix M), and both individuals asserted that the rating scale accurately reflected the levels of performance and the criteria that they had observed in the data. As Harsch and Martin (2012) reported, "the construct is operationalized in the rating scale categories (or assessment criteria), and most directly in the descriptors" (p. 234), which reinforces the importance of the revision process as well as the value of investing in the collection of validity evidence before implementing a particular assessment

tool. Ultimately, the revised rating scale represented the overlapping of several layers of input. Actual samples of genre-based writing were superimposed by the expert judgments of trained teacher-scholars and the mixed-methods analysis of quantitative and qualitative data. This comprehensive approach to rubric revision should result in an improved description of the online, academic blog post and an increase in rating quality. Even though the rubric validation process is not yet concluded, the work explained in this chapter provides important validity evidence supporting the rubric's depiction of various levels of performance within the academic blog, or discussion post genre. Additional discussions around validity will take place in the third phase of the research, where rater behavior in line with the revised rubric is examined.

CHAPTER 6: RESULTS PHASE III

In this chapter, I present the results of the longitudinal portion of the study. The data set consisted of 93 academic blog posts written by 31 participants. Each of the 31 participants produced three academic blog posts spaced evenly apart over the course of approximately two years. In order to answer the final research question regarding the ways in which the longitudinal development of the EFL writers differed between learners who had access to the revised rubric and learners who wrote their posts without the guidelines provided by the rating scale descriptors, participants' texts were examined via rater scores, a selection of variables related to syntactic and lexical sophistication, and an analysis of rhetorical moves. The chapter begins by presenting the results of a two-way Analysis of Variance (ANOVA), followed by a consideration of nine carefully-selected linguistic variables and a rhetorical moves analysis. Both the rhetorical moves analysis and the linguistic analysis serve as a triangulation of the significant results obtained via the two-way ANOVA. Seen together, the mixed-methods longitudinal results present a robust picture of the written development of a group of foreign language learners within a newly defined online genre.

6.1 ANOVA Results

Potentially the most important component of this project centered on whether or not the provision of a well-designed rubric would have a significant impact on the writing performance of English as a Foreign language learners within an emerging digital genre. To answer this foundational question, a mixed between-within-participants ANOVA, also known as a two-way ANOVA and as a repeated measures factorial design, was utilized. Fundamentally, analysis of variance involves comparing two estimates of variation: the variance among scores within each group and the differences among group means. In this case, any differences among group means

would reflect one or both of the treatment conditions (use of the rubric and time) compared to the appropriate error term. Inclusion in the rubric versus the non-rubric group constitutes the between-participants independent variable, and the time of composition serves as the within-participants, repeated independent variable (IV). Whereas the within-participants IV contains three levels (Time 1, Time 2, and Time 3), the between-participants IV has only two levels (no rubric and rubric). The null hypothesis states that the mean scores of participants in the rubric group would be equal to the mean scores of participants in the non-rubric group at Time 2 and Time 3, and alpha was set at 0.01. A visual representation of the six cells in this two-way ANOVA is presented in Table 6.1, with P_n corresponding to a randomly assigned participant number from 1 to 31. It is worthwhile to point out that the raters responsible for assigning a score to each post were not aware of the time at which a particular text had been written, nor were they informed of the group membership (rubric vs. non-rubric) of the post's author.

Table 6.1

Assignment of Participants in Between-Within-Participants ANOVA

		Time		
		T1	T2	T3
Groups	No rubric	P_1-P_{16}	P_1-P_{16}	P_1-P_{16}
	Rubric	P_{17}-P_{31}	P_{17}-P_{31}	P_{17}-P_{31}

Prior to analysis, the data were screened for accuracy and missing values. The statistical program SPSS was used to evaluate whether or not the data met the necessary assumptions for running a two-way repeated-measures ANOVA. First, each independent variable consisted of two or more groups, and the dependent variable, average score, was continuous in nature. In addition, the dependent variable approximated a normal distribution in each of the cells formed by the interaction of time and use of the analytic rubric. Descriptive statistics for the data housed

within each cell are shown in Table 6.2, where it is possible to see that although no cell boasted values for skewness and kurtosis equal to zero, the ratios of skewness to its standard error and kurtosis to its standard error were all within an acceptable range for treating the data as normal ($-2 < rT(\frac{S}{ses}\ or\ \frac{K}{sek}) < 2$). In addition to verifying the normality of the data set, homogeneity of variance was assessed by Levene's Test of Equality of Variances. Since none of the values were significant, the null hypothesis—that the error variance was equal across groups—stood. Homogeneity of variance-covariance matrices was confirmed via a non-significant result for Box's M, and seeing as the within-subjects independent variable consisted of three levels (Time 1, Time 2, and Time 3), sphericity was verified using Mauchly's Test of Sphericity. The value for Mauchly's *W* was not significant, indicating that the data also exhibited homogeneity of covariance.

Table 6.2

Descriptive Statistics for Rubric and Non-Rubric Groups by Time (1, 2, or 3)

Group	Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
No Rubric	1	16	14.06	2.06	10.00	16.67	-1.04	0.56	-1.86	0.05	1.09	0.04
	2	16	14.75	1.61	11.33	18.00	-0.33	0.56	-0.59	1.14	1.09	1.05
	3	16	15.52	1.43	13.00	18.00	0.17	0.56	0.30	-0.68	1.09	-0.62
Rubric	1	15	13.71	2.58	8.67	17.33	-0.27	0.58	-0.47	-0.93	1.12	-0.83
	2	15	17.96	1.36	15.33	19.67	-0.69	0.58	-1.19	-0.46	1.12	-0.41
	3	15	18.00	1.87	14.00	20.00	-0.93	0.58	-1.60	0.04	1.12	0.04

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

A mixed between-within participants ANOVA was performed using IBM SPSS Repeated Measures General Linear Model. The ratio between the two estimates of variance described above provided an F value, which was tested against critical values for F, to identify any effects of the rubric, of the time of composition, and of the interaction between the rubric and time. In Table 6.3, sums of squares (*SS*) refers to the "sums of squared differences between scores and

their means" (Tabachnick &Fidell, 2013, p. 38). It is also important to note that in estimating the effect of the rubric, the error value is attributable to variation among participants in each group, whereas for the within-participants analysis, variation arises due to the effect of time and to the interaction of time with the rubric.

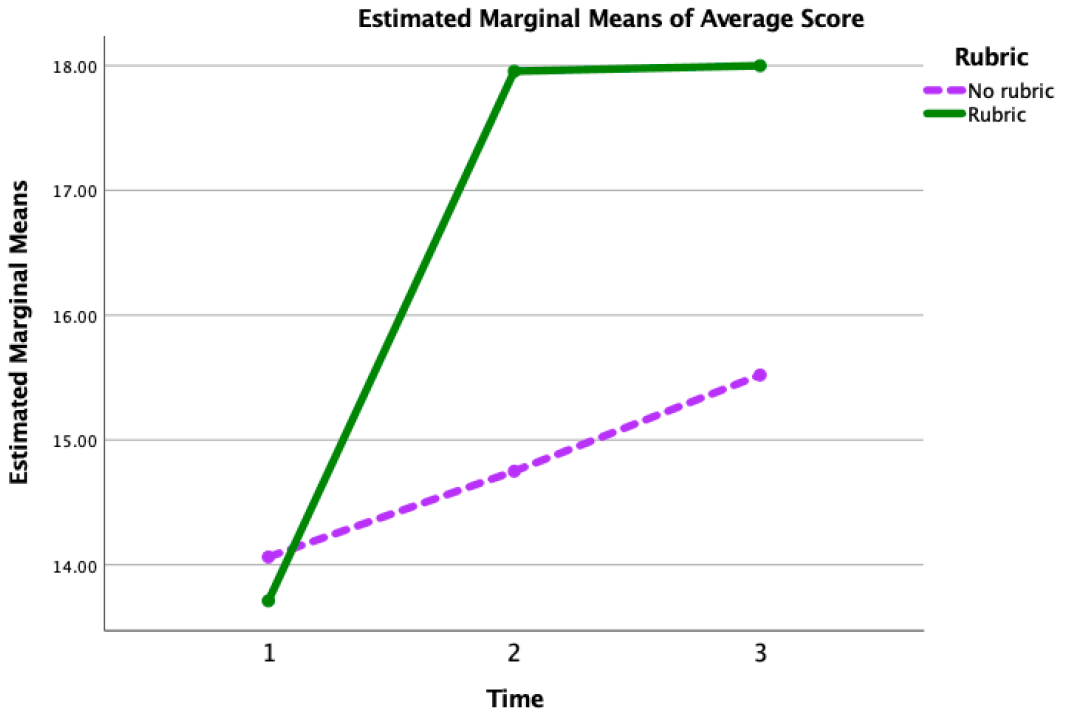
Table 6.3

ANOVA Source Table for Average Scores by Time and Rubric

Source	*SS*	*df*	*MS*	*F*	*p*	Partial Eta Squared	Power
Between-Participants Effects							
Rubric	73.41	1	73.41	16.58	< 0.001	0.36	0.98
Error	128.37	29	4.43				
Within-Participants Effects							
Time	149.73	2	74.86	25.03	< 0.001	0.46	1.00
Time*Rubric	54.69	2	27.34	9.14	< 0.001	0.24	0.97
Error	173.50	58	2.99				

Figure 6.1

Estimated Marginal Means of Average Score Over Time



Since the *p* values for the between- and within-participants effects and their interaction were all significant, the null hypothesis stating that the means across groups were equivalent was rejected. The dependent variable, average score, was significantly affected by use of the rubric, by time, and by the interaction of time and rubric. A partial eta-squared of 0.36 for the between-participants variable indicated that 36% of the between-participants variance in the scores assigned to texts written by participants in the rubric versus the non-rubric group could be attributed to the presence or absence of the rubric. Individuals in the rubric group performed significantly better on the online, genre-based task than their peers in the non-rubric group. In terms of time, 46% of the within-participants variance in average scores across groups was attributed to the time of composition and 24% of the within-participants variance was ascribed to the interaction of the two independent variables: time and use of the rubric. As demonstrated in

the graph of the estimated marginal means of average score over time (Figure 6.1), once participants in the rubric group had been acquainted with the revised rating scale—after the composition of their first post and before the composition of their second post—they outperformed by a significant margin a group of learners who had scored higher initially. Furthermore, whereas on average individuals in the non-rubric group received continually higher scores across iterations of the task, the interaction between the passage of time and the availability of the rubric enabled participants in the rubric group to produce academic blog posts that aligned more closely to data- and expert-derived descriptors of successful performance in that genre. In spite of the modest sample size, the two-way repeated-measures ANOVA had remarkable power. The effects of the presence or absence of the rubric had a power of 0.98, and the within-participants effects had a power of 1.00 for time and 0.97 for the interaction of time and the presence or absence of the rubric.

6.2 Linguistic Analysis

In order to explore further the written development of the writers in both the rubric and non-rubric groups, several language-specific aspects of participant texts were examined via nine linguistic indices. The indices were selected in terms of their representativeness of distinct aspects of language and due to their saliency in previous research. In reference to the lexicon, one index measuring lexical diversity, MATTR, along with six variables appraising different aspects of lexical sophistication (LDT, CWF, McD, USF, noun-adjective bigrams, and verb-direct object bigrams) were selected. In addition, the ratio of dependent clauses to the total number of clauses, DC/C, was selected as a representative of syntactic complexity, and main verb frequency served as a quantitative measure of syntactic sophistication. The natural language processing tools created by Kristopher Kyle, including TAALES (Kyle et al., 2018), TAASSC

(Kyle, 2016), and TAALED (Kyle et al., 2020), were used to calculate numerical values for each of the aforementioned linguistic indices for each text in the corpus (*N*=93). Index-specific data were examined by time of composition for normality, and line plots of the mean values with 95% confidence intervals enabled a visual appreciation of any potentially significant changes over time.

6.2.1 Lexical Diversity

The index chosen as a measure of lexical diversity in participant texts was the moving-average type-token ratio, or MATTR. The expectation for this index, as well as for the other linguistic variables examined in this chapter, was that learners' written language would become more diverse (or more complex, or more sophisticated) over time. Kyle et al. (2021) explained that in regard to language development, "the default hypothesis has been that proficient learners will produce more linguistically complex structures as a function of language proficiency" (p. 2). In the case of MATTR, the values should rise over time as learners incorporate an increasing number of unique words into their online compositions. Whereas the descriptive statistics for MATTR revealed a relatively normal distribution of values for lexical diversity at each collection point (Table 6.4), Figure 6.2 shows that the mean values remained stable across time. Furthermore, there were no observable differences in mean values for MATTR between the rubric and non-rubric group in spite of the fact that the revised rubric referenced variety in word choice within the cell corresponding to the highest possible value in the language use category.

Table 6.4

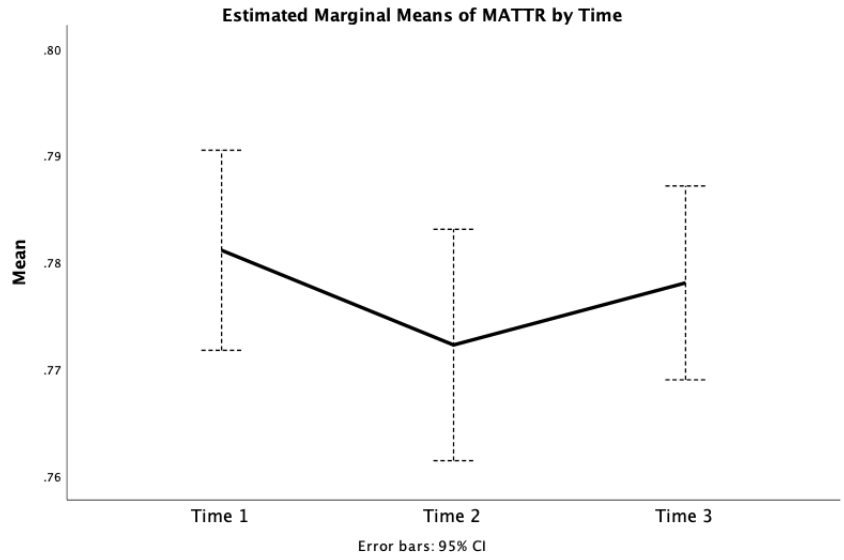
Descriptive Statistics for MATTR by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	0.781	0.026	0.738	0.825	0.131	0.421	0.311	-0.872	0.821	-1.062
2	31	0.772	0.030	0.709	0.810	-0.790	0.421	-1.876	-0.325	0.821	-0.396
3	31	0.778	0.025	0.711	0.843	-0.167	0.421	-0.397	1.459	0.821	1.777

Note: ses = standard error of skewness; sek = standard error of kurtosis

Figure 6.2

Estimated Marginal Means of MATTR Over Time



6.2.2 Lexical Sophistication

In an attempt to address the various ways in which lexical sophistication has been theorized, six unique indices were calculated. Regarding the first index, lexical decision time, the expectation was that values would increase as learners produced more sophisticated texts. The gradual, though not significant increase in values for lexical decision time observed here suggest that learners' written production moved in the anticipated direction (Figure 6.3). The second index selected as a measure of lexical sophistication was content word frequency. In contrast to the expectation for lexical decision time, the hope here was that learners would utilize words less

frequently encountered in the reference corpus over time. Nevertheless, in spite of the normality of the data set (Table 6.6), the English learners appear to have used more frequently encountered words, according to the British National Corpus, in their second and third posts (Figure 6.4).

Table 6.5

Descriptive Statistics for LDT by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	629.459	6.153	617.571	640.715	-0.033	0.421	-0.078	-0.597	0.821	-0.727
2	31	630.075	5.568	617.871	643.212	0.113	0.421	0.268	0.409	0.821	0.498
3	31	630.929	6.737	617.836	645.067	0.406	0.421	0.964	-0.379	0.821	-0.462

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Figure 6.3

Estimated Marginal Means of Lexical Decision Time Over Time

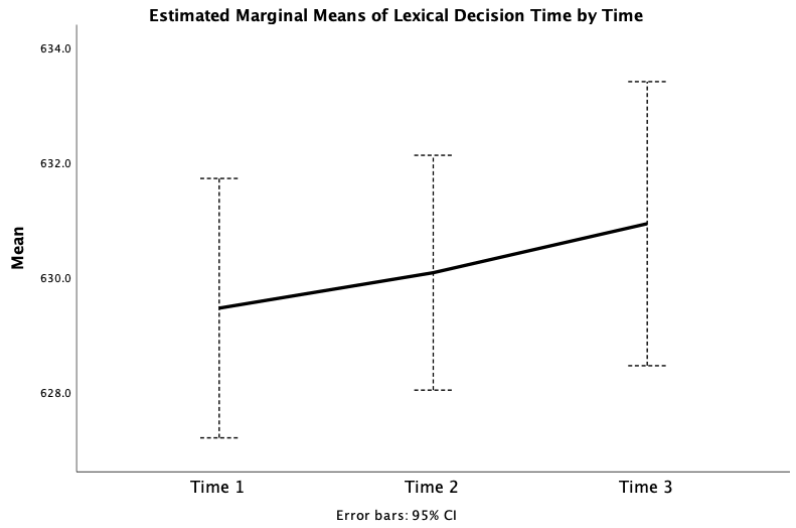


Table 6.6

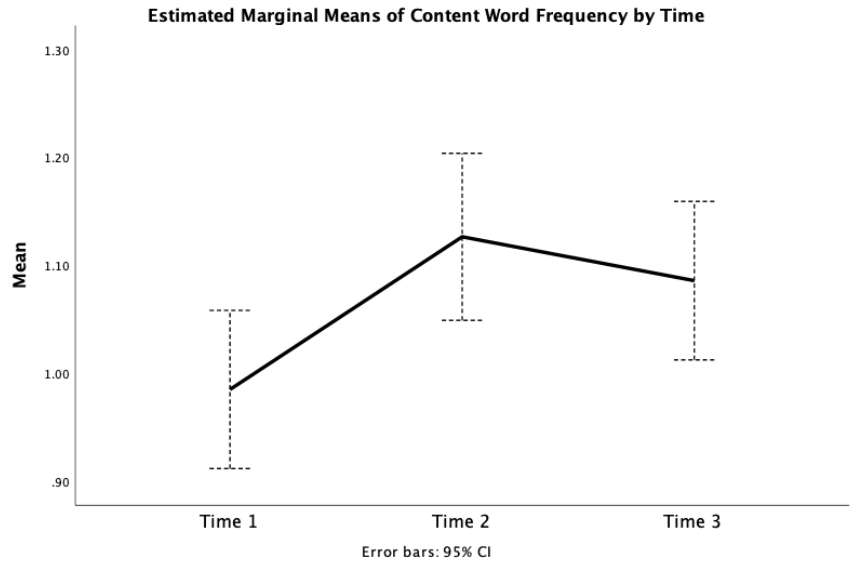
Descriptive Statistics for CWF by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	0.985	0.200	0.574	1.524	0.639	.421	1.518	1.027	0.821	1.251
2	31	1.127	0.211	0.767	1.642	0.300	.421	0.713	-0.421	0.821	-0.513
3	31	1.086	0.200	0.789	1.447	0.106	.421	0.252	-1.232	0.821	-1.500

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Figure 6.4

Estimated Marginal Means of Content Word Frequency Over Time



Still under the umbrella of lexical sophistication, the next two indices, McD and USF, provided measures of contextual distinctiveness for each of the 93 electronic texts. The numerical values furnished by the corpus-based index created by McDonald and Shillcock (2001) should increase as second language writers incorporate words that are more contextually distinct. In contrast, the scores supplied by the University of South Florida's behavioral index should decrease as learners utilize words with greater contextual distinctiveness. Interestingly, in this data set, the mean values for McD decreased over time and the mean values for USF

increased over time. In other words, in successive repetitions of the online, genre-based task, participants increased their utilization of words that were less contextually distinct (i.e., words more likely to be produced in a larger variety of occasions). This increase in words of lesser distinctiveness even appeared significant in the case of the USF index, but significance tests were not run as values obtained during the third data collection point violated assumptions of normality (Table 6.8).

Table 6.7

Descriptive Statistics for McD by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	0.936	0.074	0.809	1.054	-0.026	0.421	-0.062	-1.221	0.821	-1.487
2	31	0.905	0.072	0.739	1.046	-0.619	0.421	-1.470	0.182	0.821	0.222
3	31	0.913	0.077	0.776	1.084	0.253	0.421	0.601	-0.110	0.821	-0.134

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Table 6.8

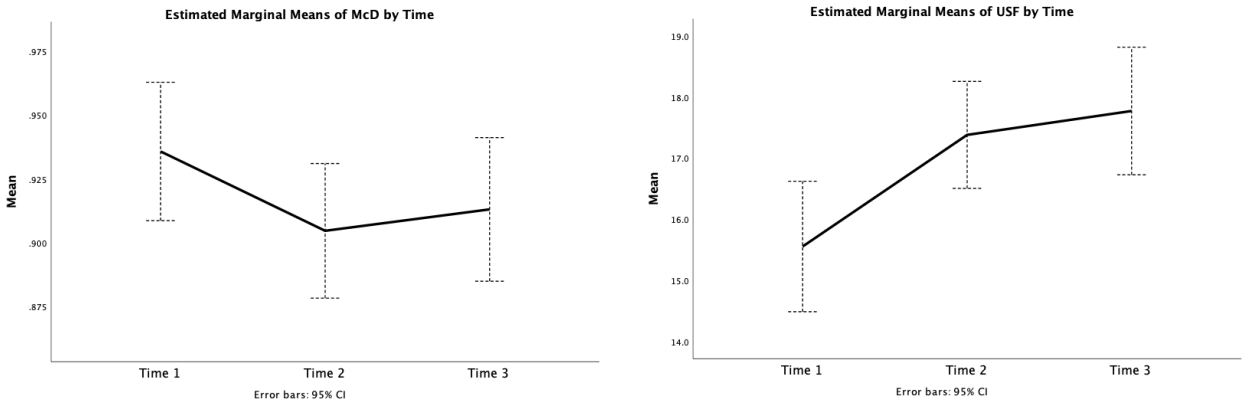
Descriptive Statistics for USF by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	15.557	2.906	11.348	22.569	0.646	0.421	1.534	0.087	0.821	0.106
2	31	17.380	2.382	12.094	21.203	-0.309	0.421	-0.734	-0.855	0.821	-1.041
3	31	17.771	2.843	14.113	26.821	1.331	0.421	**3.162**	2.526	0.821	**-3.077**

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Figure 6.5

Estimated Marginal Means of McD and USF Over Time



The remaining two indicators related to lexical sophistication cover an area beyond the use of less frequent words. Specifically, the noun-adjective and verb-direct object dependency bigram indices measured the degree to which the participants in this study incorporated combinations of words that occur frequently *together*. Seeing as previous studies have reported a positive relationship between phraseological competence, as measured by various dependency bigrams, and writing quality, the expectation for this research was to see an increase in the mean values for both indices over time. This expectation was met in terms of noun-adjective two-word sequences; however, scores related to the use of frequently encountered verb-direct object collocations decreased. The contradiction presented by the results obtained on these indices warranted further examination, particularly seeing as the 95% confidence intervals represented by the dotted lines extending vertically around each mean value in Figure 6.6 did not appear to overlap from Time 1 to Time 2. Thus, two one-way within-subjects ANOVAs were run to identify whether or not the respective increase and decrease in mean values were significant over time. The normality of both data sets was verified using the values for skewness, standard error of skewness, kurtosis, and standard error of kurtosis found in Tables 6.9 and 6.10, and alpha was

again set at 0.01. Furthermore, values for Mauchly's Test of Sphericity were not significant in either case, indicating that each variable exhibited homogeneity of covariance. The results of the two ANOVAs indicated that the within-participants effect was significant, and the null hypothesis stating the means were equivalent across time was rejected. Both dependent variables, noun-adjective and verb-direct object dependency bigrams, were significantly affected by time. The partial eta-squared of 0.19 for noun-adjective bigram scores and 0.25 for verb-direct object bigrams revealed that 19% and 25% of the variance in respective scores could be attributed to the time of composition. The discussion section will address further these significant findings.

Table 6.9

Descriptive Statistics for Noun_Adj_Bigram by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	1.460	0.736	-0.042	3.490	0.490	0.421	1.164	1.027	0.821	1.251
2	31	1.907	0.478	1.170	3.396	0.973	0.421	**2.311**	1.820	0.821	**2.217**
3	31	1.975	0.438	1.191	2.901	0.119	0.421	0.283	-0.666	0.821	-0.811

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Table 6.10

Descriptive Statistics for Verb_DO_Bigram by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	1.743	0.465	0.673	2.540	-0.332	0.421	-0.789	-0.324	0.821	-0.395
2	31	1.272	0.517	0.372	2.953	0.962	0.421	**2.285**	2.243	0.821	**2.732**
3	31	1.318	0.462	0.539	2.299	0.107	0.421	0.254	-0.585	0.821	-0.713

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Figure 6.6

Estimated Marginal Means of Noun-Adj Bigrams and Verb-DO Bigrams Over Time

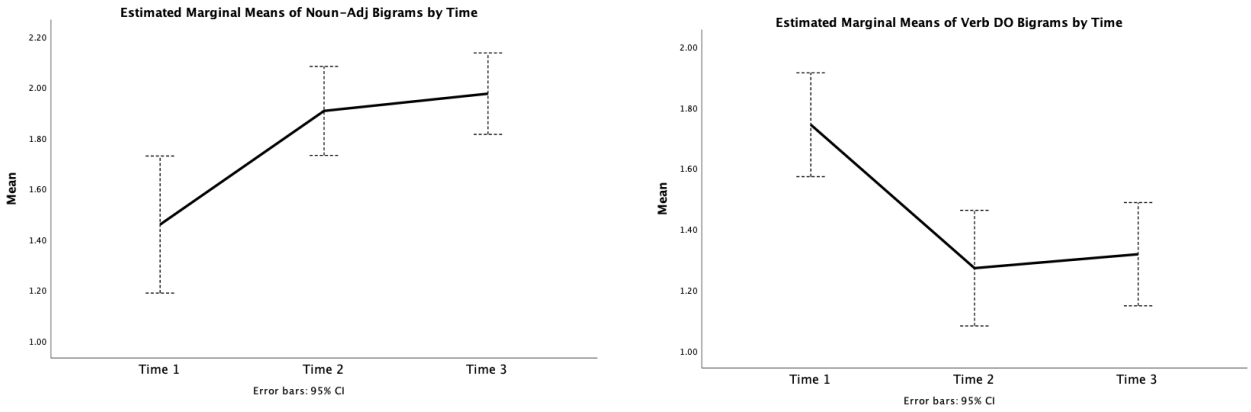


Table 6.11

ANOVA Source Table for Noun-Adj and Verb-DO Bigrams by Time

Source	*SS*	*df*	*MS*	*F*	*p*	Partial Eta Squared	Power
Within-Participants Effects for Noun-Adj Bigrams							
Time	4.87	2	2.43	7.19	< 0.002	0.19	0.92
Error	20.32	60	0.34				
Within-Participants Effects for Verb-DO Bigrams							
Time	4.18	2	2.09	9.83	< 0.001	0.25	0.98
Error	12.77	60	0.21				

6.2.3 Syntactic Complexity and Syntactic Sophistication

A review of the descriptive statistics and line graphs for the index representing syntactic complexity—the ratio of dependent clauses to total clauses—and the index measuring syntactic sophistication—main verb frequency—revealed little change over the course of the project. Although the participants used more dependent clauses as a proportion of the total number of clauses in their texts, the increase in the syntactic complexity of their written production was not meaningful. Similarly, whereas Kyle et al. (2021) found that the grade-school participants used

less frequent main verbs over time, the electronic texts in this data set did not show a comparable developmental trajectory. In fact, the slight change in the mean values of the calculated scores for syntactic sophistication showed a very slight increase in the use of more commonly encountered main verbs.

Table 6.12

Descriptive Statistics for DC/C by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	0.451	0.089	0.250	.615	-0.132	0.421	-0.314	-0.381	0.821	-0.464
2	31	0.480	0.100	0.286	.646	0.040	0.421	0.095	-0.916	0.821	-1.116
3	31	0.474	0.101	0.238	.700	0.081	0.421	0.192	0.883	0.821	1.076

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Table 6.13

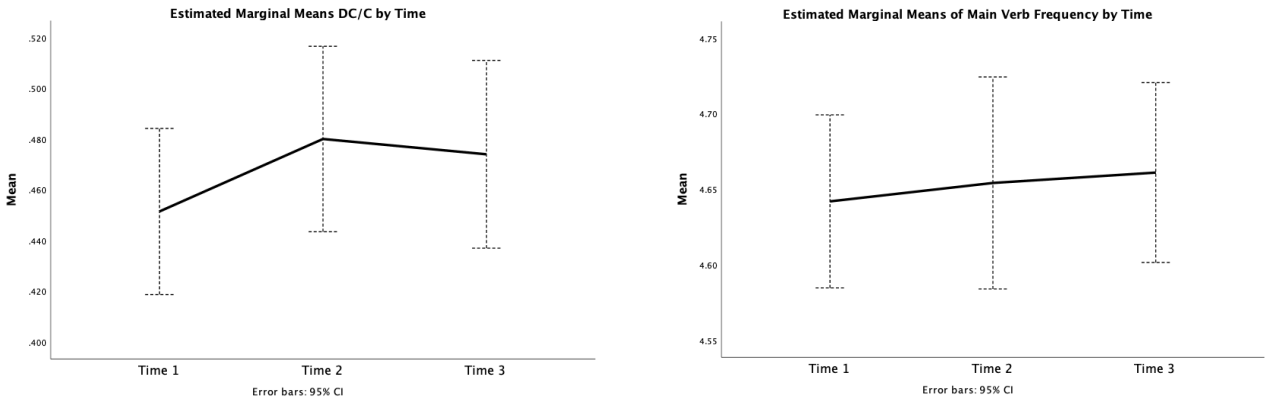
Descriptive Statistics for MVF by Time (1, 2, or 3)

Time	*N*	*M*	*SD*	Min	Max	Skew	*ses*	S/ses	Kurt	*sek*	K/*sek*
1	31	4.642	0.156	4.399	5.007	0.472	0.421	1.121	-0.443	0.821	-0.540
2	31	4.654	0.191	4.329	5.053	0.252	0.421	0.599	-0.526	0.821	-0.641
3	31	4.661	0.162	4.304	5.004	-0.220	0.421	-0.523	-0.093	0.821	-0.113

Note. *ses* = standard error of skewness; *sek* = standard error of kurtosis

Figure 6.7

Estimated Marginal Means of DC/C and MVF Over Time



6.2.4 Correlations Among Linguistic Variables

In light of the previous examination of nine, carefully selected linguistic indices, a Pearson correlation matrix was created to gain a clearer understanding of the ways in which the different variables related to one another, to the average total scores, and to the average language use scores assigned by the raters (Table 6.14). The two indices measuring syntactic complexity and syntactic sophistication did not correlate at all with each other, but rather DC/C had a significant negative correlation with lexical decision time ($r = -0.25$, $p < 0.05$) and MVF had a significant negative association with MATTR, the measure of lexical diversity ($r = -0.29$, $p < 0.05$). In addition, both DC/C and MVF had positive correlations with content word frequency. Of particular interest is the significant positive correlation ($r = 0.24$, $p < 0.05$) between the use of more frequently encountered noun-adjective dependency bigrams and the receipt of a higher total score. The significance in that correlation does not transfer, however, when comparing noun-adjective bigram scores with average scores in the language use category. In terms of scores assessed for language use, only MATTR and one index of lexical sophistication, lexical decision time (LDT), had significant correlations, $r = 0.26$ and $r = 0.24$, respectively ($p < 0.05$). Though not included in Table 6.14, participant scores on the Online Oxford Placement Test were also examined in the matrix, with the only significant correlation arising between the OOPT and average scores for language use ($r = 0.27$, $p < 0.05$). Before investigating further the relationship between the average scores for language use and the three significantly correlated variables, LDT, MATTR, and the OOPT, descriptive statistics for language use were examined for any violations of normality. Since the ratio of the values corresponding to skewness and standard error of skewness was more than -2.0, a histogram of the data was requested, revealing an outlying score of 1.67. Specifically, Participant 18 received an average score for language use of

1.67 on the second academic blog post, when none of the individuals in the longitudinal study scored less than 2.00. As a result, this case was eliminated from the data set, and a new histogram with the remaining 92 data points was pulled (Figure 6.8). A subsequent evaluation of the descriptive statistics showed no violations of normality.

Table 6.14

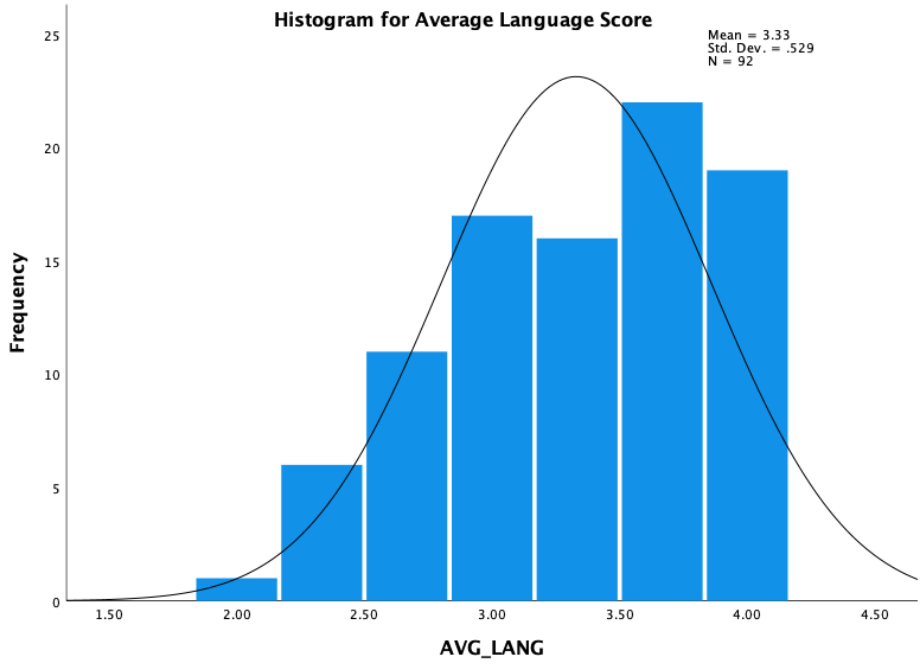
*Correlation Matrix with Linguistic Variables and Average Scores (*N*=93)*

Measure	MATTR	LDT	CWF	McD	USF	N-A	V-DO	DC/C	MVF	Avg Score	Avg Lang
MATTR	1	0.05	***-0.32**	0.01	-0.11	0.07	0.08	0.02	***-0.29**	0.14	***0.26**
LDT		1	***-0.32**	***0.67**	-0.15	***0.38**	0.14	***-0.25**	-0.11	0.01	***0.34**
CWF			1	***-0.34**	***0.33**	-0.03	***-0.47**	***0.20**	***0.52**	-0.01	-0.18
McD				1	-0.07	0.15	***0.28**	-0.11	-0.05	-0.17	0.10
USF					1	0.13	***-0.22**	0.04	0.14	0.17	-0.06
N-A						1	-0.12	0.03	-0.01	***0.24**	0.23
V-DO							1	-0.18	***-0.25**	-0.04	-0.00
DC/C								1	0.02	-0.13	-0.13
MVF									1	-0.06	-0.10
Avg Sc										1	***0.63**
Avg La											1

*$p < 0.05$

Figure 6.8

Histogram of Average Language Score after Removing Outlier



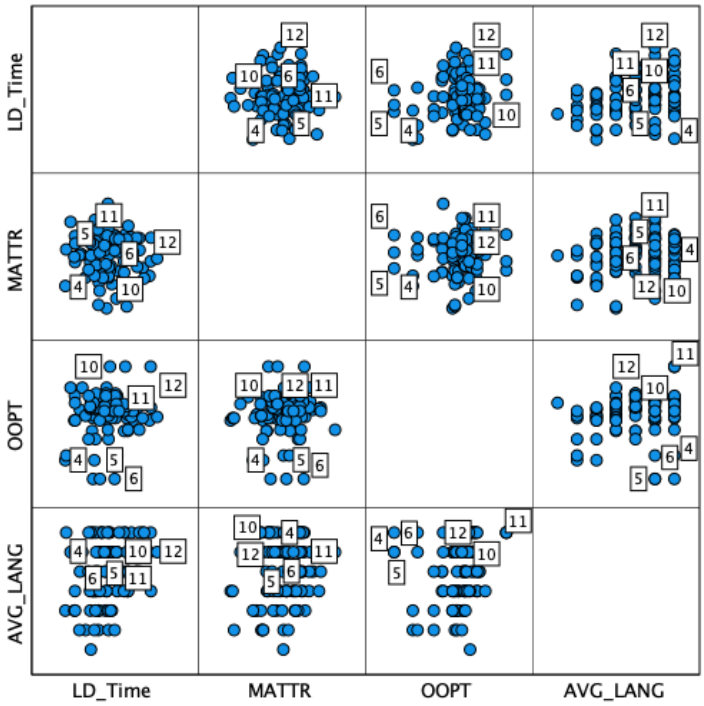
6.2.5 Regression Model Predicting Language Use

The purpose of performing a multiple regression analysis was to investigate the effects of lexical decision time, lexical diversity—as measured by MATTR—and the Oxford Online Placement Test on the scores assigned in the Language Use category of the revised rubric. Before running any statistical tests, I requested a plot matrix with the four variables of interest: LDT, MATTR, OOPT, and AVG LANG (Figure 6.9). Based on the data points in the plot matrix, it seemed that two participants, Participant 2 and Participant 4, corresponding to cases 4, 5, and 6, and cases 10, 11, and 12, respectively were disproportionately affecting the regression results. An analysis of unusual leverage values confirmed this observation, and the two outlying participants were removed from the data set. After a review of the descriptive statistics and scatter plots corresponding to the clean data set, it did not appear that the data had a huge

problem with heteroscedasticity. The scatter plots of "average language score" and each of the dependent variables presented a pattern that could result in heteroscedasticity; however, in regression the primary concern is that the error has constant variance.

Figure 6.9

Scatter Plot Matrix of LDT, MATTR, OOPT, and AVG_LANG Prior to Cleaning



Sequential regression was employed to determine whether the addition of information provided by the Online Oxford Placement Test improved prediction of a score for language use beyond the contribution of an index for lexical diversity and an index for lexical sophistication. Statistics concerning multicollinearity were excellent, with values for tolerance above 0.97 and values for the variance inflation factor (VIF) between 1.00 and 1.03. After the first step, with only lexical decision time as part of the equation, the *R*-squared value was 0.08, $F = 6.97$ (1, 84), $p < 0.05$. After adding MATTR to the regression equation, the R-squared value increased significantly to 0.12, $F = 4.41$ (1, 83), $p < 0.05$, and the addition of the third variable, the OOPT,

raised the value for *R*-squared to 0.19, $F = 6.33$ (1, 82), $p < 0.05$. Since the incremental *F*-test was significant for all three variables, the third model was selected. Table 6.15 displays the results of the incremental *F*-test, and Table 6.16 presents the unstandardized regression coefficients, the standardized regression coefficients, and the semi-partial correlations. The adjusted *R*-squared value of 0.16 indicates that 16% of the variability in language use scores can be predicted by an index for lexical diversity, MATTR, an index for lexical sophistication, LDT, and scores on the OOPT. Finally, the regression assumption of normal distribution of error was confirmed via the histogram of frequency of unstandardized residuals in Figure 6.10 and the normal percentile-percentile (P-P) plot of unstandardized residuals in Figure 6.11.

Table 6.15

Regression Model Summary

Model	*R*	R^2	Adj R^2	*SEE*	R^2 Chg	*F* Chg	*df*1	*df*2	Sig. *F* Change
1	0.28	0.08	0.07	0.51	0.08	6.97	1	84	0.01
2	0.35	0.12	0.10	0.50	0.05	4.41	1	83	0.04
3	0.43	0.19	0.16	0.48	0.06	6.33	1	82	0.01

Table 6.16

Regression Coefficients for Three-Variable Model

	Unstandardized		Standardized			Partial	Part	Tolerance	VIF
Predictor	*β*	S.E.	*β*	*t*	Sig	Correlation	Correlation		
Constant	-14.29	5.69		-2.51	0.01				
LDT	0.02	0.01	0.26	2.58	0.01	0.27	0.26	0.99	1.00
MATTR	3.47	2.00	0.18	1.74	0.09	0.19	0.17	0.98	1.03
OOPT	0.01	0.00	0.26	2.52	0.01	0.27	0.25	0.97	1.03

Figure 6.10

Histogram of Unstandardized Residual

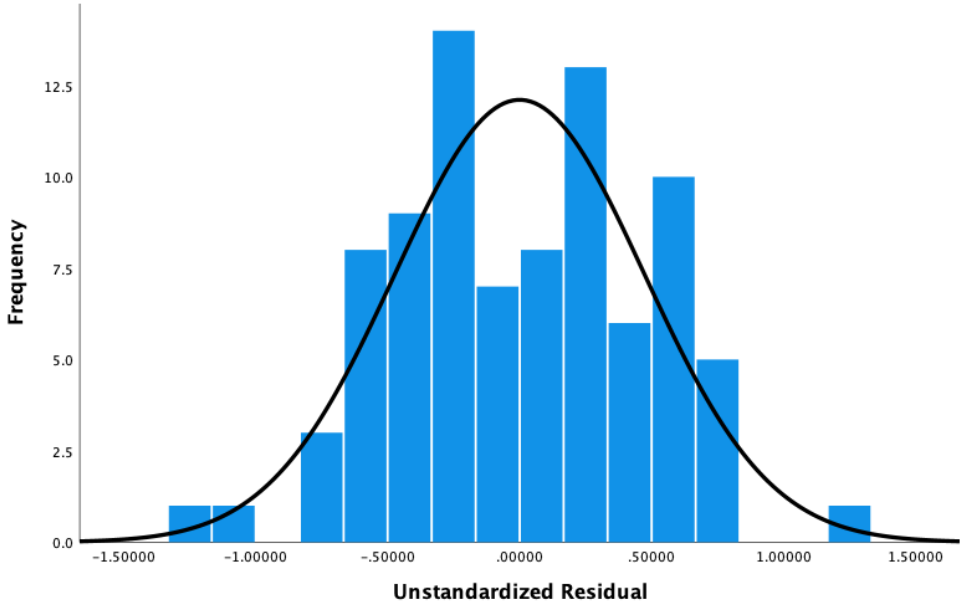
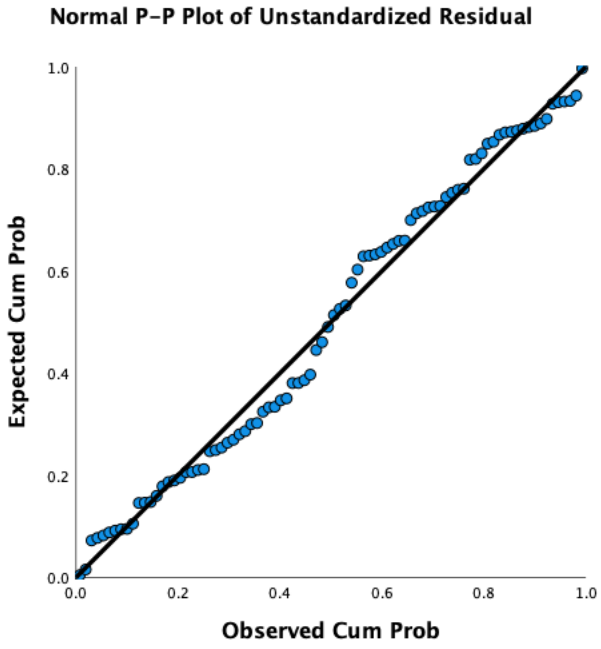


Figure 6.11

Normal P-P Plot of Unstandardized Residual



6.3 Rhetorical Moves Analysis

In order to enhance our understanding of the emerging genre of the academic blog or discussion board post, a rhetorical moves analysis of six sample texts was particularly suited for the task. The complete results of the analysis, in table form, can be found in Appendix N. The following presentation of qualitative findings begins with the elucidation of the steps identified as obligatory or common to the genre. Accompanying these results is an explanation of the conceptualization of overarching moves and an interpretation of the logic behind the inclusion of certain optional steps in the final model. As much as possible, examples stemming directly from participant texts are used to illustrate the progression of rhetorical moves. Table 6.17 presents a summary of the communicative functions found in the genre of the academic blog post.

Table 6.17

Presentation of Communicative Functions in an Academic Blog Post

Communicative Function	Classification	Example
Acknowledging readers	**Move**	
Addressing potential readership	Step (C)	Hello everyone! (S6)
Conveying relevance of topic	**Move**	
Stating importance of topic	Step (O)	Education is a controversial issue in every country (S5)
Clarifying previous proposition	Step (C)	Each person has an idea of what an acceptable education is, but also each human has an opinion about the finest way to educate. (S5)
Summarizing key background information	**Move**	
Reporting background information without citation	Step (O)	Teachers from all levels, from primary to secondary education and beyond, send homework to students so as to help them study whatever it is that they are teaching. (S6)
Introducing main topic of post	Step (O)	This post will be aimed at the influence of technology on these fields (S2)
Reporting additional background information	*Step (Op)*	*The way in which foreign languages are currently taught (and learnt!) has changed drastically over the last 20 years (S2)*
Connecting background information to present composition	*Step (Op)*	*such is the case that technology can't just be left aside" (S1)*
Answering first rhetorical cue	**Move**	
Stating answer to first rhetorical cue	Step (O)	The situation in Spain is quite diverse. (S4)

Presenting first example in support of answer	Step (O)	It is common to see overhead projectors that enable teachers and students to have a different learning experience far from the classic teacher-centered and blackboard-centered second or foreign language class (S1)
Presenting second example in support of answer	Step (C)	We can also see how learners can upload information or tasks for class into different collaborative or open spaces on the internet, such as wikis or forums. (S1)
Providing hyperlink to learn more about subject	*Step (Op)*	*If you are interested in expanding your knowledge on the matter, read this article (it is highly insightful) Church and State in Spain. (S3)*
Answering second rhetorical cue	**Move**	
Announcing transition to next rhetorical cue	Step (C)	You may get the impression that I am against the teaching of religion at high school or school but that is not the case. (S3)
Stating answer to second rhetorical cue	Step (O)	Homework assignments have many positive effects. (S5)
Presenting first example in support of answer	Step (O)	Nowadays, I am teaching an online course and students learn the majority of the content of the class by doing assignments and homework. (S5)
Justifying previous proposition	Step (O)	This mode of learning is flexible since they choose when they do their homework according to their time and pace. (S5)
Presenting second example in support of answer	Step (C)	Moreover, it is helpful because in the actual classes they put in practice what they have learned, in real-life situations and actions. (S5)
Justifying previous proposition	Step (C)	In this case, homework will be useful as long as the student does them. (S5)
Answering third rhetorical cue	**Move**	
Announcing transition to next rhetorical cue	Step (O)	Nevertheless… (S2 & S4)
Stating answer to third rhetorical cue	Step (O)	I can see why people may be against it and it depends on the teacher. (S3)
Presenting first example in support of answer	Step (O)	I remember my religion teacher being all supportive about us teenagers and all the problems that we were facing during this difficult period (S3)
Justifying previous proposition	Step (C)	We even talked about sex in class and there were LGBT people in the class that felt welcomed and found a place to be themselves (S3)
Clarifying previous proposition	Step (C)	So obviously, those aforementioned exercises are the ones teachers should avoid at all costs. (S6)
Presenting second example in support of answer	Step (C)	Additionally, homework can be useless if the student does not do them and they are part of the course (S5)
Exemplifying previous proposition	*Step (Op)*	*There was a gay boy in the class and he was even asked to leave the class. (S3)*
Answering fourth rhetorical cue	**Move**	
Announcing transition to next rhetorical cue	Step (C)	In this sense… (S1)

Stating answer to fourth rhetorical cue	Step (O)	The fact that religion is somehow present and not compulsory in syllabuses certainly helps students in such a complex stage of their lives. (S4)
Presenting first example in support of answer	Step (O)	In my personal opinion, the younger they are, the less homework they should have (S5)
Exemplifying previous proposition	Step (C)	Children need to experience real-life and play(S5)
Clarifying previous proposition	Step (C)	…although homework may help them to learn, it could not be as useful as it is with older people. (S5)
Presenting second example in support of answer	Step (C)	It can open a whole world of possibilities as well… (S1)
Exemplifying previous proposition	*Step (Op)*	*by exposing learners to situations in which real-life language is used and facilitated with subtitles and online videos (S1)*
Evaluating information presented previously	**Move**	
Summarizing overarching stance on topic	Step (O)	To cut a long story short, religion is good for students whether they are believers or not. (S3)
Restating position	Step (C)	As I have mentioned, it can help them understand why they are believers or why they are not and chose the religion that best suits their needs and ideas. (S3)
Clarifying previous proposition	Step (C)	Every student is different and everyone has diverse abilities and characteristics. (S6)
Calling for future research or encouraging reader comments	*Step (Op)*	*Do not hesitate on posting your opinion on the comments! (S2)*

Note. (O) = obligatory; (C) = common; (Op) = optional; (S#) = sample number

The first communicative function identified in the sample texts was the acknowledgment of a specific group of readers. Even though this genre-specific feature was included in the revised rubric, two of the three participants who utilized this function were in the non-rubric group, meaning that they were never advised to employ this trope. Perhaps even more noteworthy than addressing a potential readership is the near universal inclusion of a sentence communicating the importance of the topic about to be discussed. Even the non-rubric participants, who were not alerted to the role of an introductory paragraph in the overall organization of the piece, chose to incorporate this tactful step. For example, the author of Sample 6, a member of the non-rubric group, followed an address to the readers with the concise statement, “Today I am going to discuss with all of you, my readers, an interesting topic related to studying and learning.” Although further elaboration on the interestingness of the topic was

found in only four of the six samples, this step rounded out the move that has been labeled, "conveying relevance of topic."

Whether a result of participant interest in the topics selected for the writing prompts or prior experience with academic writing in English, five of the six sample texts summarized key background information prior to highlighting the main topic of the composition. Although none of the participants supported their assertions with citations or links to outside sources, a sentence articulating at least one bit of widely known information related to the topic emerged as an obligatory step in the model. In addition to the inclusion of background information, nearly all of the texts referenced explicitly the main topic of their blog post. Whereas one author moved directly from introducing the topic, "in this brief entry we will try to give a solid opinion regarding several aspects closely related to religion," to responding to the first rhetorical cue, "the situation in Spain is quite diverse," the other authors included a variety of intermediary steps. Two of the participants chose to report additional background information, and one participant provided an outline of the sub-topics that she planned to address in the body of the text. In addition, two skillful individuals managed to connect the background information they had provided earlier to the purpose of the current composition. Though not present in three or more of the sample posts, the additional information and the connection between that information and the writing task enriched the posts of the participants who utilized those functions. For this reason, both steps were added as optional components of the move classified as "summarizing key background information."

The first three moves identified in the academic blog posts were generally included in a single paragraph. In most cases, the second paragraph of the composition began by providing a concise answer to the first rhetorical cue. For example, the initial prompt on technology and

language learning included the rhetorical cue, “How is technology utilized today for language learning?” The author of Sample 1 stated explicitly, “technology can be used in several ways for language learning,” and the author of Sample 2 wrote, “a foreign language classroom cannot be imagined without any sort of technological component.” In five of six cases, these statements were followed directly by an example that supported the author’s answer. Although four of those examples were presented in general terms (i.e., “it is common to see overhead projectors” and “in public schools it is a subject that is part of the curriculum”), one participant utilized effectively a personal experience to support her stance. Interestingly, only one participant elaborated on the first example, with four of the six authors proceeding directly to the provision of a second example. Although the examples in this first paragraph were rarely justified (in contrast to participant responses to subsequent rhetorical cues), one author took the resourceful step of providing a hyperlink to an article discussing the interaction of Church and State in Spain. This unique, though effective, communicative function was added as an optional component of the move covering the answer to the first rhetorical cue.

The fifth, sixth, and seventh moves addressed the process of responding to the second, third, and fourth rhetorical cues, respectively. Even though the progression of steps within these three moves was not identical, the first step involved the use of a transition signaling a shift of sub-topic. In some cases, the transitions consisted of single words, such as “besides” or “nevertheless,” and in other cases, the transition between sub-topics constituted a complete clause, as in “but not only computers or overhead projectors are useful for the language classroom…” (Sample 1). Interestingly, the one author who elected not to provide transitions between subtopics was the author who presented a clear outline of the organization of her post in the introduction. Nevertheless, following the transition, all authors who addressed a particular

cue provided at least one example supporting their answer, with most participants supplying two examples and subsequently justifying or clarifying those illustrations. A finding of interest is that of the four rhetorical cues included in each prompt, only two participants made a point of responding to all four. In fact, the rhetorical moves analysis exposed the ways in which some authors dedicated considerably more textual space to certain questions without necessarily sacrificing the quality of the whole. The author of the first sample, for example, avoided the third rhetorical cue. In lieu of addressing that question, which asked respondents to consider situations in which technology might not be helpful for language learning, the participant provided several examples in support of his stance on the first, second, and fourth rhetorical cues. Whereas the other authors included a maximum of three examples in reference to a given question, this writer added six different examples of the ways in which technology could contribute to diverse aspects of language learning, including making learning more interesting, encouraging self-motivated learning, and facilitating writing. Thus, even though the rubric defines a top performance in the category of task fulfillment and relevancy as one that has addressed all questions from the prompt, a skilled writer who is extremely knowledgeable about a particular topic—as evidenced by his use of discipline-specific vocabulary such as "teacher-centered," "corpora," and "self-scaffold"—may not need to respond to every rhetorical cue to accomplish the task successfully.

The final move, evaluating information presented previously, revealed the greatest variety in the individual steps taken to accomplish this communicative function. Whereas five of the six authors included a statement summarizing their overarching stance on the main topic, there was a lot of diversity in the means by which they brought their compositions to a close. After summarizing their stance, four of the participants immediately restated their position on the topic, yet the similarities end there. One participant incorporated a famous quotation to

communicate her belief that religion was not a subject to be taught. Another author utilized a metaphor in which he compared technology to an automobile, and a different writer presented a qualification of her initial summary statement followed by a decisive final opinion that "homework assignments have more positive elements than negative ones." Whereas each of these methods was valid, as neither the prompts nor the rubric identified the components of an effective conclusion, two steps that elicited favorable reactions from the raters included encouraging reader comments via specific questions and calling for future research, a component of particularly advanced scholarly work.

In addition to identifying the fundamental steps through which performance within a particular genre is likely to be recognized, Jacobson et al. (2021) commented that "noticing language features helps you more closely analyze *how* certain moves are carried out and *to what effect*" (p. 226, emphasis original). Throughout the process of classifying and comparing the moves employed by these second language writers, a linguistic feature not mentioned in the rubric emerged as a salient characteristic across sample texts, and with greater frequency in the more highly-rated texts. This feature involved the effective navigation between use of personal pronouns such as "I," "you," and "we" and non-human subjects, including "technology," "religion," and "homework." For example, the author of Sample 1 wrote, "**apps** are easy to download and install in your mobile phone or tablet so **you** can quickly start using them" (emphasis added). The skillful alternation between non-human or generic subjects and personal pronouns imbued the electronic compositions with authority while simultaneously connecting the author to a desired readership. Even the author of a conservatively-scored post—Sample 6—concluded his piece by shifting from the general statement, "every student is different and everyone has diverse abilities and characteristics," to the more intimate comment, "well, I hope

you liked my insight on this topic and see you soon!" Based on the samples analyzed here, this vacillation between pronoun categories served, at least in part, to bridge the gap between purely academic genres (e.g., research articles and doctoral dissertations) and more informal written genres (e.g., messages exchanged between friends). It is also worth noting that in the instances in which an author utilized an emoji, the emoji appeared at the end of a paragraph, rather than between sentences within a single paragraph or even mid-sentence.

6.4 Summary

For the third and final phase of this project, three raters scored 93 academic blog posts according to a revised, four-level analytic rubric. A two-way repeated measures ANOVA showed that the presence of the rubric, time, and the interaction of the rubric and time had a significant positive impact on average scores. Participants' longitudinal written development was also analyzed via nine linguistic indices, and the mean values for noun-adjective and verb-direct object dependency bigrams demonstrated a significant change over time. As a supplement to the quantitative results, a qualitative rhetorical moves analysis supplied a sequence of optimal steps for constructing an academic blog post. In sum, the mixed-methods results of the longitudinal phase of this research provide a well-rounded representation of the impact of a genre-specific analytic rubric on the written development of 31 EFL learners within the academic blog, or discussion board post genre.

CHAPTER 7: DISCUSSION

This study has outlined a new method of rubric creation that combines actual student texts and the expertise of knowledgeable individuals (in this case, English language teachers with advanced education in Applied Linguistics). Through the application of a mixed-methods procedure that capitalized on quantitative and qualitative data, it was possible to create a genre-specific analytic rubric. Rather than asking potential reviewers, or raters, to come to an agreement on a set of binary questions, rater quality judgments and comments were used to construct a rating scale sensitive to the constraints of an online, academic task. With the initial rubric, six raters scored the posts of 163 English as a Foreign language learners, and a subsequent FACETS analysis revealed several areas in need of improvement. Thus, the number of performance levels was reduced from six to four, and various descriptors were eliminated, relocated, or revised. This revised rubric was used in the longitudinal portion of the research, which compared the written performance of two groups of learners, with only one of the groups allowed access to the rating scale during the composition of two more academic blog posts. The participants in the rubric group performed significantly better than a peer group of matched ability on the second and third instantiations of the task. In addition, an analysis of the linguistic development of learner texts and a rhetorical moves analysis contributed valuable information related to the participants' performance. In this chapter, the aforementioned findings are discussed at length, along with possible interpretations of the results. This discussion is followed by a consideration of the implications for research and practice.

7.1 Interpretation of Findings

The first research question set out to identify the features that should be included in analytical rating criteria used to assess written performance on an online, genre-based task. Two

sub-questions focused that analysis on aspects of student performance that were salient to expert reviewers and on the way in which qualitative data could support the inductive formation of rubric categories. The second research question was concerned with the extent to which evidence could support the use of the analytic rubric as a measure of second language writing performance. In order to answer this question, three sub-questions concentrated the analysis on relations among examinees, raters, and categories when placed on the same logit scale, rater bias, and scale functioning within each category. The third research question centered on comparing the longitudinal written performance of two groups of EFL university students: a treatment group that was given access to the rubric and a control group that repeated the online genre-based task without access to the rating scale. The fourth, and final, research question aimed to examine the ways in which a linguistic analysis of participants' longitudinal performance as well as a rhetorical moves analysis of their genre-specific texts could contribute to our understanding of the written development of both groups of participants and the nature of this emerging genre.

7.1.1 Features Included in Analytical Rating Criteria

7.1.1.1 Salient Aspects of Participant Performance. The first research question sought to identify characteristics of student performance on an online writing task, which proved salient to a group of expert reviewers. A factor analysis conducted on the reviewers' sorting behavior revealed that there were at least two groups of reviewers, with Reviewer 1 and Reviewer 6 making dramatically distinct assessments of student texts. Although Reviewer 3 made quality decisions more in line with the preferences of Reviewer 6, the folder assignment patterns of Reviewers 2, 4, and 5 retained characteristics of both groups. In other words, the results of this study corroborate decades of research in writing assessment, which have acknowledged that even expert assessors may deviate in their evaluations of learner performance on a given writing task.

Absent a clear rubric, the reviewers made assessments of student texts based on individual determinations of post quality, which may or may or may not have aligned with fellow reviewers' perceptions of those same texts. Cognizant that even experienced raters will view participant writing in diverse ways, the novelty of this study rests in demonstrating that rater disagreement may lead to the creation of a more robust, inclusive rating scale. Particularly as instructors and researchers struggle to describe features of emerging genres like blogs and discussion boards, a data-based mixed-methods approach to rubric creation may ameliorate certain challenges imposed by alternative rubric-creation techniques.

Perhaps surprisingly, given the evaluative freedom granted each reviewer, the average folder scores for 148 student texts resulted in a normal distribution. This normal distribution allowed the identification of texts representative of each of the score bands formed by moving one, two, or three standard deviations above and below the mean. Essentially, the normality of the data eliminated the need for reviewers to come to an agreement on judgments of particular posts, and in turn, the rubric creation process valued the expertise and diversity of individual language teaching professionals. Although at times the reviewers were drawn to different attributes of the 148 sample texts, the organization and consolidation of their comments led to the identification of five primary categories of interest: task fulfillment and relevancy, content, organization and balance, genre specific features, and language use.

7.1.1.2 The Inductive Formation of Level Descriptors. Approaching the formation of rubric categories via an inductive approach led to a rating scale free of vague or generic language, such as "adequately developed" and "well organized." In contrast, each category contains precise descriptors that connect the aims of the online task with the textual attributes observed in student performance at distinct levels. Deviating from frequently encountered

composition rubrics, this rating scale makes a distinction between fulfillment of the writing task and quality of content. Scrutiny of reviewer comments revealed that performance in one category did not necessarily equate with achievement in the other. For example, a participant may have included several thoughtful ideas that were supported with a variety of detailed examples. At the same time, such engaging content may not have addressed the questions raised in the writing prompt. Whereas many popular rubrics consider topic coverage and content to be one in the same, the qualitative descriptions provided by the reviewers in this study pointed to a fine-grained distinction between the two categories.

Another novelty of this blog-specific analytic rubric rests in the identification of several features unique to this digital genre. The consolidation of reviewer comments pointed to a preference for pieces in which the author acknowledged a particular group of readers and engaged them via well-placed questions. Student authors with greater control of the academic blog genre exploited the interactive nature of online writing by including hyperlinks to images or to source texts containing additional information on the topic. Various reviewers also commented positively on participant inclusion of emojis and the occasional use of personal experience as support for a central idea. In grouping reviewer comments by average score level, the presence or absence of these genre-specific features rose to the forefront. In turn, such details became the basis of rubric descriptors, thereby capturing the unique voice of each reviewer.

7.1.2 Evidence Supporting Rubric's Use as a Measure of Writing Performance

7.1.2.1 Relationships Among Examinees, Raters, and Categories. Participant responses to the first prompt and raters' rubric-based scores of those same texts enabled the calculation of logit values associated with dependable estimates of examinee ability, category difficulty, and rater severity. The wide distribution of participant abilities on the online genre-

based task was supported by a high degree of reliability and sound fit statistics. In addition, the category measurement report confirmed that the five categories (task fulfillment and relevance, content, organization and balance, genre-specific features, and language use) posed distinct levels of difficulty. As a result, each of the categories contributed differing degrees of variance to distinguishing among participant ability levels. The strong fit statistics reported for the categories confirmed, however, that each of the categories worked toward the measurement of a single latent trait, substantiating the suitability of the one-parameter Rasch model. If the fit statistics had not been as strong, consideration of a two-parameter logistic model would have served as a logical next step, and future research may wish to consider use of a such a model or of a generalized partial credit model.

7.1.2.2 Rater-Category Interactions. Seeing as the first phase of the project involved the creation of a new, genre-based analytic rubric, the raters encountered the initial rubric for the first time upon opening the materials that had been sent to them. In other words, in contrast to studies that have employed widely used rubrics, such as the TOEFL holistic rubric for independent writing or Jacobs et al.'s (1981) ESL composition profile, the raters employed for this project, though experts in their field, did not have previous experience working with the criteria included on this rating scale. Nevertheless, none of the raters proved to be problematic. These knowledgeable individuals were able to interpret the rubric descriptors and to apply the scale levels with a reasonable degree of consistency.

Of course, inter-rater reliability does not necessarily mean that the raters were free of bias. The bias report presented in section 5.3.2.2 showed how different raters were biased for or against distinct categories. For example, the harshest rater overall was actually the most lenient rater in judging the portion of participant performance connected to task fulfillment and

relevancy. In contrast, one of the two most lenient raters scored examinees particularly harshly on language use. Of the five categories, raters were most consistent in assigning scores for content, whether as a result of the clarity of the content-related descriptors or a shared understanding of distinct levels of quality within this category. In spite of rater differences in severity, the stable fit statistics showed that the raters were internally consistent and that none of the raters deviated significantly from model expectations. As a result, the information provided by the raters in this phase of the project proved useful for investigating longitudinal differences in the performance of EFL writers on the online genre-based task.

A noteworthy finding of the rater measurement report concerns the respective leniency and severity of the native English-speaking and non-native English-speaking raters. Although the research did not set out to examine differences in the behavior of native- and non-native speaking raters, the FACETS analysis necessitated a closer look at rater behavior along these lines. Whereas Youn (2013) and Patharakorn (2018) determined that the first language of the raters did not affect their internal consistency or overall severity, the non-native speaking raters in this study were positioned on either end of the severity-leniency continuum. The harshest rater was a non-native speaker of English, as were the two most lenient raters. This finding is also in contrast to the results presented in Kachchaf and Solano-Flores (2012), who found a statistically significant difference in the scores assigned to students' Spanish and English open-ended responses by raters who were native speakers of English and raters who were native speakers of Spanish. The native English-speaking raters consistently scored the Spanish-English bilingual students' written products more highly than their L2 English peers.

Although the student participants in Kachchaf and Solano-Flores (2012) were much younger than the university-aged participants in this study, it may be worthwhile to examine the

ways in which the context of rating (i.e., either English as a second language or English as a foreign language) influences the relative leniency and severity of L1 and L2 raters. It could be that the very strict rater in this study was an anomaly and that L2 English raters in an English as a foreign language context tend to be more lenient than L2 raters in an English as a second language context. At the same time, it would be premature to suggest context of instruction as a definitive explanation for the observed differences, seeing as Kachchaf and Solano-Flores utilized eight raters (4 NS and 4 NNS) and this study employed only six raters (3 NS and 3 NNS) in the rubric revision stage. The results of the bias analysis among raters and categories in this study do contradict a suggestion forwarded by Kachchaf and Solano-Flores in their concluding remarks, namely that "bilingual teachers may be more likely to value functional communicative aspects of language whereas teachers of English as a foreign language may be more likely to value formal aspects of language such as spelling and syntax" (p. 174). The bias analysis in this dissertation revealed that in terms of language use, the non-native English-speaking EFL instructors were again positioned on either end of the leniency-severity continuum. The two harshest raters on participants' use of language were non-native speakers, but so too was the most lenient rater in this category.

7.1.2.3 Functioning of Rating Scale Levels. The examinee measurement report that accompanied the first FACETS analysis indicated that participant performance could be separated reliably into four statistically distinguishable levels, as opposed to the six levels identified on the original rubric. Similar to results presented in Janssen et al. (2015), category probability curves for this rubric showed that there were likely a few redundant score levels in each category. An examination of these curves provided the clearest picture of instances of overlap between adjacent score levels. For example, in the case of language use, the Level 4

category probability curve was entirely subsumed by the curves corresponding to Levels 3 and 5, thereby rendering that fourth level as superfluous. As a result, the initial rating scale was condensed from six levels per category to four levels per category, thereby presenting a more meaningful progression of performance indicators from one level to the next. It is possible that a three-step rubric could have worked even better for the participants in the longitudinal phase of the project, as they represented a smaller range of ability levels, but for pedagogical purposes, the number of levels was left at four. In other words, even though descriptors corresponding to the lowest level of performance were not often observed in the written products of the longitudinal participants, it was important to acknowledge that it was possible for a text to lack basic organizational principles or to be devoid of any genre-specific features. This lowest level could be applicable to the performance of individuals in a different context, or more specifically to a group of learners who had not volunteered to compose an academic blog post in English. For other programs or bigger contexts in which the variety of performance levels is much larger, maybe the five- or six-level rubric, common to large-scale examinations (Knoch, 2011), would provide a better illustration of the distribution of student ability levels.

Janssen et al. (2015) concluded their research by suggesting that future studies focus on the revision of individual category descriptors, which became a key focus of revisions to the six-level analytic rubric created in Phase I in this study. Harsch and Martin (2012) identified four ways in which the wording of individual descriptors listed in the Common European Framework of Reference needed to be adapted to their local context. These areas included subsuming or splitting the original descriptors, tackling inconsistencies in verbiage, removing statements of probability or possibility, and adding more specific rater instructions. Similarly, to reduce the number of levels in the rating scale, various rubric descriptors were eliminated, redistributed, or

revised. For example, within the task fulfillment and relevancy category, a descriptor originally aligned with a Level 4 performance—*may* contain some irrelevant points and information—was eliminated due to its inclusion of a marker of probability. Descriptors expressing possibilities were challenging to interpret, and their subsequent elimination or revision improved the clarity and distinctiveness of individual rubric cells. As in Harsch and Martin, inconsistencies in verbiage were also addressed, particularly in the case of the category describing various genre-specific features. Rather than retaining a mix of descriptors, which highlighted different genre-related characteristics depending on the performance level, in the revised rubric, each of the four cells corresponding to genre-specific features stated, "makes use of…" none, one, two, or three or more, respectively "…of the following: acknowledgement of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience/opinion." Certainly, depending on the topic or the unique requirements of a particular academic blog post, including several genre-specific features in one text could be seen as incongruous. Although participant responses to the task employed in this research were enriched by such additions, the appropriateness and desirability of various genre-specific features ought to be determined by the local context of use.

7.1.3 Differences in Longitudinal Performance Between Rubric and Non-rubric Group

The results of the two-way repeated measures analysis of variance indicated that repetition of the online genre-based task had a significant positive impact on their written performance over time. In addition to the significant effect of time, learners who had been provided with a detailed analytic rubric performed significantly better on the online written task as opposed to learners who had not been supplied with the rubric. In spite of the modest sample size, the statistical power was high, suggesting that the improvements in participants' written texts were sufficiently straightforward, so as not to warrant the recruitment of additional second

language writers. Andrade et al. (2010) confirmed that despite the intuitive appeal of rubrics, research supporting the claim is scant. In fact, an extensive search for empirical literature from outside the field of Applied Linguistics revealed only two articles with evidence in favor of student-utilization of analytic rubrics. In a study with first language (L1) Turkish speakers enrolled in a Turkish university's science education program, Kocakülah (2010) found a very large difference in the scores students received on a task involving Newton's Laws of Motion. Whereas the control group and the experimental group performed similarly on the task prior to instruction, from a mean of 23.25 for the experimental group and a mean of 23.72 for the control group, after the analytic rubric 'treatment,' mean scores of participants in the experimental group rose to 84.95, compared to 33.86 for the control group. Similar to the results found in this dissertation, even though the experimental group's average ability on the task was slightly lower than the aptitude of the control group at the first data collection point, use of the genre-specific rubric enabled the experimental group to perform significantly better than their peers on successive iterations of the task.

In another study, this time with first language English speakers at middle schools in the United States, Andrade et al. (2010) found a statistically significant positive effect on student performance resulting from the combined use of model essays, student involvement in rubric creation, and the subsequent application of the rubric during writing. In spite of the significance of Andrade et al.'s findings, Brookhart and Chen (2015) and Jonsson (2014) warned of the potential of conflating the impact of an analytic rubric when the rubric is utilized in combination with other instructional or assessment tools. Seeing as no additional interventions were included in this dissertation, the results demonstrate clearly that providing students with a genre-specific rubric correlates with significantly higher scores on the final written product. Furthermore, to my

knowledge, this dissertation is the first study within the field of Applied Linguistics to provide an empirical justification for the use of well-designed rubrics in the teaching of second language writing. Specifically, the presence of the rubric accounted for 36% of the between-participants variance in learners' total scores. The positive consequences of applying the analytic rubric in this study surely outweighed any potential negative consequences, and in turn, these findings contribute an important piece of validity evidence in support of the consequences of assessment (Lane & Stone, 2002). In sum, both the repetition of a genre-based task and the use of a well-designed rubric contributed to a statistically significant positive difference in the assessed written products of EFL learners.

7.1.4 Contributions of Linguistic Analysis and Rhetorical Moves Analysis

7.1.4.1 Linguistic Analysis of Longitudinal Written Development. In this study, nine linguistic variables measuring various aspects of lexical diversity, lexical sophistication, syntactic complexity, and syntactic sophistication were calculated for participants' longitudinal texts. The measure for lexical diversity, moving-average type-token ratio (MATTR), stayed relatively stable over time, as did two traditional measures for lexical sophistication, lexical decision time (LDT) and content word frequency (CWF). Although the indices related to contextual distinctiveness, McD and USF, were not analyzed in terms of significance (McD due to a lack of any noticeable change and USF for a violation of normality), the developmental trajectory of each variable proceeded in the opposite direction of the results reported in Kyle and Eguchi (2021). Whereas the expectation would be for McD values to increase and USF values to decrease as participants incorporated words that were more contextually distinct, learners in this study steadily inserted words with lower levels of contextual distinctiveness. The final two variables covering lexical sophistication, noun-adjective two-word sequences and verb-direct

object collocations, did demonstrate significant changes over time. The positive direction of the change in noun-adjective dependency bigrams aligned with prior literature, which has shown that more proficient writers use two-word sequences that are more closely associated with each other (Kyle & Eguchi, 2021). In contrast, the index for verb-direct object dependency bigrams decreased significantly over time. In other words, whereas the increase in the mean scores for noun-adjective bigrams could suggest participant utilization of more strongly related two-word sequences, the decrease in average scores for verb-direct object dependency bigrams could indicate participant utilization of more fossilized language or greater experimentation with English. Further analyses of individual participants' development in their use of these two types of dependency bigrams might clarify the contradiction in the quantitative findings.

In terms of the syntactic indices, values for DC/C, the selected measure of syntactic complexity, and MVF, the measure utilized for syntactic sophistication, demonstrated little variation over time. One explanation for this lack of syntactic development in participant texts is found in Ortega's (2012) discussion of the dynamic-synoptic style continuum arising from studies in Systemic Functional Linguistics. Within this theory, increases in subordination are noticeable in early stages of language development, but their utility declines at more advanced proficiency levels in favor of nominalization and grammatical metaphor. Another explanation for the absence of meaningful changes in participants' syntactic complexity and sophistication could be the high level of correspondence among the three prompts, both in terms of the rhetorical cues and the prompt domain. In several studies that have reported significant changes in syntactic complexity and syntactic sophistication, it is not always clear that the prompts to which participants responded were equivalent. For example, in Bulté and Housen's (2014) research on short-term changes in second language writing complexity, one prompt was used at the

beginning of the time period and a different prompt was administered at the end of the course. In Lu's (2011) cross-sectional study, which set out to identify syntactic complexity measures that could distinguish reliably among school levels, the researcher acknowledged that although students in the same level at the same school received the same prompt, the nature of the writing prompt differed by level and across institutions. In a final example, the essays examined in Kyle et al. (2021) were written in response to a series of topics that started with "write a short story about your new school, friends, and teachers" at Time 1 and ended with "pretend your school principal has stated that from now on anyone should wear a school uniform [and] write him/her a short letter to explain why you agree/do not agree with this new rule" at Time 6. While it is certainly possible that the prompts used in the abovementioned research elicited comparable texts, it would also be worthwhile to consider whether or not dissimilar tasks can lead to variation in written production.

In addition to examining the written development of participants' academic blog posts, the nine linguistic indices enabled an investigation of the degree of correspondence between rater scores for language use and measures of lexical diversity, lexical sophistication, syntactic complexity, and syntactic sophistication. From the review of a correlation matrix containing the nine linguistic variables, participant scores on the Oxford Online Placement Test (OOPT), and rater scores for language use, it became apparent that only three variables correlated significantly with scores assigned for language use: lexical decision time (LDT), the moving-average type-token ratio (MATTR), and OOPT scores. A sequential regression analysis demonstrated that 16% of the variability in language use scores could be predicted by MATTR, an index of lexical diversity, LDT, an index of lexical sophistication, and scores on the OOPT. Interestingly, even though values for noun-adjective and verb-direct object dependency bigrams changed

significantly over time, neither index contributed to the model predicting scores for language use. The same result was reported by Bulté and Housen (2014) who attributed the finding to a possible halo effect, in which participant performance in one area, whether poor or excellent, influences rater scoring of other categories. It could also be that linguistic features besides the indices highlighted here affected raters' assignment of scores for language use. Jarvis (2013) raised an important question concerning whether or not human raters should receive training in the components of lexical diversity, for example, before they are asked to assign scores. Although rater training will be addressed in Section 7.3, Jarvis' (2013) call for further research into the ability of various linguistic indices to predict judgments of human raters is a worthwhile avenue for gaining insight into the discrepancy between scores for language use and measures of lexical sophistication, syntactic complexity, and syntactic sophistication in this study. Finally, even though a period of nearly two years passed between Time 1 and Time 3, additional points of data collection could have uncovered a more fine-grained appreciation of learners' written development over time.

7.1.4.2 Rhetorical Moves Analysis of the Academic Blog Post Genre. The rhetorical moves analysis revealed several interesting characteristics of the texts produced in reference to the online genre-based task. Three of these revelations include the complexity of introductory and concluding paragraphs, the connection between the prompt's rhetorical cues and the written products, and the identification of points at which writer completion of a particular move was unclear. Participants in both groups had had some familiarity with standard essay writing in English and members of the rubric group had been alerted to the importance of an introduction in the organization of their composition. Nevertheless, the moves analysis uncovered a complex series of steps involved in establishing the importance of the topic and supplying background

information before responding directly to rhetorical cues included in the prompt. In other words, rather than begin a post by directly answering the first question, participants introduced their readers to the topic under consideration and then used general information about the topic or personal experience to construct a shared knowledge base from which to begin deliberation of the required rhetorical cues. The variety of content included in participants' introductory paragraphs suggested that a lot more goes into preparing a response to a writing prompt than merely supplying an answer. For example, whereas a question posed verbally in an in-person or synchronous online course could be answered with a raise of the hand and a direct answer, responding to these rhetorical cues required preparation of a shared knowledge base. Whether this finding is unique to the academic blog post or is also found in student responses to open-ended questions on exams is left to be examined.

The second point of interest is the one-for-one connection between the order in which the rhetorical cues were presented in the prompt and the sequence of the moves in the final model of the academic blog post. Even though individual students chose not to, or simply forgot to respond to one of the rhetorical cues, the order in which they answered them did not deviate from the order in which they were presented. This connection between the prompt and the response was the focal point of Chapman's (2016) doctoral dissertation, and the importance of the prompt on student output has been observed in other experimental research. Finally, it is important to note that the process of identifying the individual step corresponding to a particular text fragment was not always a straight-forward affair. For example, actions such as clarifying a previous preposition and summarizing the author's stance on the main topic were fairly easy to recognize. On the other hand, the point at which a particular writer presented a direct response to one of the prompt's rhetorical cues was not always clear. Interestingly, the rubric did not specify that direct

answers to the prompt's rhetorical cues needed to be identified clearly, but future research could address how the clarity of such an action might affect a rater's perception of task completion.

7.2 Implications for Theory and Research

7.2.1 Transparency in the Communication of Scale Development Procedures

Several researchers have reported on the scarcity of literature documenting the manner by which rating scales are constructed (Al-Hoorie & Vitta, 2019; Fulcher, 2003; Knoch, 2011; Turner, 2000). Fulcher (2003) even identified the use of intuition by the "educated native speaker" as a common practice of test developers and language teachers. Much preferred to the intuitive method is scale development on the basis of performance data. This study has carefully documented the process behind the construction of an analytic rubric for an academic blog post, and it has also presented a case in favor of developing task-specific rubrics on the basis of actual samples of the genre in question. These sample texts, combined with the judgment of expert raters, served as the backbone for developing a rubric that accurately reflected distinct levels of participant performance. In addition to demonstrating the utility of approaching writing assessment from a genre-based perspective, this research presented a careful examination of validity evidence in support of the rubric's use with the population of learners included in the study. Future scientific investigation in the area of second language writing assessment ought to present at least two types of validity evidence, whether evidence based on test content, test consequences, internal structure, response processes and/or relations to other variables. Panadero and Jonsson (2020) maintained that a failure to provide validity information concerning the use of a rubric would obfuscate claims made in relation to the effects of that assessment instrument. The rating scale developed in this research has presented validity evidence based on test content, test consequences, internal structure, and relations to other variables, and future genre-based

assessment projects may wish to adopt similar procedures in the moment of constructing a new rubric and of communicating the results of its implementation.

7.2.2 Benefits Arising from the Use of Mixed Methods Research

Not only is it important to document the procedures involved in rubric construction and to present validity evidence in support of the interpretation of rubric scores, but also the utilization of Mixed Methods Research contributes to the legitimacy of the findings. The exploratory, sequential mixed-methods design (Creswell & Creswell, 2018) adopted for this project, which necessitated the purposeful integration of qualitative and quantitative modes of inquiry, added to the quality of the rubric descriptors and enabled a more thorough investigation of the effect of the rubric on second language writing development. The quantitative data provided by the raters in the second and third phases of the research allowed for an examination of the internal structure of the rubric and for the application of a two-way, repeated measures analysis of variance, which reported a statistically significant difference in the performance of participants in the rubric group versus the non-rubric group. Qualitative data supplied by the reviewers in the first phase of the project enabled the identification of clear rubric descriptors, and participant texts, when treated as qualitative data, allowed for an analysis of rhetorical moves. This analysis uncovered a series of steps contributing to the enactment of the academic blog post genre. Both methods of inquiry and analysis worked to minimize the weaknesses inherent in the other source. In turn, the mixed methods design provided a stronger argument in favor of the use of this genre-specific analytic rubric to support longitudinal development in second language writing.

7.2.3 Connections Between Prompt and Product

The results of this study have highlighted the connection between the writing prompt and the written product. Specifically, the model of text-based actions arising from the analysis of participant rhetorical moves corresponded directly to the order in which the four rhetorical cues were presented in the writing prompts. The writing prompt is likely to affect not only the structure of a student's text, but also the diversity of a learner's vocabulary. Zenker and Kyle (2021) uncovered a small effect of the essay prompt on values corresponding to lexical diversity. Specifically, a prompt asking participants to respond to a statement about banning smoking indoors inspired the use of a greater range of vocabulary than a prompt containing a statement about the importance of part-time jobs whilst attending university. The reason for which a prompt about smoking, as opposed to one about part-time jobs. would lead to greater lexical diversity in learner texts may not be immediately obvious; however, the connection between prompt structure and written output ought to be a consideration in the early stages of research design. The topic, the order of presentation of rhetorical cues, and the expected audience of the composition can all lead to variations in written performance. Furthermore, it may result that while intending to investigate the nature of emerging digital genres, both the researcher and the participants recur to more familiar, albeit less genuine, formulas that resemble the five-paragraph essay. Indeed, the structure of steps uncovered in the rhetorical moves analysis shared a number of features with the familiar 'essay' format. It could be that both the rhetorical cues and the learners' recurrence to familiar formulas for academic writing influenced the nature of the digital texts investigated in this research, and it would be worthwhile to explore additional avenues for examining emerging written genres.

7.2.4 Implications for Rater Training (or not)

Training raters prior to the scoring of written products is a familiar practice in the field of Applied Linguistics. In fact, most studies examining the process by which raters make scoring decisions conclude with recommendations for continued rater training surrounding the application of rubric criteria to individual writing samples. Kachchaf and Solano-Flores (2012) are not the only researchers to have attributed rater consistency to rater training sessions despite the absence of a control group, yet the raters in this study received zero training apart from an instruction sheet. Notwithstanding the lack of rater training, the raters in Phase II and in Phase III all achieved acceptable levels of inter-rater reliability. Although rater reliability is an important consideration in the assessment of written performance tasks, Harsch & Martin (2012) have argued that focusing on the quality of level descriptors would be a more efficient way to improve rater consistency. In other words, the results of the present study demonstrate that rather than adhering to a paradigm in which raters assume responsibility for interpreting rubric descriptors, it would be worthwhile to consider the impact of directing resources spent on rater training to the creation of more effective rubrics. It is important that raters are able to interpret the rating scale descriptors as intended, thus the rationale in this study for using the reviewers' language as the basis of rubric construction. For language programs with financial constraints, emphasizing the development of clear descriptors may be a reasonable alternative to extensive rater training.

7.3 Implications for Practice

7.3.1 Utilization of Analytic Rubrics

This research project has demonstrated that a genre-based analytic rubric can have a positive impact on students' written products. As a result, if an instructor's goal is to assist language learners in becoming more proficient writers, providing them with well-designed

analytic rubrics will help them to realize that objective. Panadero and Jonsson (2020) explained that "rubrics are probably the most common way of sharing explicit assessment criteria with students" (p. 1), and by doing so, teachers can help students approach an unfamiliar genre with increased confidence. Especially considering that "there is almost always a need for students to comply with the teacher's expectations" (Panadero & Jonsson, 2020, p. 9), the use of a genre-based analytic rubric can increase student awareness of the criteria according to which their texts will be assessed. At the same time, the transparency inherent in the communication of explicit assessment criteria need not restrict the display of a writer's individuality. For example, the rhetorical moves analysis conducted for this study identified the ways in which students were able to express their creativity as well as the points at which it was appropriate to insert original ideas or content. Though time-consuming, language assessment researchers and writing instructors can follow the procedures outlined in this research to develop context-sensitive rubrics that accurately depict a range of student performance levels within new digital genres.

7.3.2 Student Training in Using Rubrics

There is a wide body of literature concerning rater training, but noticeably less on training students to interpret rubric descriptors and to apply rubric criteria to their writing. Genre-based analytic rubrics present students with the opportunity to reflect on their written performance; however, merely supplying learners with a rating scale may not be the decisive factor (Panadero & Jonsson, 2020). In the present study, the group of learners who had access to the rubric also had the opportunity to review the rating scale criteria and to ask questions about descriptors that were not entirely clear. Furthermore, these learners were advanced users of English at the university level, and as Panadero and Jonsson (2020) reported, in studies with university-level participants, "students have managed to use rubrics to improve their

performance and/or self-regulate without any substantial training" (p.1). On the other hand, if the rubric created in this research were to be introduced in an elementary school or used with learners of lower English proficiency, additional student training in applying the rubric to their written work would be warranted. Then, as students become familiar with the rubric criteria, their awareness of the connection between the rubric descriptors and the manifestation of those criteria in their written work may support further language development. Nevertheless, even if learners fall short of complete comprehension of rubric descriptors, at the very least, the explicit provision of those criteria can open a discussion between the language teacher and the learners, thereby initiating learners' entrance into their desired community of practice (Panadero & Jonsson, 2020). Certainly, a clear analytic rubric will make the interpretation process easier, and in turn, augment learners' explicit genre awareness.

7.3.3 Task and Prompt Design

In the review of literature, five characteristics of effective genre-based tasks were identified, mainly that tasks should be explicit, genuine, recurrent, social, and varied. The results of this study have highlighted the value of explicitly communicating the expectations for a particular task and of providing opportunities for task repetition. Two additional points to keep in mind at the moment of designing a technology-mediated task are the affordances of the electronic medium and the role of the teacher in supporting students' appropriation of those resources. For example, although the EFL instructor in Vurdien's (2013) research sought to engage students in academic blogging, the eight tasks were mirrored on tasks appearing in the written portion of the Cambridge CAE exam, including an essay, a report, and a review. Several participants in the study reported "that the exam-oriented nature of the tasks was not stimulating" (Vurdien, 2013, p. 135), and the author recommended that future studies consider carefully the

type of task employed, so as to maximize the capabilities of the digital environment. Although the blog platform utilized in this study did not support the embedding of multimodal resources, such as videos or audio files, it would be worthwhile to explore other platforms and to examine the ways in which such resources may enhance learners' written texts. It is also important to reconsider the definition of "expert" in the ambit of an emerging genre such as the academic blog post. In fact, as instructors and language learners create within the medium, they participate in defining the characteristics of that genre. Thus, by setting aside previously held conceptions surrounding the existence of a genre expert, teachers and students can investigate together the effects of different technology-mediated tasks on language learners' written products. The heightened genre awareness resulting from such a project could facilitate revision of the initial rhetorical cues and experimentation within the genre. Whichever pedagogical strategy is selected, the results of this research suggest that repeated engagement with an online, genre-based task can lead to significant gains in written performance over time.

7.4 Closing Remarks

Harsch and Martin (2012) explained that "in general, once a rating scale has been drafted, it has to be trialed for its new context and purpose as part of the scale validation process" (p. 232). In line with their recommendation, it is worth stating that the revised rubric in this study is not meant to serve as a generic or standardized rating scale for all second language contexts. This genre-specific analytic rubric would likely need to be adapted in accordance with the needs of individual classrooms. In addition, future researchers may wish to pilot translated versions of the rubric for use with languages other than English. The collection of validity evidence involves "ongoing, iterative work" (Janssen et al., 2015, p. 65), and that work is necessary if we wish to make meaningful interpretations of rubric-based scores in distinct contexts. A one-size-fits-all

approach to rubric design is not always the best answer to local assessment concerns, but knowledgeable practitioners can follow the guidelines outlined in this project to create appropriate second language writing assessments focused on genres of import to the specific learner population.

CHAPTER 8: CONCLUSION

This research project set out to address various gaps in Applied Linguistics literature, including the outsized representation of the oral mode in studies on task-based language teaching, the need for further analyses of digital texts, which have become increasingly prevalent in academic and non-academic contexts, and the absence of studies on the effects of utilizing genre-based analytic rubrics on student written performance. In responding to these gaps, this study selected an online, genre-based writing task as its focus, and it documented the process by which a genre-specific rubric was created to fit the technology-mediated task. Finally, the experimental, mixed-methods design used to collect longitudinal data demonstrated that a genre-based rubric had a significant positive effect on participants' written performance. In this chapter, the reader will encounter a summary of the project sequence and corresponding results, followed by a discussion of the study's limitations and suggestions for future research. Ultimately, I would like to judge the success of this project in terms of the degree to which it inspires continued research in emerging genres, digital tasks, and the formative assessment of language learner written production in those tasks.

8.1 Project Summary

A literature-based needs analysis of the written work assigned to university-level learners in the United States and Spain led to the identification of the educational blog or discussion board post as a genuine, written digital task. Further research into the effect of written prompts on language learner production resulted in the creation of three, largely equivalent prompts containing four rhetorical cues. For the first phase of the project, 148 participants recruited at the University of Murcia composed academic blog posts in response to a prompt on language learning and technology. Next, six experts examined the 148 academic blog posts, separating

them into six levels according to merit. The learners' performance on the first genre-based task served as the basis from which to construct an analytic rubric describing actual performance levels. The sequential mixed-methods approach to rubric creation led to the identification of five primary features around which the reviewers assessed the learners' written performance: task fulfillment and relevancy, content, organization and balance, genre specific features, and language use. The reviewers' qualitative comments became the foundation of the detailed descriptors used to populate the cells formed by the intersections of rubric categories and score levels. After the rubric's formation, 15 additional participants wrote an academic blog post in response to the first prompt, bringing the total number of texts to 163. Next, these 163 texts were scored by six raters according to the initial six-level rubric. The resulting fully-crossed design allowed for the examination of validity evidence based on internal structure. Specifically, a many-facets Rasch analysis enabled an examination of participant ability levels, rater severity, and category difficulty. The fit indices provided by FACETS indicated that the one-parameter Rasch model was appropriate for this study's data and that the five rubric categories were functioning independently of one another. Finally, a detailed inspection of the category probability curves as well as the wording used for individual rubric descriptors substantiated the reduction of scale levels from six to four.

For the third phase of the project, 31 of the original 163 participants wrote two additional academic blog posts separated by the span of approximately one year. At the end of the data collection period, three raters who were unaware of the order in which the posts were written scored the 93 texts (three per participant) according to the revised, four-level rating scale. Average total scores were calculated for each post, and within- and between-group means were examined via a two-way repeated measures analysis of variance (ANOVA). The results of the

mixed between-within participants ANOVA showed that the learners' average score was significantly affected by the presence of the rubric, by time, and by the interaction of time and the rubric. In other words, not only did repetition of the online genre-based task have a significant positive effect on participant performance, but also the application of the analytic rubric contributed significantly to the superior performance of the individuals who had been assigned to the rubric group. To supplement the statistical analysis, the longitudinal written development of participants' academic blog posts was analyzed via nine linguistic indices covering lexical diversity, lexical sophistication, syntactic complexity, and syntactic sophistication. The mean values for two linguistic variables, noun-adjective and verb-direct object dependency bigrams, demonstrated a significant change over time; however, the direction of the change in participant use of verb-direct object dependency bigrams did not correspond to results encountered in previous literature. Furthermore, two linguistic indices that showed no significant change over time, the moving-average type-token ratio (MATTR) and lexical decision time (LDT), ended up forming part of a regression model that could predict 16% of the variance in participant scores. Finally, a rhetorical moves analysis of six participant texts revealed a clear sequence of obligatory, common, and optional steps for constructing an academic blog post. The information provided by these supplemental analyses could contribute to further revisions of the four-level analytic rubric.

8.2 Limitations

Despite the rigorous nature of this study's research design, it is important to acknowledge certain limitations of the project. The three limitations addressed in this section include adherence to an equally divided, non-weighted grid in both the initial and the revised rubrics, challenges presented by volunteer participants, and requirements related to the analysis of

linguistic variables. In reference to the rating scale, a potential limitation rests in having restricted its format to a five-by-six grid. Humphry and Heldsinger (2014) and Knoch (2011) have alluded to the possibility that categories included in an analytic rubric may require distinct divisions of performance levels. In other words, whereas the revised rating scale distinguished among four levels of performance for each category, it would be worthwhile to consider whether or not different categories would perform better with three levels, or even five or six levels. Humphry and Heldsinger (2014) explained that partitioning rubric categories in the same manner is largely due to convenience rather than "because equal numbers of gradations faithfully capture the distinguishable performance levels for separate criteria" (p. 256). Thus, even though the rating scales constructed for this research conform to the traditional matrix design, future investigations that incorporate the rubric may want to consider collecting validity evidence either to support or to contradict the inclusion of an equal number of performance levels for each category.

A second limitation concerning the construction of the rating scale is that the five categories were not weighted. Although the categories surfaced from an examination of reviewer comments, it is possible that the reviewers weighted certain characteristics more heavily when assigning posts to one of the numbered folders. One method for uncovering such weighting would be to interview the reviewers either during or after they have completed the task to identify the attributes that most impacted their folder assignments. It would also be feasible to calculate the number of words the reviewers used in comments referencing each of the five categories. However, seeing as a primary goal of this research was to identify features of student performance that could be included on an analytic rubric used to assess academic blog posts, the categories were not weighted. Ultimately, the context in which the rating scale is to be employed,

including the nature of the assignment and the instructor's goals for the course, should guide the selection of individual criteria and decisions surrounding the weighting of different categories.

In reference to the longitudinal portion of the study, an anticipated limitation was the modest number of second language learners who returned to write a second and a third post. Nevertheless, the power reported in the two-way repeated measures ANOVA indicated that even with 31 participants, the significant between-group differences and the significant increase in average scores over time were exceedingly clear. The participant-related variable arising as a potential limitation rests in the fact that the longitudinal participants were volunteers. The average ability level of the longitudinal participants was higher than the average ability reported for participants in the first phase. In addition, the ability levels of the longitudinal participants spanned a relatively smaller range of logit values. Seeing as syntactic development in texts produced by advanced language learners tends to progress more gradually than with learners of lower proficiency levels, the higher language ability of the longitudinal participants could have impacted the observed change or lack of change in the mean values for the nine linguistic variables examined in this study. Furthermore, it would be worthwhile to investigate whether or not the proficiency level of participants may affect their ability to apply the information contained in an analytic rubric.

The final limitation is not related to the results obtained through this research, but rather a commentary on the complexity of conducting linguistic analyses of learner texts. Cobb and Horst (2015) commented that a practitioner or action researcher may start a project "encouraged by the apparent 'doability' of corpus research, only to watch this disappear into mathematical complexity" (p. 205). Even though I began the linguistic analysis of participants' academic blog posts with an honest appreciation of the intricacy of the process, I faced a steep learning curve in

interpreting the results. Whereas the linguistic tools created by Kristopher Kyle and colleagues have made the calculation of indices of lexical and syntactic sophistication quite expedient, the process of attaching meaning to the numerical values these tools provide required a broadening of my mental faculties. Not only did the linguistic analysis presented in this research necessitate the acquisition of a wide range of vocabulary, including "bigram," "Zipfian," and "lemma," but also it required the application of rather advanced mathematical knowledge to understand the nuances of different calculation methods, such as "mutual information" and "delta-p." Although I am thankful to have had the opportunity to overcome these challenges, at present, a potential limitation of corpus linguistics could be the sizeable amount of background knowledge needed to appreciate the subtleties of a linguistic analysis and the benefits it can offer to individual language teachers.

8.3 Future Research

Perhaps the most obvious direction for future research would be to pilot the tasks and the rating scale developed for this project with English language learners in another context. Students learning English as a second language or learners with a different first language background might construct their academic blog posts in a different way or find the rubric descriptors less applicable to their needs. In addition to analyzing the digital texts produced by different learner populations, a fruitful area of research would be to examine the degree to which the descriptors included on the analytic rubric in this research aligned with written products constructed in response to prompts with unique rhetorical cues or situated outside of the educational domain. To avoid the unintended consequence that the academic blog post rubric might encourage a narrowing of the construct or the possibilities of this digital genre, it is vital to continue collecting samples of learner texts from a variety of contexts and in response to unique

tasks. Likewise, researchers could compare the findings of the rhetorical moves analysis in this research to an analysis of expert or professional examples of educational blogs. Along this vein, there is ample room for needs analyses focused on identifying the types of English writing tasks assigned at the university level in different areas of the world, particularly in view of the educational changes occasioned by the ongoing pandemic. In the Spanish context, given the authority retained by each of the 17 autonomous regions (Castelló et al., 2012), it would be beneficial to explore the nature of writing tasks assigned in universities located outside of the population centers of Barcelona and Madrid.

Another fruitful, yet underexplored area of the effect of analytic rubrics on learners' written production would be to investigate potential gender-based differences. Whereas Kocakülah (2010) did not find a meaningful gender difference in the rubric-based scores of a scientifically-themed written task, Andrade et al. (2009) uncovered a significant effect in the interaction of rubric use and time on girls' self-efficacy, as measured by an adapted version of the Writing Self-Efficacy Scale. Specifically, female participants in the rubric group reported significantly higher self-efficacy values compared to female participants in the control group and male participants in both groups. Seeing as the ratio of men to women in the longitudinal phase of this study was 7:24, the current data set would not be useful for an investigation of any gender-based differences. In fact, the scarcity of male participants underscores the importance of this line of research, without which it would not be possible to generalize the significant findings on rubric use to learners who do not identify as female.

In addition to investigating potential gender-based differences in rubric use, a worthy avenue for future research would be to compare learner writing processes when composing an academic blog post both with and without a rubric. Researchers in task-based language teaching

have long been concerned with exploring the impact of different combinations of task characteristics and task conditions on the processes of task-based interaction (Bygate et al., 2001), and bringing this line of research into the domain of second language writing assessment could inspire other researchers to investigate the impact of rubric use. Recently, Galbraith and Vedder (2020) reported on several advances in the investigation of second language writing processes, including think-aloud protocols, keystroke logging, and eye-tracking. For example, López-Serrano et al. (2019) employed think-aloud protocols during writing to investigate the strategies participants used to solve various language-related episodes (LREs). Their research led to the creation of a complex coding scheme for LREs, which could be applied to investigations of learner processing behavior involving a rubric. Investigations of second language learners' pausing behavior via keystroke logging (Barkaoui, 2019), could reveal whether or not learners who have access to a rubric employ more strategic pausing behavior, and eye-tracking studies, such as the one carried out by Révész et al. (2019) could shed light on the points at which learners refer to the rubric for guidance, whether at the planning stage or during writing. In sum, research into the effectiveness of analytic rubrics would be enriched by data obtained via the aforementioned methods, and information about learner processes could provide further validity evidence in favor of the interpretation of rubric scores for a particular use.

A fourth line of research could expand the scope of the project to address more directly the social nature of personal and educational blogging. For example, Vurdien (2013) addressed the language learning potential hidden in the commenting feature provided by most blog platforms, and Reinhardt (2019) acknowledged the ability of academic blogs to foster a shared reading-writing experience for invested learners. Amid the isolation brought about by the COVID-19 pandemic, online genre-based tasks that can encourage learner-learner

communication without having to meet face-to-face acquire even greater legitimacy. In addition to investigations of learners' written or multi-modal comments, researchers may want to consider the ways in which the analytic rubric developed in this study could be adjusted to incorporate the dialogical nature of the commenting feature. Finally, the social aspect of educational blogging could be extended to the co-creation of genre-based analytic rubrics. The rubric used in Kocakülah (2010) was created in conjunction with the student participants, and it is possible that involving learners in the creation of a genre-based rating scale will set the stage for the transfer of those skills to the analysis of an unfamiliar genre.

8.4 Closing

The ambitious research project described in this dissertation sought to answer relevant questions concerning the effectiveness of a genre-based analytic rubric and the means by which teachers and learners can uncover the expectations of potentially unfamiliar digital genres. The final project contributes to literature on scale development and revision, to investigations on the impact of a well-designed rubric, and to research on linguistic measures of written longitudinal development. Specifically, this study and the rubrics presented herein address several practical concerns related to the assessment of academic blogs in secondary and tertiary education around the world. Even though the nature of the academic blog or discussion board post is not likely to remain stable over the coming years, both the initial and the revised rubrics may serve to connect future investigations concerned with student performance in this increasingly employed format. Whether these investigations tackle the applicability of the rating scale to different contexts, the nature of learner processing behavior while using the rubric, or the means by which practitioners can capitalize on the social nature of the academic blog, this study makes a meaningful contribution to efforts supporting transparency in the assessment of second language writing.

REFERENCES

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR impact factors. *Language Teaching Research, 23*(6), 727-744. https://doi.org/10.1177/1362168818767191

Alcón-Soler, E. (2018). Effects of task-supported language teaching on learning how to mitigate email requests. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics*. John Benjamins.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Alwi, N. A. N. M., Adams, R., & Newton, J. (2012). Writing to learn via text chat: Task implementation and focus on form. *Journal of Second Language Writing, 21*(1), 23-39. https://doi.org/10.1016/j.jslw.2011.12.001

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andon, N., Dewey, M., & Leung, C. (2018). Tasks in the pedagogic space: Using online discussion forum tasks and formative feedback to develop academic discourse skills at master's level. In V. Samuda, K. Van den Branden, & M. Bygate (Eds.), *TBLT as a researched pedagogy* (pp. 235-264). John Benjamins.

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17*(2), 199-214. https://doi.org/10.1080/09695941003696172

Andrade, H. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research, 102*(4), 287-302. https://doi.org/10.3200/JOER.102.4.287-302

Andrade, M. S., & Evans, N. W. (2013). *Principles and practices for response in second language writing: Developing self-regulated learners*. Routledge.

Aquino-Cutcher, A., Asplin, W., Bohlke, D., & Lambert, J. (2016). *Final draft 3*. Cambridge University Press.

Arslan, R. Ş, & Şahin-Kızıl, A. (2010). How can the use of blog software facilitate the writing process of English language learners? *Computer Assisted Language Learning, 23*(3), 183-197. https://doi.org/10.1080/09588221.2010.486575

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin.

Ashby-King, D. T., Iannacone, J. I., Ledford, V. A., Farzad-Philliphs, A., Salzano, M., & Anderson, L. B. (2021). Expanding and constraining critical communication pedagogy in the introductory communication course: A critique of assessment rubrics. *Communication Teacher*. https://doi.org/10.1080/17404622.2021.1975789

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1983). The construct validity of the FSI oral interview. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 154-169). Newbury House.

Bachman, L., & Palmer, L. (2010). *Language assessment in practice*. Oxford University.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257-279. https://doi.org/10.1177/026553229601300303

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions.* Heinle.

Balota, D.A., Yap, J. Y., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445-459. https://doi.org/10.3758/BF03193014

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2019). What can L2 writers' pausing behavior tell us about their L2 writing processes? *Studies in Second Language Acquisition, 41*, 529-554. https://doi.org/10.1017/S027226311900010X

Beaumont, J. (2012). *Focus on writing 4*. Pearson.

Black, R. W. (2006). Language, culture, and identity in online fanfiction. *E-Learning, 3*(2), 170-184. https://doi.org/10.2304/elea.2006.3.2.170

Blake, R. J. (2008). *Brave new digital classroom : Technology and foreign language learning*, Georgetown University. https://ebookcentral.proquest.com/lib/uhm/detail.action?docID=547784.

Bloch, J. (2007). Abdullah's blogging: A generation 1.5 student enters the blogosphere. *Language Learning & Technology, 11*(2), 128-141. http://dx.doi.org/10125/44107

Bloch, J. (2008). *Technologies in the second language composition classroom.* University of Michigan.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Brindley, G. (2009). Task-centered language assessment in language learning: The promise and the challenge. In J. Norris, M. Bygate, & K. Van den Branden (Eds.), *Task-based language teaching* (pp. 435-454). John Benjamins.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343-368. http://dx.doi.org/10.1080/00131911.2014.929565

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Heinle & Heinle.

Brown, J. D. (1998). Language testing: Purposes, effects, options, and constraints. *The TESOLANZ Journal*, *6*, 13-30.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

Brown, J. D. (2009). Principal components analysis and exploratory factor analysis—Definitions, differences, and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(1), 26-30. https://hosted.jalt.org/test/PDF/Brown29.pdf

Brown, J. D. (2012a). Developing rubrics for language assessment. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 1-9). University of Hawaii National Foreign Language Resource Center.

Brown, J. D. (2012b). Introduction to rubric-based assessment. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 13-31). University of Hawaii National Foreign Language Resource Center.

Brown, J. D. (2014). *Mixed methods research for TESOL.* Edinburgh University Press.

Brown, J. D., Hudson, T. D., Norris, J., & Bonk, W. (2002). *An investigation of second language task-based performance assessments.* University of Hawaii.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42-65. https://doi.org/10.1016/j.jslw.2014.09.005

Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies. *CALICO Journal, 33*(1), 117-141. https://doi.org/10.1558/cj.v33i1.26374

Bygate, M. (2018). Introduction. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 1-26). John Benjamins.

Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Pearson.

Byrnes, H. (2009). The role of task and task-based assessment in a content-oriented collegiate FL curriculum. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* (pp. 477-493). John Benjamins.

Byrnes, H. (2011). Beyond writing as language learning or content learning: Construing foreign language writing as meaning-making. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 133-157). John Benjamins.

Byrnes, H. (2012). Conceptualizing FL writing development in collegiate settings: A genre-based systemic functional linguistic approach. In R. M. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 191-219). Walter de Gruyter.

Byrnes, H. (2014). Theorizing language development at the intersection of 'task' and L2 writing: Reconsidering complexity. In H. Byrnes, & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 79-103). John Benjamins.

Byrnes, H., & Manchón, R. M. (2014). Task-based language learning: Insights from and for L2 writing: An Introduction. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 1-23). John Benjamins.

Cangelosi, J. (2000). *Classroom management strategies: Gaining and maintaining students' cooperation*. John Wiley & Sons.

Canto, S., de Graaff, R., & Jauregi, K. (2014). Collaborative tasks for negotiation of intercultural meaning in virtual worlds and video-web communication. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 183-212). John Benjamins.

Caplan, N. A. (2019). Have we always taught the five-paragraph essay? In N. A. Caplan & A. M. Johns (Eds.), *Changing practices for the L2 writing classroom: Moving beyond the five-paragraph essay* (pp. 2-23). University of Michigan.

Casanave, C. P. (2017). *Controversies in second language writing: Dilemmas and decisions in research and instruction*. University of Michigan.

Castelló, M., Mateos, M., Castells, N., Iñesta, A., Cuevas, I., & Solé, I. (2012). Academic writing practices in Spanish universities. *Electronic Journal of Research in Educational Psychology, 10*(2), 569-590. https://doi.org/10.25115/ejrep.v10i27.1517

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V. Ageyev, & S. Miller (Eds.), *Vygotsky's educational theory and practice in cultural context.* Cambridge University Press.

Chapman, M. (2016). The effect of the prompt on writing product and process: A mixed-methods approach [PhD Thesis, University of Bedfordshire]. PhD e-theses. http://hdl.handle.net/10547/621846

Chen, R. T.-H. (2015). L2 blogging: Who thrives and who does not? *Language Learning & Technology, 19*(2), 177-196. http://dx.doi.org/10125/44423

Chen, W.-C., Shih, Y.-C. D., Liu, G.-Z. (2015). Task design and its induced learning effects in a cross-institutional blog-mediated telecollaboration. *Computer Assisted Language Learning, 28*(4), 285-305. https://doi.org/10.1080/09588221.2013.818557

Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Learner corpus research* (pp. 185-206). Cambridge University Press.

Conference on College Composition and Communication [CCCC]. (2009a). CCCC statement on second language writing and writers. http://www.ncte.org/cccc/resources/positions/secondlangwriting

Conference on College Composition and Communication [CCCC]. (2009b) Writing assessment: A position statement. http://cccc.ncte.org/cccc/resources/positions/writingassessment

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94-100. https://doi.org/10.1080/09296171003643098

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches.* SAGE.

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), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43-56. https://doi.org/10.1016/j.asw.2016.03.001

Crusan, D., & Ruecker, T. (2019). Standardized testing pressures and the temptation of the five-paragraph essay. In N. A. Caplan & A. M. Johns (Eds.), *Changing practices for the L2 writing classroom: Moving beyond the five-paragraph essay* (pp. 201-220). University of Michigan Press.

Davies, M. (2010). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159-190. https://doi.org/10.1075/ijcl.14.2.02dav

Davis, L., & Kondo-Brown, K. (2012). Assessing student language performance: Types and uses of rubrics. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 33-55). University of Hawaii National Foreign Language Resource Center.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8. https://doi.org/10.3102/0013189X032001005

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability. Research Bulletin 61-15*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Ducate, L. C., & Lomicka, L. L. (2008). Adventures in the blogosphere: From blog readers to blog writers. *Computer Assisted Language Learning, 21*(1), 9-28. https://doi.org/10.1080/09588220701865474

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*, 88-115. https://doi.org/10.1016/j.asw.2009.04.001

EBAU Inglés II. Argumentative Writing Scoring Rubric. (2021, September 15). Universidad de Murcia. https://www.um.es/en/web/vic-estudios/contenido/acceso/pau/ebau-materias-coordinadores/ingles/documentacion

Elgort, I. (2017). Blog posts and traditional assignments by first- and second-language writers. *Language Learning & Technology, 21*(2), 52-72. https://dx.doi.org/10125/44611

Elola, I., & Oskoz, A. (2010). Collaborative writing: Fostering foreign language and writing conventions development. *Language Learning and Technology, 14*(3), 30-49. http://dx.doi.org/10125/44226

Elola, I., & Oskoz, A. (2017). Writing with 21st century social tools in the L2 classroom: New literacies, genres, and writing practices. *Journal of Second Language Writing, 36*, 52-60. https://doi.org/10.1016/j.jslw.2017.04.002

Ferris, D. R. (2011). *Treatment of error in second language student writing*. University of Michigan Press.

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing, 19*, 6-23. https://doi.org/10.1016/j.asw.2013.09.004

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 composition: Purpose, process, and practice.* Routledge.

Fulcher, G. (1988). *Lexis and reality in oral evaluation* (Annual Meeting of the International Association of Teachers of English as a Foreign Language ED 298 759; pp. 1-62). https://files.eric.ed.gov/fulltext/ED298759.pdf

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. Pearson.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction.* SAGE.

Galbraith, D., & Vedder, I. (2019). Methodological advances in investigating L2 writing processes. *Studies in Second Language Acquisition, 41*(si3), 633-645. https://doi.org/10.1017/S0272263119000366

Gee, J. P. (2005). *An introduction to discourse analysis: Theory and method* (2nd ed.). Routledge.

Godwin-Jones, R. (2006). Emerging Technologies. Tag clouds in the blogosphere: Electronic literacy and social networking. *Language Learning & Technology, 10*(2), 8-15. http://dx.doi.org/10125/44055

Godwin-Jones, R. (2018). Second language writing online: An update. *Language Learning & Technology, 22*(1), 1-15. https://dx.doi.org/10125/44574

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70-105. https://doi.org/10.1080/09588221.2012.700315

González-Lloret, M. (2014). The need for needs analysis in technology-mediated TBLT. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 23-50). John Benjamins.

González-Lloret, M., & Ortega, L. (2014). Towards technology-mediated TBLT: An introduction. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 1-22). John Benjamins.

Graves, K. (2000). *Designing language courses: A guide for teachers*. Heinle & Heinle.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs. Research Report RR-95-44*. Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01678.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* SAGE.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295-303. https://doi.org/10.1177/026553229701400306

Harklau, L. (2002). The role of writing in classroom second language acquisition. *Journal of Second Language Writing, 11*, 329-350. https://doi.org/10.1016/S1060-3743(02)00091-7

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17,* 228-250. https://doi.org/10.1016/j.asw.2012.06.003

Hedgcock, J. S., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts.* Routledge.

Hindley, D., & Clughen, L. (2018). 'Yay! Not another academic essay!' Blogging as an alternative academic genre. *Journal of Writing in Creative Practice, 11*(1), 83-97. https://doi.org/10.1386/jwcp.11.1.83_1

Hinkel, E. (2015). *Effective curriculum for teaching L2 writing: Principles and techniques.* Routledge.

Humphry, S. M., Heldsinger, A. M. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263. https://doi.org/10.3102/0013189X14542154

Hyland, K. (2003). *Second language writing*. Cambridge University Press.

Hyland, K. (2004). *Genre and second language writing.* University of Michigan Press.

Hyland, K. (2011). Learning to write: Issues in theory, research, and pedagogy. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 17-35). John Benjamins.

Hyland, K. (2016). *Teaching and researching writing*. Taylor & Francis.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly, 30*(4), 693-722. https://doi.org/10.2307/3587930

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

Jacobson, B., Pawlowski, M., & Tardy, C. M. (2021). Make your "move": Writing in genres. In D. L. Driscoll, M. Heise, M. K. Stewart, & M. Vetter (Eds.), *Writing spaces: Readings on writing*, Volume 4, (pp. 217-238). https://writingspaces.org/?page_id=790&fbclid=IwAR2GZREBSYuebvHI6M7CSA_SW7G1iYjyECY_cltcnVyprlLf-3iBIy-KIx4

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51-66. https://doi.org/10.1016/j.asw.2015.07.002

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(s1), 87-106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johns, A. M. (2018, March). *The five-paragraph essay can't serve as a security blanket!* Paper presented at the TESOL International Convention, Chicago.

Johns, A. M. (2019). Writing in the interstices: Assisting novice undergraduates in analyzing authentic writing tasks. In N. A. Caplan & A. M. Johns (Eds.), *Changing practices for the L2 writing classroom: Moving beyond the five-paragraph essay* (pp. 133-149). University of Michigan Press.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840-852. https://doi.org/10.1080/02602938.2013.875117

Kachchaf, R., & Solano-Flores, G. (2012). Rater language background as a source of measurement error in the testing of English language learners. *Applied Measurement in Education, 25*, 162-177. https://doi.org/10.1080/08957347.2012.660366

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481-504. https://doi.org/10.1177/0265532219849522

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120-141. https://doi.org/10.1111/modl.12447

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81-96. https://doi.org/10.1016/j.asw.2011.02.003

Kocakülah, M. S. (2010). Development and application of a rubric for evaluating students' performance on Newton's Laws of Motion. *Journal of Science Education and Technology, 19*, 146-164. https://doi.org/10.1007/s10956-009-9188-9

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(2), 3-31. https://doi.org/10.1191/0265532202lt218oa

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3*, 231-255. https://doi.org/10.1016/1060-3743(94)90018-3

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321-336. https://doi.org/10.1177/0265532216663991

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine-grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. GSU ScholarWorks https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513-535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition, 43*(4), 781-812. https://doi.org/10.1017/S0272263120000546

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly, 18*(2), 154-170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K. & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora*. Multilingual Matters.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practices, 21*(1), 23-30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring.* Heinle.

Larsen-Freeman, D. (2013). A promising combination: Complexity theory, design-based research, and CALL. In J.C. Rodríguez & C. Pardo-Ballester (Eds.), *Design-based research in CALL* (pp. 23-30). CALICO.

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*, 34-50. https://doi.org/10.1016/j.jslw.2012.11.003

Leki, I. (2011). Learning to write in a second language: Multilingual graduates and undergraduates expanding genre repertories. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 85-109). John Benjamins.

Lightbown, P., & Spada, N. (2006). *How languages are learned.* Oxford University Press.

Linacre, J. M. (2019). *FACETS, 3.82*. MESA.

López-Serrano, S., Roca de Larios, J., & Manchón, R. M. (2019). Language reflection fostered by individual L2 writing tasks: Developing a theoretically motivated and empirically based coding system. *Studies in Second Language Acquisition, 41*(3), 503-527. https://doi.org/10.1017/S0272263119000275

Long, M. (2015). *Second language acquisition and task-based language teaching.* Wiley.

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. https://doi.org/10.5054/tq.2011.240859

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Macaro, E. (2014). Reframing task performance: The relationship between tasks, strategic behaviour, and linguistic knowledge in writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 53-77). John Benjamins.

Manchón, R. M. (2009). Broadening the perspective of L2 writing scholarship: The contribution of research on foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research.* Multilingual Matters.

Manchón, R. M. (2011a). Situating the learning-to-write and writing-to-learn dimensions of L2 writing. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 3-14). John Benjamins.

Manchón, R. M. (2011b). Writing to learn the language: Issues in theory and research. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 61-82). John Benjamins.

Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 27-52). John Benjamins.

Manchón, R. M., & Roca De Larios, J. (2011). Writing to learn in FL contexts: Exploring learners' perceptions of the language learning potential of L2 writing. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 181-208). John Benjamins.

Manchón, R. M., Roca De Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing processes. Implications for theory. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 102-129). Multilingual Matters.

Martin, J. R. (1999). Modelling context: A crooked path of progress in contextual linguistics. In M. Ghadessy (Ed.), *Text and context in functional linguistics* (pp. 25-61). John Benjamins.

Martin, J. R. (2002). Meaning beyond the clause: SFL perspectives. *Annual Review of Applied Linguistics, 22*, 52-74. https://doi.org/10.1017/S026719050200003X

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Matsuda, P. K. (2012). Let's face it: Language issues and the writing program administrator. *WPA: Writing Program Administration, 36*(1), 141-163. http://associationdatabase.co/archives/36n1/36n1matsuda.pdf

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392. https://doi.org/10.3758/BRM.42.2.381

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295-323. https://doi.org/10.1177/00238309010440030101

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. https://doi.org/10.1037/1082-989X.1.1.30

McGuire, P. L. (1996). Language Planning and Policy and the ELT Profession in Selected Central American Countries. *TESOL Quarterly 30*(3), 606-611. https://www.jstor.org/stable/3587703

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

Melzer, D. (2014). *Assignments across the curriculum: A national study of college writing*. Utah State University Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23. https://doi.org/10.3102/0013189X023002013

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1177/026553229601300302

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256. https://doi.org/10.1177/026553229601300302

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*, 151-167. https://doi.org/10.1080/00335638409383686

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes, 50*, 40-63. https://doi.org/10.1016/j.esp.2017.11.006

Nelson, D. L., McEvoy, C., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402-407. https://doi.org/10.3758/BF03195588

Nitta, R., & Baba, K. (2014). Task repetition and L2 writing development: A longitudinal study from a dynamic systems perspective. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 107-136). John Benjamins.

Nitta, R., & Baba, K. (2018). Understanding benefits of repetition from a complex dynamic systems perspective: The case of a writing task. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 279-310). John Benjamins.

Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *The Modern Language Journal, 90*(6), 576-583. https://doi.org/10.1111/j.1540-4781.2006.00466_2.x

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36,* 230-244. https://doi.org/10.1017/S0267190516000027

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002a). Examinee abilities and task difficulty in task-based L2 performance assessment. *Language Testing, 19*(4), 395-418. https://doi.org/10.1191/0265532202lt237oa

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. (2002b). *Designing second language performance assessments.* University of Hawaii.

Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly, 17*(1), 60-84. https://doi.org/10.1080/15434303.2019.1674854

Ohta, A. (2000). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In J. Lantolf (Ed.), *Sociocultural theory and second language learning.* Oxford University Press.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2009). Studying writing across EFL contexts: Looking back and moving forward. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 232-255). Multilingual Matters.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127-155). De Gruyter.

Ortega, L. (2019). *Reconciling ethics and politics with methodological rigor: Tips for (quantitative) researchers*. International Doctoral Summer School, University of Malta, June 14, 2019.

Ortmeier-Hooper, C. (2019). Rethinking the five-paragraph essay as a scaffold in secondary school. In N. A. Caplan & A. M. Johns (Eds.), *Changing practices for the L2 writing classroom: Moving beyond the five-paragraph essay* (pp. 89-115). University of Michigan.

Oskoz, A., & Elola, I. (2014). Promoting foreign language collaborative writing through the use of Web 2.0 tools and tasks. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 115–148). John Benjamins.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. SAGE.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 1-19. https://doi.org/10.1016/j.edurev.2020.100329

Patharakorn, P. (2018). *Assessing interactional competence in a multiparty roleplay task: A mixed-methods study* (Publication No. 13423190) [Doctoral Dissertation, University of Hawaii at Manoa]. ProQuest Dissertations and Theses.

Peyton, J. (2000). *Dialogue journals: Interactive writing to develop language and literacy* (ED450614). ERIC. https://files.eric.ed.gov/fulltext/ED450614.pdf

Pinkman, K. (2005). Using blogs in the foreign language classroom: Encouraging learner independence. *The JALT CALL Journal, 1*(1), 12-24. https://doi.org/10.29140/jaltcall.v1n1.2

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, *55*, 72-75. https://www.ascd.org/el/articles/whats-wrong-and-whats-right-with-rubrics

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Association for Supervision and Curriculum Development.

Rebuschat, P., Hamrick, P., Sachs, R., Reistenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures. In J. Bergsleithner, S. Frota, & J. Yoshioka (Eds.), *Noticing: L2 studies and essays in honor of Dick Schmidt* (pp. 249-270). National Foreign Language Resource Center, University of Hawaii at Manoa.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16. https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Reeder, K. (2010). Edubba: Real-world writing tasks in a virtual world. In M. Thomas & H. Reinders (Eds.), *Task-based language learning and teaching with technology* (pp. 176-196). Continuum.

Reeves, T. C., & McKenney, S. (2013). Computer-assisted language learning and design-based research: Increased complexity for sure, enhanced impact perhaps. In J.C. Rodríguez & C. Pardo-Ballester (Eds.), *Design-based research in CALL* (pp. 9-22). CALICO.

Reinhardt, J. (2019). Social media in second and foreign language teaching and learning: Blogs, wikis, and social networking. *Language Teaching, 52*(1), 1-39. https://doi.org/10.1017/S0261444818000356

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition*, *41*(si3), 605-631. https://doi.org/10.1017/S027226311900024X

Richards, J. C., & Rodgers, T. S. (1986). *Approaches and methods in language teaching.* Cambridge University.

Robinson, P. (2011). Second language task complexity, the cognition hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3-38). John Benjamins.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning--Insights from and for L2 writing* (pp. 163-189). John Benjamins.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Palgrave Macmillan.

Sánchez López, A. J. (2018). *External task repetition: The role of modality, written corrective feedback and proficiency* (Unpublished doctoral dissertation). University of Murcia, Murcia, Spain.

Sauro, S. (2012). L2 performance in text-chat and spoken discourse. *System, 40*(3)*,* 335-348. https://doi.org/10.1016/j.system.2012.08.001

Sauro, S. (2014). Lessons from the fandom: Technology-mediated tasks for language learning. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 239-261). John Benjamins.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77-101). Multilingual Matters.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209-231.

Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*(4), 513-521. https://doi.org/10.2307/330053

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University.

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning & Technology, 4*(1), 77-110. http://dx.doi.org/10125/44470

Sotillo, S. M. (2016). An update on discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning & Technology, 20*(2), 166-171. http://dx.doi.org/10125/44470

Stevens, D. D., & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus.

Stockwell, G. (2010). Effects of multimodality in computer-mediated communication tasks. In M. Thomas & H. Reinders (Eds.), *Task-based language learning and teaching with technology* (pp. 83-104). Continuum.

Sun, Y.-C. (2010). Extensive writing in foreign-language classrooms: A blogging approach. *Innovations in Education and Teaching International, 47*(3), 327-339. https://doi.org/10.1080/14703297.2010.498184

Sun, Y.-C., & Chang, Y. J. (2012). Blogging to learn: Becoming EFL academic writers through collaborative dialogues. *Language Learning & Technology, 16*(1), 43-61. http://dx.doi.org/10125/44274

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon & M. S. Berns (Eds.), *Initiatives in communicative language teaching* (pp. 185-201). Addison-Wesley.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (2009). The concept of task. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* (pp. 41-55). John Benjamins.

Sykes, J. M. (2014). TBLT and synthetic immersive environments: What can in-game task restarts tell us about design and implementation? In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 149-182). John Benjamins.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tardy, C. M. (2012). A rhetorical genre theory perspective on L2 writing development. In R. M. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 165-190). Walter de Gruyter.

Tardy, C. M. (2016). *Beyond convention: Genre innovation in academic writing*. University of Michigan.

Tardy, C. M. (2018, March). *Building linguistic and rhetorical flexibility in first-year writing.* Paper presented at the TESOL International Convention, Chicago.

Tardy, C. M. (2019). Is the five-paragraph essay a genre? In N. A. Caplan & A. M. Johns (Eds.), *Changing practices for the L2 writing classroom: Moving beyond the five-paragraph essay* (pp. 24-41). University of Michigan.

Thomas, M., & Reinders, H. (2010). Deconstructing tasks and technology. In M. Thomas & H. Reinders (Eds.), *Task-based language learning and teaching with technology* (pp. 1-16). Continuum.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22. https://doi.org/10.1177/0265532215594830

Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 359-372). Springer International. https://doi.org/10.1007/978-3-319-02261-1_24

Turner, J. (1997). Creating content-based language tests: Guidelines for teachers. In M. Snow & & D. Brinton (Eds.), *The content-based classroom.* Longman.

Turner, J. L. (2014). *Using statistics in small-scale language education research: Focus on non-parametric data*. Taylor & Francis.

Turner, C.E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review, 56*(4), 555-584. https://doi.org/10.3138/cmlr.56.4.555

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70. https://doi.org/10.2307/3588360

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12. https://doi.org/10.1093/elt/49.1.3

Van den Branden, K. (2006). Introduction: Task-based language education in a nutshell. In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 1-16). Cambridge University.

Van den Branden, K., Bygate, M., & Norris, J. (2009). Task-based language teaching: Introducing the reader. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* (pp. 1-13). John Benjamins.

Van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective.* Kluwer Academic.

Van Lier, L., & Matsuo, N. (2000). Varieties of conversational experience looking for learning opportunities. *Applied Language Learning, 11*(2), 265-287. https://www.dliflc.edu/wp-content/uploads/2014/04/all-v11-n2.pdf

Vandergriff, I. (2016). *Second-language discourse in the digital world*. John Benjamins.

Verspoor, M., Schmid, M., & Xu, X. (2012). A dynamic usage-based perspective on L2 writing. *Journal of Second Language Writing, 21*, 239-263. https://doi.org/10.1016/j.jslw.2012.03.007

Vurdien, R. (2013). Enhancing writing skills through blogging in an advanced English as a Foreign Language class in Spain. *Computer Assisted Language Learning*, *26*(2), 126-143. https://doi.org/10.1080/09588221.2011.639784

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Harvard University.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology and Research Development, 53*(4), 5-23. https://doi.org/10.1007/BF02504682

Ward, C. (2012). *Focus on writing 3*. Pearson.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing.* Cambridge University.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*, 194-209. https://doi.org/10.1016/j.jslw.2007.07.004

Weigle, S. C. (2016). Second language writing assessment. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of Second and Foreign Language Writing* (pp. 473-493). De Gruyter.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

Willis, J. (1996). *A framework for task-based learning*. Longman.

Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *The English Journal, 96*(4), 62-66. https://doi.org/10.2307/30047167

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Winke, P. M. (2014). Formative, task-based oral assessment in an advanced Chinese-language class. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 263-294). John Benjamins.

Winke, P. M., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 38-54. https://doi.org/10.1016/j.asw.2015.05.002

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii.

Yanguas, Í. (2010). Oral computer-mediated interaction between L2 learners: It's about time! *Language Learning & Technology, 14*(3), 72-93. http://dx.doi.org/10125/44227

Yasuda, S. (2012). *The implementation of genre-based tasks in foreign language writing instruction: A longitudinal study of writers' rhetorical awareness, writing quality, and lexicogrammatical choices* [Doctoral dissertation, University of Hawaii at Manoa]. UHM Scholar Space. http://hdl.handle.net/10125/101043

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods* (Publication No. 3577270) [Doctoral dissertation, University of Hawaii at Manoa]. ProQuest Dissertations Publishing.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47,* 2-15. https://doi.org/10.1016/j.asw.2020.100505

Ziegler, N. (2016). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics, 36*, 136-163. https://doi.org/10.1017/S0267190516000039

APPENDIX A: DISSERTATION CONSENT FORM

Hi! You are being asked to participate in a research study conducted by Marta González-Lloret and Kristin Rock from the Department of Second Language Studies at the University of Hawaii at Manoa. The results of this study will contribute to Kristin's doctoral dissertation.

What am I being asked to do?
If you participate in this project, you will be asked to complete an asynchronous writing task at several points throughout the semester.

Taking part in this study is your choice.
You can choose to take part or you can choose not to take part in this study. You also can change your mind at any time. If you stop being in the study, there will be no penalty or loss to you.

Why is this study being done?
The purpose of this project is to examine the role of rubrics in promoting second language writing development in asynchronous online tasks. I am asking you to participate because you are a university student who is also an English language learner.

What will happen if I decide to take part in this study?
If you decide to participate in this study, you will be asked to do the following:
First, you will be asked to read through this consent form and to indicate whether or not you choose to take part in the study, which should take about 5 minutes. At this point, you will be asked to fill out a background information questionnaire that should take no more than 10 minutes to complete. Next, you will take a computer-adaptive English proficiency test through Oxford English Testing, which should last approximately 20 minutes. Finally, you will complete a writing task in English at a time and location of your choosing, as long as the date of choice falls within specified collection periods. Only you and I will be present as you work on the writing task, and completion may take anywhere from 30 minutes to 2 hours. Your total participation will take one to two hours. After the first data collection point, we will schedule two more meetings (midway through the semester and at the end of the semester) during which you will complete a similar series of tasks. You will be one of about 60 people in this study.

What are the risks and benefits of taking part in this study?
I do not foresee any risks to participation in this study. However, should participation in the study cause you undue stress or anxiety, you are free to take a break or to withdraw from the project at any time.

The direct benefit to you is likely to be improvement in your written production in English. In addition, the results of this study will advance knowledge of how individuals learn to write in a second language within specific genres. Insight gained from the project could also be applied to future curriculum and test development projects.

Privacy and Confidentiality:
Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. Confidentiality will be maintained by means of destroying (i.e., shredding) physical copies of this consent form. A scanned copy of the consent form will be saved on a password protected computer,

and only I and my advisor at the University of Hawaii will have access to it. Data obtained throughout the research will only be linked to randomly assigned ID numbers, which will not be linked back to you. This data, which includes your written work, will be stored on the same password protected computer, and only I and my advisor at the University of Hawaii will have access to it.

Other agencies that have legal permission have the right to review research records. The University of Hawaii Human Studies Program has the right to review research records for this study. When I report the results of my research project, I will not use your name. I will not use any other personal identifying information that can identify you. I will use pseudonyms (fake names) and report my findings in a way that protects your privacy and confidentiality to the extent allowed by law.

Future Research Studies:
Identifiers will be removed from your identifiable private information and after removal of identifiers, the data may be used for future research studies or distributed to another investigator for future research studies and we will not seek further approval from you for these future studies.

Compensation:
After each meeting, you will receive a $5 gift card for every hour of your participation. If this research leads to creation of new tests or commercially available curricula, you will not receive any payment.

Questions: If you have any questions about this study, please email me at rockk@hawaii.edu. You may also contact my advisor, *Marta González-Lloret* at marta@hawaii.edu. You may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu. to discuss problems, concerns and questions; obtain information; or offer input with an informed individual who is unaffiliated with the specific research protocol. Please visit http://go.hawaii.edu/jRd for more information on your rights as a research participant.

If you agree to participate in this project, please sign and date the following signature page and return it to Kristin Rock. Keep a copy of the informed consent for your records and reference.

Signature(s) for Consent:
I give permission to join the research project entitled, *Exploring the Use of Analytic Rubrics for L2 Writing Development in Online Tasks.*

Name of Participant (Print): _____

Participant's Signature: _____

Signature of the Person Obtaining Consent: _____

Date: _____

Thank you!

APPENDIX B: DISSERTATION CONSENT FORM (SPANISH)

¡Hola! Se le pide que participe en un estudio de investigación realizado por Marta González-Lloret y Kristin Rock del Departamento de Estudios de una Segunda Lengua en la Universidad de Hawai en Manoa. Los resultados de este estudio contribuirán a la tesis doctoral de Kristin.

¿Qué me piden que haga?
Si participa en este proyecto, se le pedirá que complete una tarea de escritura asincrónica en varios momentos a lo largo del semestre.

Participar en este estudio es su elección.
Puede elegir participar o puede optar por no participar en este estudio. También puede cambiar de opinión en cualquier momento. Si deja de participar en el estudio, no habrá ninguna sanción ni pérdida para usted.

¿Por qué se está haciendo este estudio?
El objetivo de este proyecto es examinar el papel de las rúbricas en la promoción del desarrollo de la escritura en un segundo idioma en tareas en línea asíncronas. Le estoy pidiendo que participe porque usted es un estudiante universitario que también aprende inglés.

¿Qué pasará si decido participar en este estudio?
Si decide participar en este estudio, se le pedirá que haga lo siguiente:
Primero, se le pedirá que lea este formulario de consentimiento e indique si decide participar o no en el estudio, lo cual debe tomar unos 5 minutos. En este momento, se le pedirá que complete un cuestionario de antecedentes que no debe tomar más de 10 minutos para completarlo. A continuación, tomará un examen de nivel de conocimiento de inglés adaptativo en computadora a través de las Pruebas de Inglés de Oxford, el cual debe durar 20 minutos aproximadamente. Por último, completará una tarea de escritura en inglés a una hora y en un lugar de su elección, siempre que la fecha de la elección se encuentre dentro de los períodos de recogida específicos. Solo usted y yo estaremos presentes mientras usted trabaja en la tarea de escritura, y la finalización puede tomar de 30 minutos a 2 horas. Su participación total durará de una a dos horas. Después del primer punto de recogida de datos, programaremos dos reuniones más (a mitad del semestre y al final del semestre) durante las cuales usted completará una serie de tareas similares. Usted será una de las aproximadamente 60 personas en este estudio.

¿Cuáles son los riesgos y beneficios de participar en este estudio?
No preveo ningún riesgo para la participación en este estudio. Sin embargo, si la participación en el estudio le causa estrés o ansiedad excesivos, puede tomar un descanso o retirarse del proyecto en cualquier momento.

El beneficio directo para usted probablemente sea una mejora en su producción escrita en inglés. Además, los resultados de este estudio mejorarán el conocimiento de cómo las personas aprenden a escribir en un segundo idioma dentro de géneros específicos. La información obtenida del proyecto también podría aplicarse al futuro plan de estudios y proyectos de desarrollo de pruebas.

Privacidad y confidencialidad:
Cualquier información que se obtenga en relación con este estudio y que pueda identificarse con usted permanecerá confidencial y se divulgará solo con su permiso o según lo exija la ley. La

confidencialidad se mantendrá mediante la destrucción (es decir, la destrucción) de copias físicas de este formulario de consentimiento. Se guardará una copia escaneada del formulario de consentimiento en una computadora protegida con contraseña, y solo yo y mi asesor en la Universidad de Hawai tendremos acceso a ella. Los datos obtenidos a lo largo de la investigación solo se vincularán a números de identificación asignados al azar, que no se vincularán a usted. Estos datos, que incluyen su trabajo escrito, se almacenarán en la misma computadora protegida con contraseña, y solo yo y mi asesor en la Universidad de Hawai tendremos acceso a ellos.
Otras agencias que tienen permiso legal tienen el derecho de revisar los registros de investigación. El Programa de Estudios Humanos de la Universidad de Hawai tiene el derecho de revisar los registros de investigación para este estudio. Cuando informe los resultados de mi proyecto de investigación, no usaré su nombre. No utilizaré ninguna otra información de identificación personal que pueda identificarle. Usaré seudónimos (nombres falsos) e informaré mis resultados de una manera que proteja su privacidad y confidencialidad en la medida que lo permita la ley.

Futuros Estudios de Investigación:
Los identificadores se eliminarán de su información privada identificable y después de la eliminación de los identificadores, los datos podrán ser utilizados para futuros estudios de investigación o se distribuirán a otro investigador para futuros estudios de investigación y no solicitaremos su aprobación adicional para estos futuros estudios.

Compensación:
Después de cada reunión, recibirá una tarjeta de regalo de 5 libras por cada hora de su participación. Si esta investigación conduce a la creación de nuevas pruebas o planes de estudio disponibles en el mercado, usted no recibirá ningún pago.

Preguntas: Si tiene alguna pregunta sobre este estudio, por favor, envíeme un correo electrónico a rockk@hawaii.edu. También puede contactar con mi asesora, Marta González-Lloret, en marta@hawaii.edu. Puede comunicarse con el Programa de Estudios Humanos de la Universidad de Hawai al 808.956.5007 o uhirb@hawaii.edu para discutir problemas, inquietudes y preguntas; obtener información; u ofrecer información con una persona informada que no esté afiliada con el protocolo de investigación específico. Por favor, visite http://go.hawaii.edu/jRd para obtener más información sobre sus derechos como participante de investigación.
Si acepta participar en este proyecto, por favor, firme y feche la siguiente página de firma y devuélvala a Kristin Rock. Conserve una copia del consentimiento informado para sus registros y referencias.

Firma (s) de consentimiento:
Doy permiso para unirme al proyecto de investigación titulado *Explorar el uso de rúbricas analíticas para el desarrollo de escritura en una segunda lengua en tareas en línea*.

Nombre del participante (en letra de imprenta): _____

Firma del participante: _____

Firma de la persona que obtiene el consentimiento: _____

Fecha: _____

¡Gracias!

APPENDIX C: BACKGROUND INFORMATION FORM

1. *Study ID Number /* Número de identificación de estudio*:*

2. *Age /* Edad*:*

3. *Gender /* Género*:*

4. *Do you have normal vision, or if you need glasses or contacts, are you wearing them? Please* **circle** *your answer.*
   ¿Tiene una visión normal o, si necesita gafas o lentes de contacto, las está usando? Por favor, ***rodee*** *su respuesta.*

   Yes No

5. *Please indicate the highest level of education you have completed by* **circling** *the appropriate line.* If you are still in school, *please include your year of study (i.e. first, second, third, fourth).*
   Por favor, indique el nivel de educación más alto que haya completado **rodeando** la línea correspondiente. *Si aún está en la escuela*, por favor, incluya su año de estudio (es decir, primero, segundo, tercero, cuarto).

   Secondary school or high school / *Secundaria o Bachillerato*

   University: Year _____ (if applicable) / *Universidad: Año _____ (si corresponde)*

   Some graduate work / *Algún trabajo de postgrado*

   Master's degree / *Máster*

   Doctoral degree / *Doctorado*

6. *Did you major, or are you majoring in English philology?*
   ¿Se especializó o se está especializando en Filología Inglesa?

7. *Have you taken linguistics or language analysis courses in the past?*
   ¿Ha tomado cursos de lingüística o de análisis de idiomas en el pasado?

8. *Are you currently enrolled in any English courses at the University of Murcia?* If yes*, please state the course title(s).*
   ¿Actualmente está matriculado en algún curso de inglés en la Universidad de Murcia? *En caso afirmativo*, indique el título del curso.

9. *Do you write regularly in English?* If yes*, how often do you write in English, to whom do you write, and in what genres (i.e., a journal, a blog, university work, emails, letters, etc.)?*
   ¿Escribe regularmente en inglés? *En caso afirmativo*, ¿con qué frecuencia escribe en inglés, a quién escribe y en qué géneros (es decir, una revista, un blog, trabajo universitario, correos electrónicos, cartas, etc.)?

10. *Do you maintain a blog in English or Spanish?* If yes*, in which language? How often do you upload posts? What is the blog about?*
¿Mantiene un blog en inglés o en español? En caso afirmativo, ¿en qué idioma? ¿Con qué frecuencia subes publicaciones? ¿De qué trata el blog?

11. *Do you follow any bloggers currently?* If yes*, whom? What are the blogs about?*
¿Sigue a algún bloguero actualmente? Si es así, ¿a quién? ¿Sobre qué tratan los blogs?

12. *What is your first language? If there are two or more, please list all languages that you consider are native.*
¿Cuál es su primera lengua? Si hay dos o más, por favor, enumere todas las lenguas que considere que son nativas.

13. *Is your first language spoken at home? /* ¿Se habla su primera lengua en casa?

14. *Do you know any additional languages? /* ¿Conoces algún idioma adicional?

If yes, please indicate your level of familiarity with each additional language below:
En caso afirmativo, por favor, indique a continuación su nivel de familiaridad con cada idioma adicional:

Language / Idioma*:*

Context of learning/knowledge: (Please highlight all contexts in which you have been exposed to the language, and then indicate the approximate number of hours you would attribute to each context)
Contexto de aprendizaje / conocimiento: (Por favor, subraye todos los contextos en los que ha estado expuesto al idioma, y luego indique la cantidad aproximada de horas que atribuiría a cada contexto).

Elementary School / Primaria _____
Middle School / Secundaria _____
High School / Bachillerato _____
University / Universidad _____
Private tutor / Tutor privado_____
With family members / Con miembros de la familia _____
Computer software / Software de ordenador _____
Study abroad or in-country / Estudios en el extranjero o en el país _____
Other (please describe) / Otro (por favor, describa) _____

How would you rate your speaking, listening, reading and writing abilities in this language, with 1 signifying beginner and 5 signifying advanced/near-native fluency?
¿Cómo calificaría sus habilidades de hablar, escuchar, leer y escribir en este idioma, con 1 significando principiante y 5 significando con fluidez avanzada / casi nativa?

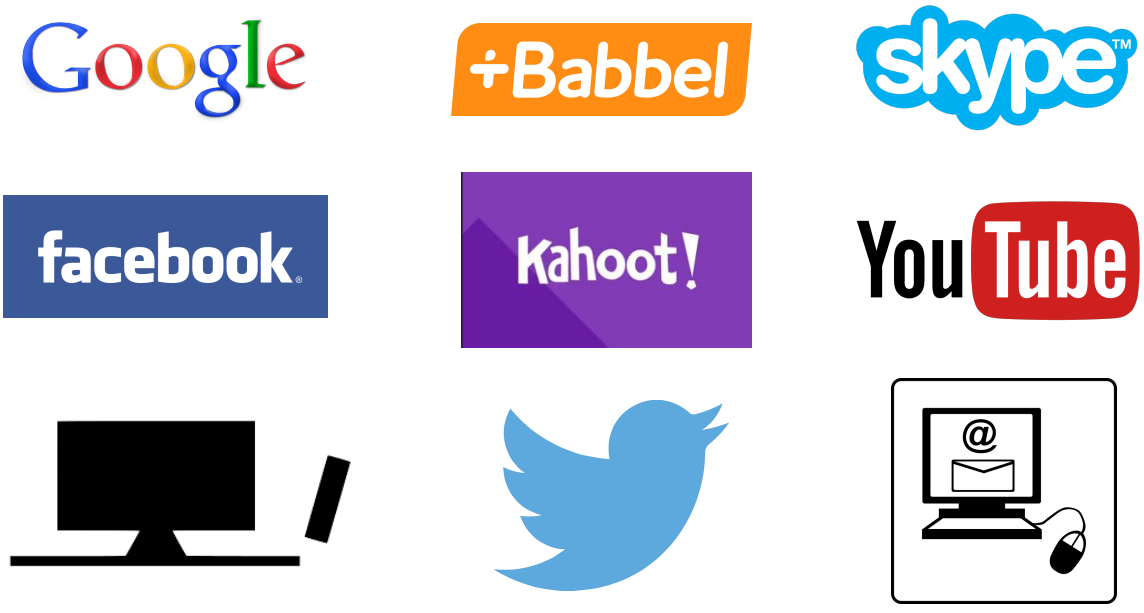
Speaking / *Habla*: _____ Listening / *Escucha*: _____
Reading / *Lectura*: _____ Writing / *Escritura*: _____

APPENDIX D: WRITING PROMPT 1

You are enrolled in a language course at an English-medium university. As part of the course requirements, your professor has asked you to maintain an online journal—a blog—in which you expand upon key topics presented in class. Your blog will be read by the professor, other students in the class, and even Internet users from around the world will be able to access your writing. As such, you may or may not want to name specific individuals or companies in your commentary. To complete this assignment, your blog post should respond to each of the questions found in the prompt, though you do not need to follow a specific format for organizing your ideas. Finally, to avoid plagiarism—and thereby, unnecessary criticism of your work—the professor has asked you not to utilize material directly from the Internet or other sources. You may access an electronic dictionary (either monolingual or bilingual) for words or specific terms, and you may incorporate hyperlinks should they be appropriate. There is no word limit; however, you cannot spend more than one and a half hours on your post, as you need to move on to assignments from other courses.

For your first blog entry, respond to the following questions:

> How is technology utilized today for language learning? In what ways is technology useful for learning languages? When do you think it is not useful? Why might technology be more useful for some areas of language learning than others? Give your opinion and support it with detailed reasons.

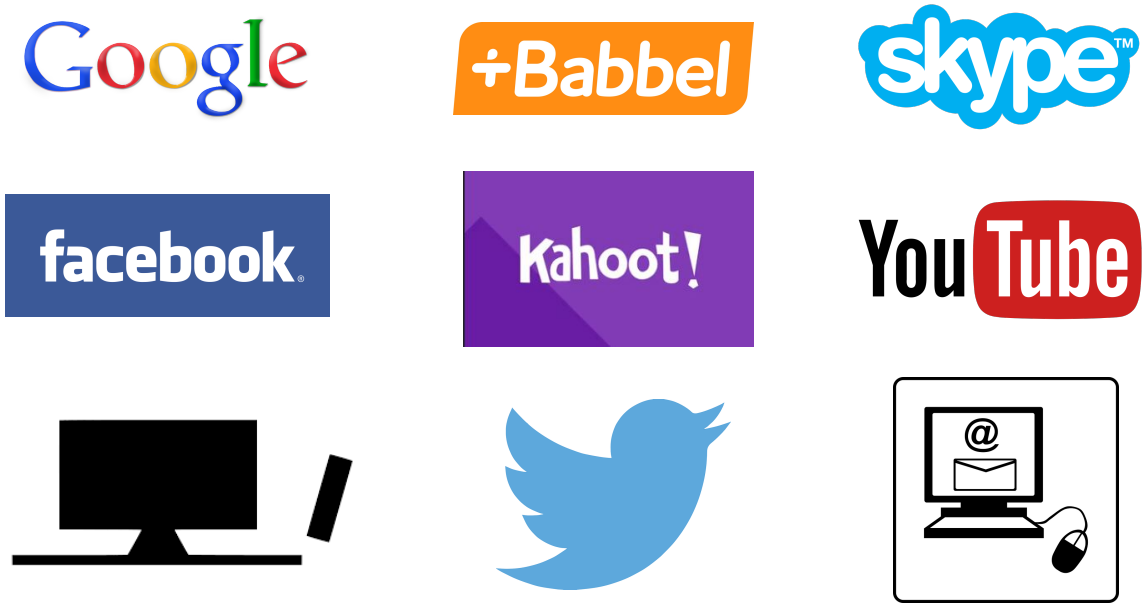


Tema de escritura 1

Estás matriculado en un curso de idiomas en una universidad de lengua inglesa. Como parte de los requisitos del curso, tu profesor te ha pedido que mantengas un diario en línea -un blog- en el que desarrolles los temas clave presentados en clase. Tu blog será leído por el profesor, otros estudiantes de la clase e incluso usuarios de Internet de todo el mundo podrán acceder a tu escrito. Como tal, puedes querer, o no, nombrar a personas o compañías específicas en tu comentario. Para completar esta tarea, tu entrada de blog debe responder a cada una de las preguntas que se encuentran en el tema, aunque no es necesario seguir un formato específico para organizar tus ideas. Por último, para evitar el plagio -y, por lo tanto, críticas innecesarias sobre tu trabajo-, el profesor te ha pedido que no utilices material directamente de Internet o de otras fuentes. Puedes acceder a un diccionario electrónico (ya sea monolingüe o bilingüe) para palabras o términos específicos, y puedes incorporar hipervínculos si son apropiados. No hay límite de palabras; sin embargo, no puedes pasar más de una hora y media en tu entrada, ya que necesitas continuar con las tareas de otros cursos.

Para tu primera entrada de blog, responde a las siguientes preguntas:

> ¿Cómo se utiliza la tecnología hoy en día para el aprendizaje de idiomas? ¿En qué formas la tecnología es útil para aprender idiomas? ¿Cuándo crees que no es útil? ¿Por qué la tecnología puede ser más útil para algunas áreas del aprendizaje de idiomas que para otras? Da tu opinión y apóyala con razones detalladas.



APPENDIX E: REVIEWER INSTRUCTIONS

Dear Reviewer,

Thank you for your willingness to review the blog posts I gathered for the first phase of my dissertation! I asked you to be a part of this project because you have a master's degree in Second Language Studies, Teaching English to Speakers of Other Languages (TESOL), English Education, or Applied Linguistics. You also have experience working with English language learners and with assessing second language writing.

Your role in this project is to contribute to the formation of a rubric for assessing an academic blog post, though we will be doing so on the basis of actual samples of English language learners' blog entries. Your task will involve reading the 148 short blog posts collected for the study, separating the entries into different folders, and commenting on each piece. I would ask that you complete the task in the order in which the instructions are presented below.

Instructions:

1. *Please read carefully the writing prompt given to the participants. You can locate the file, "Writing Prompt" on the shared Google Drive folder. You are free to reference the prompt often throughout the review process, as the instructions for the task may well factor into your assessment of the degree to which a participant accomplished the assignment's goals.*
2. *Familiarize yourself with the Excel spreadsheet upon which you will record data corresponding to each blog entry (see "Reviewer Marking Sheet" in the Google Drive Folder).*
3. *Start reading! The manner in which you choose to read through the blog posts is entirely up to you. You can start with Participant 1, or with Participant 148, or with any other participant. You also have the freedom to decide whether you want to read through all of the entries prior to separating them and making comments, or whether you would like to proceed one at a time.*
4. *Sort the entries into six folders in order of merit,* **placing the least successful entries in Folder 1** *and the* **most successful entries in Folder 6**. *These folders have already been created for you, and you will see them in the shared Google Drive Folder. There need not be an equal number of entries in each folder, and the only concrete rules are that* **every folder must be used** *and that* **at least four entries must appear in each folder**. *If the writing abilities of the participants were normally distributed, the number of papers in each folder, beginning with Folder 1, would be 4, 20, 50, 50, 20, and 4; however, there is no reason to expect a perfectly symmetrical distribution in this sample, as the English abilities of the participants were not normally distributed. In addition to placing the blog entries in distinct folders, please record the folder number on the Excel spreadsheet.*
5. *Record comments concerning the characteristics of each blog entry on the Excel spreadsheet. To my knowledge, there are no empirically derived rubrics describing the components of a successful academic blog post, and as such, I will refrain from suggesting particular categories upon which you ought to focus during reading. Rather, I would suggest that you draw on your experience with participating in online academic*

discussions (whether through a class discussion board or a wider academic audience) as well as your knowledge of the qualities valued in effective written communication. I recognize that you may have used a wide range of scoring rubrics in your teaching experience, and you are in no way restricted from recurring to those rubrics, if particular descriptors help you to identify the strong or weak components of different entries. At the same time, I would encourage you to refer continually to the guidance provided by the prompt and to your understanding of the nature of this new genre. There are no concrete rules regarding the length of the comments or the number of descriptors you should produce; however, targeted, short phrases will be easiest for me to parse! I would also humbly request that you attempt to avoid vague descriptors such as "good," "somewhat," or "adequate," in favor of comments that more clearly describe the nature of the piece under review. ☺

I have estimated that this process will take approximately five hours to complete, and you will be compensated $25 per hour. Should you need more time to complete the task, please contact me and we can discuss modifications to your compensation.

Finally, just as the identity of the participants is confidential to you, I will not reveal your identity nor any self-identifying characteristics, including your age, gender, and nationality, to the participants. Furthermore, the participants will not have access to the comments you make on their writing, and in any potential publications of the results of this research, you will be referred to only as "Reviewer" or "Reader" followed by a randomly assigned number. In acknowledgment of your willingness to participate in the role of reviewer, please sign the bottom of this page and scan a copy of the form with your signature into the shared Google Drive folder. I trust your judgment as to the quality of the texts you will review, and your participation is integral to the success of this project. Thank you!!

Sincerely,

Kristin Rock

_____

Printed Name

_____

Signature

APPENDIX F: ANALYTIC RUBRIC FOR AN ACADEMIC BLOG POST

User instructions: This rubric is intended as an instructional tool as well as for self- and/or instructor-assessment of an academic blog—or discussion board—post written in response to a particular set of questions. For each category in the left-hand column, there are a series of data-derived descriptors that correspond to a numerical value (1, 2, 3, 4, 5, or 6). The higher the number, the greater the quality of written work within this genre. Total scores will range from 5 to 30 points.

	1	2	3	4	5	6
Task Fulfillment & Relevancy	Does not address the questions from the prompt Text has no relevance to the prompt	Prompt addressed only minimally, with questions or certain nuances of questions skipped/missed Full paragraphs may not relate to the prompt	Addresses prompt questions in a cursory or glancing way, with one to two questions missed May contain some irrelevant information	Addresses all but one of the questions from the prompt. May contain some irrelevant points and information	Addresses all questions from the prompt A few details are not relevant or the connection between those details and the prompt is unclear	Addresses all questions from the prompt All details are relevant and directly connected to the main points (i.e., no superfluous information)
Content	Ideas are not developed No supporting details or examples provided Potentially a very short response	May contain good ideas, but only a few vague or general examples provided Text lacks solid arguments and detailed reasoning May be a short response	Provides some clear examples, though the ideas are not developed or explored fully One or two ideas may be loosely or partially supported	Includes important content, though the text lacks critical depth Concrete examples given, but argument could be strengthened with additional details	Contains many good points that are supported with personal examples Ideas are logical and specific, though text may be repetitive at times	Has numerous, thoughtful ideas that are supported with a variety of detailed examples The depth and development of evidence show strong grasp of content area

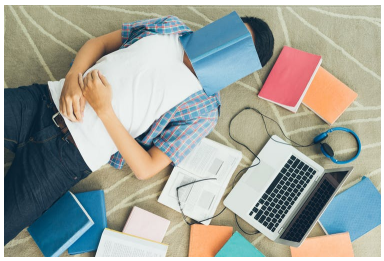
	1	2	3	4	5	6
Organization & Balance	Lacks organization Entire text could be a series of one-sentence paragraphs (list-like), or one, disjointed paragraph	Little discernible structure as writer toggles between points too many times. Ideas do not build upon each other well or progress clearly	Ideas separated into paragraphs, although within-paragraph organization may not always be clear Paragraphs are unbalanced, with points either over- or underdeveloped	Begins with strong introduction and flows well, but tapers off toward the end with rushed or abrupt conclusion Main ideas of some paragraphs may be unclear	Overall and within-paragraph organization is solid, and ideas progress logically Paragraphs are evenly weighted	Organization is clear & engaging; even explicitly stated in opening paragraph Fantastic intro and conclusion tie well-structured, balanced piece together
Genre Specific Features	Does not contain any recognizable genre characteristics (e.g., personal opinion, acknowledgment of readers, etc.)	Apart from use of personal experience or a less formal tone, piece does not contain any genre-specific characteristics, such as hyperlinks, direct questions, etc.	Acknowledges audience by directly addressing a group of readers **or** through the inclusion of rhetorical questions Contains either emojis, hyperlinks, **or** personal experience/opinion	Addresses a particular group of readers Makes use of two of the following: Emojis; hyperlinks; questions to encourage reader responses; personal experience/opinion	Addresses a particular group of readers Makes use of three of the following: Emojis; hyperlinks; questions to encourage reader responses; personal experience/opinion	A creative piece that exploits the interactive nature of the genre through well-placed emojis, hyperlinks, and personal content Encourages readers to respond via specific questions
Language Use	Usage errors hinder comprehensibility of most sentences Numerous spelling and grammar mistakes make the text difficult to read Barely controls simple syntactic structures Unspecified or undetectable voice	Several word choice errors detract from clarity of expression Many spelling and grammar mistakes give text a "rushed" quality Uses some simple structures correctly, but overall lacks syntactic variation Voice may be inconsistent	Word choice issues obfuscate meaning at times Some spelling and grammar errors present Succeeds with simple structures, but struggles to control complex syntax, as evidenced by run-on sentences, for example	Utilizes some sophisticated vocabulary, but may include the occasional awkward phrase A few sentences are difficult to parse due to spelling and grammar mistakes Some mixture of syntactic structures, though not error-free	Includes some appropriate idiomatic expressions and high-level vocabulary Contains a few grammar and spelling errors that do not impede comprehension Minimal difficulty with complex syntax	Appropriate, varied, and sophisticated word choice Nearly free of spelling and grammar errors A good mix of simple and complex syntactic structures Primarily employs active voice

APPENDIX G: WRITING PROMPT 2

You are enrolled in a language course at an English-medium university. As part of the course requirements, your professor has asked you to maintain an online journal—a blog—in which you expand upon key topics presented in class. Your blog will be read by the professor, other students in the class, and even Internet users from around the world will be able to access your writing. As such, you may or may not want to name specific individuals or companies in your commentary. To complete this assignment, your blog post should respond to each of the questions found in the prompt, though you do not need to follow a specific format for organizing your ideas. Finally, to avoid plagiarism—and thereby, unnecessary criticism of your work—the professor has asked you not to utilize material directly from the Internet or other sources. You may access an electronic dictionary (either monolingual or bilingual) for words or specific terms, and you may incorporate hyperlinks should they be appropriate. There is no word limit; however, you cannot spend more than one and a half hours on your post, as you need to move on to assignments from other courses.

For your next blog entry, respond to the following questions:

> How do teachers and professors use homework to help students learn? In what ways are homework assignments helpful for learning course material? When is homework not useful or even detrimental to student learning? How might the age of the student factor into a teacher's decision-making regarding homework? Give your opinion and support it with detailed reasons.

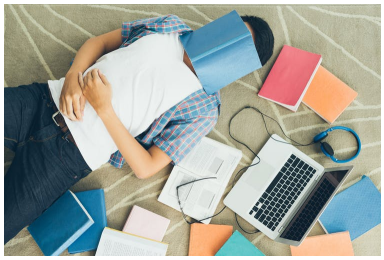


Tema de escritura 2

Estás matriculado en un curso de idiomas en una universidad de lengua inglesa. Como parte de los requisitos del curso, tu profesor te ha pedido que mantengas un diario en línea -un blog- en el que desarrolles los temas clave presentados en clase. Tu blog será leído por el profesor, otros estudiantes de la clase e incluso usuarios de Internet de todo el mundo podrán acceder a tu escrito. Como tal, puedes querer, o no, nombrar a personas o compañías específicas en tu comentario. Para completar esta tarea, tu entrada de blog debe responder a cada una de las preguntas que se encuentran en el tema, aunque no es necesario seguir un formato específico para organizar tus ideas. Por último, para evitar el plagio -y, por lo tanto, críticas innecesarias sobre tu trabajo-, el profesor te ha pedido que no utilices material directamente de Internet o de otras fuentes. Puedes acceder a un diccionario electrónico (ya sea monolingüe o bilingüe) para palabras o términos específicos, y puedes incorporar hipervínculos si son apropiados. No hay límite de palabras; sin embargo, no puedes pasar más de una hora y media en tu entrada, ya que necesitas continuar con las tareas de otros cursos.

Para tu próxima entrada de blog, responde a las siguientes preguntas:

> ¿Cómo utilizan los maestros y profesores los deberes para ayudar a que los estudiantes aprendan? ¿En qué formas los deberes son útiles para aprender el material del curso? ¿Cuándo crees que los deberes no son útiles, o aún perjudiciales para el aprendizaje del estudiante? ¿Cómo podría influir la edad del estudiante en el proceso del maestro de hacer decisiones sobre los deberes? Da tu opinión y apóyala con razones detalladas.

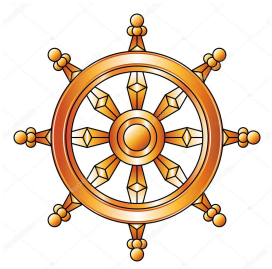


APPENDIX H: WRITING PROMPT 3

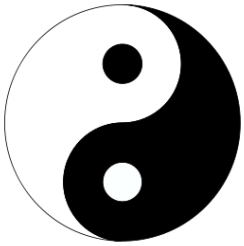
You are enrolled in a language course at an English-medium university. As part of the course requirements, your professor has asked you to maintain an online journal—a blog—in which you expand upon key topics presented in class. Your blog will be read by the professor, other students in the class, and even Internet users from around the world will be able to access your writing. As such, you may or may not want to name specific individuals or companies in your commentary. To complete this assignment, your blog post should respond to each of the questions found in the prompt, though you do not need to follow a specific format for organizing your ideas. Finally, to avoid plagiarism—and thereby, unnecessary criticism of your work—the professor has asked you not to utilize material directly from the Internet or other sources. You may access an electronic dictionary (either monolingual or bilingual) for words or specific terms, and you may incorporate hyperlinks should they be appropriate. There is no word limit; however, you cannot spend more than one and a half hours on your post, as you need to move on to assignments from other courses.

For your next blog entry, respond to the following questions:

> How is religion incorporated in public and/or private schools in Spain? In what ways is school instruction on religion constructive for student development? When is religious instruction not supportive of, or even detrimental to, a young person's education? What can students learn from different religions? Give your opinion and support it with detailed reasons.





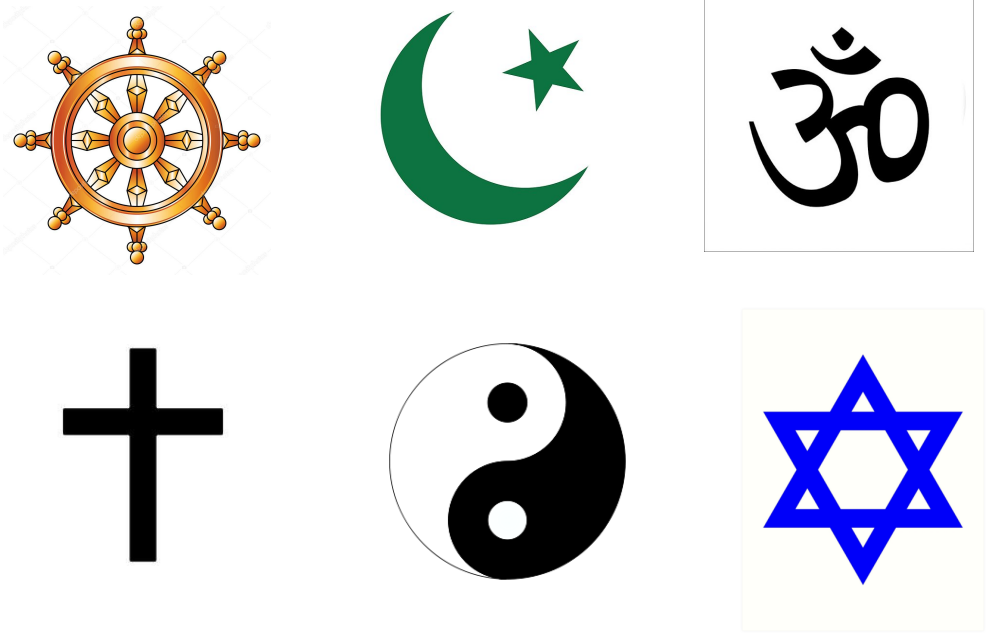



Tema de escritura 3

Estás matriculado en un curso de idiomas en una universidad de lengua inglesa. Como parte de los requisitos del curso, tu profesor te ha pedido que mantengas un diario en línea -un blog- en el que desarrolles los temas clave presentados en clase. Tu blog será leído por el profesor, otros estudiantes de la clase e incluso usuarios de Internet de todo el mundo podrán acceder a tu escrito. Como tal, puedes querer, o no, nombrar a personas o compañías específicas en tu comentario. Para completar esta tarea, tu entrada de blog debe responder a cada una de las preguntas que se encuentran en el tema, aunque no es necesario seguir un formato específico para organizar tus ideas. Por último, para evitar el plagio -y, por lo tanto, críticas innecesarias sobre tu trabajo-, el profesor te ha pedido que no utilices material directamente de Internet o de otras fuentes. Puedes acceder a un diccionario electrónico (ya sea monolingüe o bilingüe) para palabras o términos específicos, y puedes incorporar hipervínculos si son apropiados. No hay límite de palabras; sin embargo, no puedes pasar más de una hora y media en tu entrada, ya que necesitas continuar con las tareas de otros cursos.

Para tu próxima entrada de blog, responde a las siguientes preguntas:

> ¿Cómo se incorpora la religión en las escuelas públicas y/o privadas en España? ¿En qué formas la instrucción escolástica en la religión es constructiva para el desarrollo del estudiante? ¿Cuándo crees que la enseñanza religiosa no apoya, o aún perjudica, la educación de los jóvenes? ¿Qué pueden aprender los estudiantes de otras religiones? Da tu opinión y apóyala con razones detalladas.



APPENDIX I: BLOG POSTS REPRESENTING LEVELS 1 THROUGH 6

LEVEL 1 (Participant 85)

TECHNOLOGY & LANGUAGE

Nowadays technology is involved in everywhere.

On one hand, as a student, I´ve been learning languages with technologies. I only have studied English and French. In both lessons teachers always put songs in that language and series or movies to practise the language. From my point of view this is so good, because we also learn the different accents. In addition it isn´t bored. So yes! Of course it is usefull.

On the other hand as a future teacher, as the world is expanding technology, technology has to come with us for everything, also to teach languages.

Finally, say that technology is better to improve listening and speaking because we have more facilities for that. We do not need too expensive technology or something loke that.

LEVEL 2 (Participant 89)

TECHNOLOGY AS FUNNY LEARNING

Good afternoon,

Have you ever thought in the importance of technology nowadays? If not, you can reflect on that if you look around, by doing that you will discover that it is surronding us. In the 21st century technology takes part of our daily life, so...Would it be a great idea to include it in schools? Personally, I think so.

From my point of view we have to adapt the school environment to the context in which we are living in; and, as outside school we use it for too many things, why shouldn't we include it in our lessons?

As you would have noticed, I am in favour of using technology in classrooms, but not just because we live with it, but also because I consider that it has many utilities such as we have the opportunity of learning languages through it, for example, by watching videos (tutorials, lessons or conversations) that teach us some rules and/or vocabulary, playing educational and/or interactive games, doing skype with people from other places that speak in other language which is different to ours, listening educational songs or movies of another language...). As it has many advantages it is very useful for learning languages in a more dynamic and funny way. In my opinion, if we use technology as a tool for situations like projecting the book on the digital screen, it is not useful as we are not using it for doing innovations.

LEVEL 3 (Participant 96)
Learning Language Technology and Its importance in our lifes

Technology is one of the main resources which students got for improving their language skills. Also for teachers it is an important tool which they use it for create lessons and other activities to reinforce their students abilities. Social media, websides and other electronic devices that allow you to surf into internet, help you to be updated and to increase your language knowledge. As we can see technology has a lot of importance in different aspects.

When we have the necessity to improve our language skills, people normally resort to textbooks, and change the language of the films. This measures are very effective but also technology can help a lot when we are talking about improving our language skills. An amoung of applications are avaliable for the user in order to put into practise their lingüistic abilities, they offer you a sort of exercises which are very useful for perfecting your skills and to see what are your mistakes to correct them. In the case of the teachers, I personally remember my prescholar teacher that used to take a retro-proyector to show us some images with the name in english at the botton of the sheet. However, nowadays technology has evolved and now language learning apps are adapted to the level of the students and they give you more chances to learn.

Due to the fact that people want to travel around the world to discover new cultures and to visit different countries, technology give them enough means to study a foreign language in a fast and effective way.

Moreover, not everyone travel abroad for having some vacations, there are people that have the necessity to move to another state with the aim to find a job because in their native country they do not got the economic resources to obtain a work position.

In addition, the reason for learning languages it is reduce to the simple fact of increasing the knowledge of a person: for instance some retired people are learning languages because it is a healthy activity for the brain and it is a way of entertaining.

On the other hand, learning languages through technology could be not a good implement for communicating with people, as it is not the same to talk face to face with a person that talking through a microphone and a camera. Having a real life conversation allows you to express in a better way as well as to understand better a person because you can see his/her expression on the face and to analyze his/her body language properly.

Even though learning a new language could be useful in some university degrees such as the one that are related to social sciences or to computer programming. In general these career where you receive information from other international sources, it is compulsory to know english or japanese as most of the information you gain is in these languages and are the ones that domain the world of technology. Nevertheless it is not a crucial skill to know another language if you work with people from your native country or your job is based on repetitive processes.

LEVEL 4 (Participant 24)

It is a fact that the use of technology in language learning is becoming more and more frequent as a consequence of the increment in the use of mobile devices, which allows us to access to almost any information not only in an inexpensive and fast way, but also at any time and in any place.

The vast majority of people owns a smartphone or a tablet that can incorporate a great variety of apps, such as 'Kahoot' and 'Babble', which can be really useful when trying to acquire a second or third language. In the same way, other electronic applications such as 'Skype', 'Facebook' and 'Twitter' offer us the possibility of being in contact with other people or students from different nationalities with whom we can communicate by means of the language we are trying to learn.

As it is my case, and I hope yours too, when I have to wait sitting in a bench (no matter the reason why), I usually spend some time on Babble, trying to refresh my rusty French, or even on a YouTube channel, in an attempt to learn some English collocations or idioms.

Thus, regarding language learning from my own experience, the use of technology is always useful, even in class from time to time when the teacher attempts to make the lesson more entertaining by fostering the competitivity among students (e.g. 'Kahoot'). Nonetheless, electronic devices can have some drawbacks, such as the possibility of distraction when using other social networks, the possible sight damage and even the filtration of fake information.

I thereby consider that technology is far more useful when learning grammar and vocabulary rather than for the attainment of a higher level of fluency or even sometimes listening comprehension. Although it might not be the case of everyone.

To conclude, I consider the use of technology in language learning to be a paramount factor that nowadays facilitates the way students learn a language, as they can access easily to any information they desire, even to that information they were not aware before. Although the improvement may not be directly reflected in all the aspects of the language, it is still really helpful.

LEVEL 5 (Participant 35)

The use of technology in language learning

Technology may be used in many ways for language learning. In fact, it it a very useful tool that has been bening used for a long time. Having said that, we will now examine how they are being used nowadays related to each skill.

The most clear example is the use of technology with the aim of practising and testing the listening kill of our students. We all have the image of the foreign language teacher with her CD player around the high school. Moreover, listeing to music in your foreing language can also be a great deal of help in order to improve our student's vocabulary and pronunciation.

Nevertheless, the use of technology is increasing in other skills too. For instance, in reading, students are progressively getting used to being presented the texts in the screen of a laptop or a tablet, or even their phones. Usually, after reading the text they are to complete some question referring to it, regadless the question type (fill in the gaps, chose the odd one out, multiple choice questions among other types). The main purpose of this questions is to check the degree of their understanding of the text. This has usually been done in paper with the same format so it does not make a big difference.

Speaking is another skill where technologies have played an important role. Previously, the contact with native speakers of the language was very limited since it was difficult to reach one (except if the teacher was lucky to find someone living near who was willing to help). Now, there are thousands of platforms that through videocalls and chats (writing is also involved here) facilitate pretty much the contact with native speakers, foreing language teachers or ever students from another part of the world to share experiences in the foreing language about its learning. Form my point of view, this increases very much the degree of motivation of the students since it is almost like real life contact.

Finally, technology use in writing is also a crucial issue in language learning. For me, the most useful resource the Internet has provide to students is the use of online dictionaries that save a lot of the student's time since they do not have to search words in a very fat book full of them. It is very much easier and quicker to type the word and get the information you are looking for in less than a second. Cheking long texts in a electronic device can be dangerous for the teacher's eye and as a consequence, most of the teachers prefer to have them printed or handwritten. However it is true that a text written in a electronic device is easier to read because the proble of (un)integibillity.

The Internet is also very helpful as a resource (for both, teachers and students) where they can find a myriad books and exercises to learn vocabulary and particular grammatical structures. Also, it is full of authentic materials that the teacher decides to adapt more or less to his/her students's level. Every student can find a youtuber o a blogger that deals with topics of their interest because of the diverse variety of people who post daily, weekly or mothly in their social networks. This fact turns foreing language learning into almost a second language learning.

Moving now to the point on why technology is not useful, it is important to consider the harm that screen can do to our eyes. This question should not be taken as a joke. Also, in learnign grammar for instance, it does not make a great difference (as it happens with reading) since the exercises that technologies can offer do not differ in a significant degree form the tradicitional ones. An argument in favour could be the automatic check of an answer or at the end of an exercise.

Our final remark about the use of technology in language is that in general, it is very helpful but it is not totally essential. The reasonn behind this is that all life languages have been taught without technology.

LEVEL 6 (Participant 49)

IS TECHNOLOGY ENOUGH FOR FOREIGN LANGUAGE LEARNING?

Long gone are the days when the only access to foreign languages in traditionally monolingual c0untries such as France, Spain or the US was provided by foreign language teachers. These devoted professionals, non-native themselves in most of the cases, travelled many miles just to record a few minutes of natural conversation between native speakers for later use in their language classes. Though infamous among the students for its close relationship with listening tests, this material was virtually their only contact point with real language in use.

Fortunately, the arrival of new technologies, and in particular of internet, has relieved the foreign language teacher from this relentless responsibility. Nowadays, hundreds of commercial and educational digital tools are accessible by any interested learner who decides to embark on the enriching journey of learning a foreign language. Let's take a language learning app such as Duolingo, which has reached the outstanding figure of 200 million users worlwide, to exemplify the extent of their impact. In addition, the complete success of international TV series (who hasn't felt like an outcast after having missed the last eposide of GoT?), together with the interest of young generations in watching audiovisual material in original version have provided the population with a notable increase in exposure to native English (just compare with the teacher situation right above!).

Let us assume for now that new technologies provide language learners with enough (or at least more) input in the foreign language. Nevertheless, concluding that exposure alone will guarantee language learning might be hasty, at least if we look at the evidence gathered in the second language acquisition (SLA) research field. In this line, multiple studies support the idea that the opportunity for output (i.e. the opportunity to use the language themselves) might be necessary to learn a language appropriately. In addition, feedback on performance might be essential to acquire high levels of accuracy. To follow up with the examples above, neither Duolingo nor watching a TV series in English provides the learner with opportunities to use the language him or herself. In this situation, we might find too many foreign language speakers with notable receptive (reading and listening) skills, but completely unable to produce (speak and write).

Where do we go from here? This question has no easy answer. What we can conclude is that, although there is no discussion that technology increases exposure, language learning is much more than that. Provided that the language they are learning is not widely spoken in their country of residence, the only opportunity that foreign language learners might have to produce sufficient output could be to interact with native speakers or with other learners in a controlled environment (e.g. a classroom). In the end, the figure of the teacher, although with relieved responsibilities, might be essential to achieve an adequate competence in a foreign language.

APPENDIX J: DISTRIBUTION OF AVERAGE FOLDER SCORES

Distribution of Average Folder Scores

Average Score	Frequency	Percent	Cumulative Percent
1.00	1	0.7	0.7
1.17	1	0.7	1.4
1.33	1	0.7	2.0
1.50	1	0.7	2.7
1.67	5	3.4	6.1
2.00	1	0.7	6.8
2.17	4	2.7	9.5
2.33	5	3.4	12.8
2.50	11	7.4	20.3
2.67	5	3.4	23.6
2.83	7	4.7	28.4
3.00	8	5.4	33.8
3.17	5	3.4	37.2
3.33	9	6.1	43.2
3.50	8	5.4	48.6
3.67	9	6.1	54.7
3.83	15	10.1	64.9
4.00	9	6.1	70.9
4.17	3	2.0	73.0
4.33	14	9.5	82.4
4.50	8	5.4	87.8
4.67	1	0.7	88.5
4.83	5	3.4	91.9
5.00	6	4.1	95.9
5.17	3	2.0	98.0
5.33	1	0.7	98.6
5.50	2	1.4	100.0
Total	148	100.0	

APPENDIX K: COMMAND AND DATA FILE FOR FACETS

```
; RatingsFirstBlogPost.txt
title = Ratings on First Blog Post
convergence = 0.1       ; size of largest remaining marginal score residual at convergence
unexpected = 3.0        ; size of smallest standardized residual to report
arrange = M     ; arrange output tables in Measure ascending order
facets = 3      ; there are 3 facets in this analysis
noncenter = 1   ; examinee facet floats
positive = 1    ; for examinees, greater score   greater measure
Inter-rater = 2 ; facet 2 is the rater facet
pt-biserial=measure ; point-measure correlation
usort = 1,3,2   ; sort residuals by 1=Participant, 3=Rater, 2=Category
Model=
?,?B,#B,R6      ; observations are ratings in range 1-6.
                ; look for interaction/bias between rater and category
?B,?B,?,R6      ; look for examinee x rater interaction
*

anchorfile   = essayspostsa.txt ; measure file for subsidiary analyses
graphfile    = essayspostsg.txt ; values for plotting curves (better to use Graphs menu)
residualfile = essayspostsr.txt ; observations and their residuals
scorefiles   = essayspostss.txt ; the score files for each facet: essayssco1.txt, etc.

Labels=
1,examinee
 001-163        ; 163 otherwise anonymous examinees
*
2,Rater
 1-6     ; 6 otherwise anonymous raters
*
3,categories
 1= TaskFulfillment    ; 5 categories
 2= Content
 3= OrganizationBalance
 4= GenreSpecific
 5= LanguageUse
*
dvalues = 3, 1-5
data =

001,1,6,6,5,3,5
002,1,3,4,5,3,5
003,1,3,3,4,3,5
004,1,6,4,5,3,5
```

005,1,3,4,4,3,5
006,1,3,4,4,1,4
007,1,5,4,5,1,6
008,1,6,3,4,2,4
009,1,6,5,4,2,6
010,1,5,4,5,1,4
011,1,2,3,2,2,3
012,1,5,5,5,2,3
013,1,4,4,3,1,3
014,1,2,2,3,2,4
015,1,5,4,5,2,4
016,1,3,2,3,2,5
017,1,3,4,3,2,3
018,1,5,4,5,2,5
019,1,3,3,3,2,4
020,1,5,3,3,2,5
021,1,4,4,5,1,6
022,1,5,5,6,4,6
023,1,5,2,3,3,3
024,1,4,4,3,3,4
025,1,3,2,2,1,4
026,1,1,1,1,1,3
027,1,3,1,3,1,3
028,1,5,3,3,2,3
029,1,2,1,3,2,3
030,1,2,2,3,1,3
031,1,3,2,2,2,3
032,1,2,2,2,2,3
033,1,4,3,4,2,4
034,1,2,2,3,2,5
035,1,5,4,5,1,4
036,1,6,4,5,1,5
037,1,5,4,3,1,5
038,1,4,4,2,1,5
039,1,4,4,4,2,4
040,1,4,4,3,1,5
041,1,2,3,3,1,4
042,1,4,5,5,2,4
043,1,4,4,5,1,4
044,1,3,4,4,1,3
045,1,4,4,3,1,3
046,1,5,5,5,2,5
047,1,5,4,5,2,5
048,1,3,3,2,2,3
049,1,3,3,3,2,3
050,1,3,2,5,1,6

051,1,4,4,5,3,5
052,1,5,5,5,3,6
053,1,3,4,3,1,2
054,1,6,6,5,5,5
055,1,6,5,6,3,5
056,1,6,4,5,2,3
057,1,4,3,4,3,4
058,1,6,6,6,2,6
059,1,6,6,5,3,6
060,1,6,5,6,2,5
061,1,6,5,5,1,6
062,1,4,4,4,2,5
063,1,5,4,5,1,5
064,1,5,4,5,1,5
065,1,4,4,4,3,5
066,1,4,4,3,2,3
067,1,5,4,3,3,3
068,1,5,5,5,3,4
069,1,5,4,3,1,4
070,1,4,4,4,2,3
071,1,3,3,3,1,3
072,1,6,5,5,2,4
073,1,5,4,5,2,3
074,1,4,3,4,2,3
075,1,3,3,3,3,3
076,1,3,3,3,1,2
077,1,4,3,3,1,3
078,1,5,4,3,2,3
079,1,5,4,5,3,4
080,1,3,3,3,3,3
081,1,5,2,3,3,3
082,1,5,2,3,1,2
083,1,5,5,5,2,4
084,1,5,4,5,1,4
085,1,4,2,3,1,4
086,1,5,3,3,1,3
087,1,5,2,3,1,2
088,1,5,4,5,1,4
089,1,4,3,3,3,4
090,1,4,2,3,3,3
091,1,4,2,3,1,3
092,1,4,2,4,3,3
093,1,3,4,3,1,3
094,1,2,2,3,1,3
095,1,3,3,3,3,3
096,1,2,2,3,1,3

097,1,2,2,4,3,3
098,1,2,2,2,1,2
099,1,4,3,4,3,3
100,1,2,1,3,1,2
101,1,5,2,3,3,3
102,1,2,3,3,1,3
103,1,2,2,3,1,3
104,1,1,1,3,1,3
105,1,2,1,3,1,2
106,1,2,1,3,1,2
107,1,3,3,3,1,3
108,1,2,2,1,1,3
109,1,3,3,2,3,3
110,1,5,4,5,1,4
111,1,4,4,5,1,4
112,1,5,4,4,1,4
113,1,3,4,3,1,3
114,1,3,3,3,1,3
115,1,5,4,3,1,3
116,1,3,4,3,1,3
117,1,2,3,3,1,3
118,1,5,4,3,3,3
119,1,5,4,5,1,3
120,1,5,4,3,2,3
121,1,4,3,3,1,3
122,1,2,2,3,1,3
123,1,2,3,3,1,3
124,1,2,1,3,1,3
125,1,3,2,3,1,2
126,1,3,3,3,1,2
127,1,4,3,3,1,3
128,1,5,4,3,1,3
129,1,4,3,3,1,3
130,1,3,4,3,1,3
131,1,4,3,1,1,3
132,1,3,3,3,1,3
133,1,3,3,3,3,3
134,1,2,1,3,3,3
135,1,2,1,3,1,3
136,1,2,2,3,1,3
137,1,3,2,3,1,3
138,1,3,3,3,1,3
139,1,3,2,3,3,3
140,1,2,3,3,1,3
141,1,2,3,3,1,3
142,1,2,2,3,1,3

143,1,2,2,3,3,3
144,1,3,2,3,1,3
145,1,4,2,3,3,2
146,1,4,3,3,1,3
147,1,4,3,3,1,3
148,1,2,4,3,1,3
149,1,2,3,3,1,4
150,1,6,5,5,3,4
151,1,5,4,3,2,3
152,1,3,2,3,2,2
153,1,6,6,6,2,6
154,1,5,4,3,1,3
155,1,6,4,4,1,3
156,1,5,4,5,3,4
157,1,6,5,5,3,6
158,1,2,2,3,2,2
159,1,5,5,5,3,4
160,1,6,4,5,4,4
161,1,4,3,3,3,4
162,1,3,2,3,1,3
163,1,5,3,5,3,3
001,2,5,4,6,1,4
002,2,3,2,3,2,1
003,2,5,3,4,2,4
004,2,4,3,3,2,4
005,2,4,2,3,2,2
006,2,3,2,3,1,4
007,2,4,4,2,1,4
008,2,4,3,3,2,3
009,2,5,3,3,2,4
010,2,5,5,4,1,5
011,2,3,2,2,1,2
012,2,5,6,5,2,5
013,2,4,4,4,2,3
014,2,2,2,3,1,2
015,2,5,4,3,2,4
016,2,3,2,1,1,3
017,2,3,3,3,2,3
018,2,4,5,4,1,5
019,2,4,3,3,3,3
020,2,2,4,3,2,4
021,2,4,5,3,1,3
022,2,3,4,4,4,4
023,2,4,3,3,3,3
024,2,3,3,3,3,3
025,2,4,3,2,2,4

026,2,3,1,3,1,1
027,2,3,3,3,1,3
028,2,3,2,3,2,2
029,2,3,2,3,3,2
030,2,3,3,3,1,3
031,2,5,4,4,2,4
032,2,4,2,3,2,4
033,2,5,4,4,1,3
034,2,3,3,3,2,3
035,2,4,4,4,2,3
036,2,5,3,4,1,4
037,2,4,5,4,2,5
038,2,2,2,2,1,2
039,2,4,4,2,1,3
040,2,3,3,3,1,3
041,2,4,3,2,1,3
042,2,5,5,5,2,5
043,2,4,4,4,1,4
044,2,5,4,5,1,5
045,2,3,4,4,1,3
046,2,4,4,4,1,4
047,2,3,4,4,2,4
048,2,4,3,1,2,2
049,2,4,5,4,2,4
050,2,2,2,3,1,4
051,2,4,4,5,3,5
052,2,4,5,5,4,5
053,2,3,2,3,1,2
054,2,4,4,4,5,5
055,2,5,5,5,5,5
056,2,4,3,4,2,4
057,2,2,2,2,1,2
058,2,4,4,4,2,3
059,2,4,4,4,2,4
060,2,4,4,4,1,4
061,2,4,5,5,1,4
062,2,2,2,3,2,3
063,2,5,4,4,2,3
064,2,4,3,4,2,3
065,2,4,3,2,3,2
066,2,3,3,3,3,3
067,2,5,4,4,3,3
068,2,5,4,4,4,4
069,2,5,4,4,2,3
070,2,3,3,4,1,2
071,2,5,4,3,1,3

072,2,5,4,3,1,3
073,2,4,4,4,2,3
074,2,4,2,3,2,3
075,2,4,2,3,3,2
076,2,3,2,3,2,3
077,2,3,3,3,2,3
078,2,4,3,3,2,3
079,2,4,3,3,2,3
080,2,3,2,3,3,3
081,2,3,3,2,2,2
082,2,4,3,3,1,3
083,2,3,3,3,2,3
084,2,5,4,4,2,4
085,2,1,1,2,2,1
086,2,5,4,3,2,3
087,2,4,3,3,2,4
088,2,5,4,4,2,4
089,2,3,4,3,4,4
090,2,3,2,3,3,3
091,2,3,3,4,2,4
092,2,3,3,3,3,3
093,2,4,3,4,2,3
094,2,3,3,3,2,3
095,2,4,4,4,4,3
096,2,3,2,2,2,3
097,2,4,3,4,3,4
098,2,3,3,3,2,3
099,2,5,5,5,5,5
100,2,4,3,2,2,3
101,2,5,3,4,5,3
102,2,2,2,2,2,2
103,2,4,3,4,2,4
104,2,3,3,3,2,3
105,2,4,4,4,1,3
106,2,3,3,3,1,2
107,2,4,3,3,2,3
108,2,4,3,2,2,3
109,2,4,3,3,3,3
110,2,5,4,4,1,4
111,2,4,4,3,2,3
112,2,3,4,3,1,3
113,2,4,4,3,2,3
114,2,3,3,3,2,3
115,2,4,3,4,1,3
116,2,3,3,3,2,3
117,2,3,3,4,2,3

118,2,2,2,3,3,3
119,2,4,4,4,1,2
120,2,3,2,2,2,2
121,2,4,3,3,2,3
122,2,2,3,3,2,3
123,2,3,3,4,3,3
124,2,3,3,3,2,2
125,2,3,3,3,2,3
126,2,3,4,3,3,3
127,2,2,3,3,2,2
128,2,4,4,3,2,3
129,2,4,3,4,1,3
130,2,3,3,2,2,3
131,2,2,3,2,1,3
132,2,3,3,3,2,3
133,2,3,3,4,3,2
134,2,3,3,4,3,3
135,2,3,2,3,2,2
136,2,4,3,3,2,3
137,2,4,3,3,2,3
138,2,4,3,3,2,3
139,2,3,3,3,3,3
140,2,3,3,4,3,4
141,2,3,3,2,2,3
142,2,3,3,3,3,3
143,2,3,3,4,3,3
144,2,5,4,4,2,4
145,2,4,4,3,3,2
146,2,3,3,3,2,3
147,2,3,3,4,2,3
148,2,4,4,4,3,4
001,3,6,5,5,3,6
002,3,3,3,3,2,4
003,3,6,4,3,2,5
004,3,3,4,2,1,5
005,3,3,2,2,4,3
006,3,5,4,4,1,5
007,3,6,5,4,1,5
008,3,5,3,3,2,3
009,3,4,5,4,2,5
010,3,6,6,3,1,6
011,3,2,2,1,1,5
012,3,4,6,5,2,5
013,3,5,4,4,2,4
014,3,3,2,2,1,2
015,3,6,6,6,2,6

016,3,5,2,1,2,4
017,3,4,5,5,2,6
018,3,5,5,5,1,6
019,3,5,4,4,3,4
020,3,4,4,4,2,6
021,3,6,6,6,2,6
022,3,4,5,5,6,4
023,3,5,3,3,3,3
024,3,5,2,2,2,4
025,3,5,2,1,2,5
026,3,2,2,2,1,3
027,3,5,3,2,2,3
028,3,5,5,5,2,5
029,3,3,2,3,3,3
030,3,4,4,5,2,4
031,3,5,4,4,2,4
032,3,5,5,5,1,4
033,3,4,6,5,1,6
034,3,4,3,4,2,4
035,3,6,6,6,2,5
036,3,6,5,5,1,4
037,3,4,6,5,2,6
038,3,3,2,2,1,2
039,3,5,4,2,2,4
040,3,3,3,4,2,4
041,3,5,4,3,1,5
042,3,6,6,5,2,6
043,3,6,6,5,3,5
044,3,6,5,5,2,5
045,3,3,4,4,2,5
046,3,5,4,4,2,4
047,3,5,4,3,3,4
048,3,4,3,2,3,5
049,3,6,6,6,3,6
050,3,2,2,2,2,5
051,3,4,4,4,3,5
052,3,4,5,4,3,5
053,3,4,2,3,1,3
054,3,4,5,5,6,5
055,3,5,5,5,5,6
056,3,5,4,4,2,4
057,3,3,2,4,3,4
058,3,4,5,5,2,5
059,3,4,4,4,3,4
060,3,4,4,5,2,4
061,3,4,4,4,1,5

062,3,2,2,2,2,4
063,3,3,4,4,2,5
064,3,4,4,4,3,4
065,3,5,3,3,4,3
066,3,3,2,1,3,3
067,3,4,3,5,3,4
068,3,5,4,3,3,4
069,3,3,5,4,3,4
070,3,5,5,5,2,4
071,3,3,3,4,1,3
072,3,5,5,4,2,4
073,3,5,3,3,2,5
074,3,3,2,2,2,2
075,3,4,2,2,3,2
076,3,5,3,3,1,3
077,3,5,3,4,2,3
078,3,5,4,4,2,3
079,3,5,4,5,3,4
080,3,4,2,2,3,3
081,3,3,2,3,3,3
082,3,4,2,3,2,3
083,3,3,3,2,2,3
084,3,5,3,3,2,3
085,3,2,2,1,1,2
086,3,5,3,3,2,2
087,3,4,2,3,2,3
088,3,5,5,5,2,4
089,3,3,4,3,4,3
090,3,3,3,3,3,2
091,3,4,3,2,2,3
092,3,4,2,2,3,2
093,3,4,3,3,3,3
094,3,3,3,1,2,4
095,3,5,4,3,3,4
096,3,3,3,3,2,3
097,3,4,3,5,3,3
098,3,2,2,3,2,3
099,3,5,4,4,4,4
100,3,4,3,4,2,3
101,3,6,3,5,5,3
102,3,3,3,4,2,2
103,3,4,4,4,2,4
104,3,4,2,2,2,2
105,3,5,4,4,1,5
106,3,3,4,4,2,3
107,3,4,3,3,2,3

108,3,4,2,2,3,3
109,3,4,3,2,4,3
110,3,4,3,4,1,5
111,3,4,4,4,3,3
112,3,4,3,4,1,3
113,3,5,4,4,2,4
114,3,2,3,3,2,3
115,3,5,5,5,1,3
116,3,4,5,5,3,4
117,3,4,5,4,2,3
118,3,4,4,1,5,5
119,3,5,4,4,2,3
120,3,4,3,2,3,5
121,3,4,4,2,2,3
122,3,3,2,3,2,4
123,3,3,2,3,3,4
124,3,3,3,4,2,4
125,3,3,3,3,2,3
126,3,4,5,5,3,4
127,3,3,3,4,1,4
128,3,4,3,3,2,3
129,3,4,3,4,2,3
130,3,5,4,2,2,4
131,3,3,2,1,2,4
132,3,3,2,2,2,3
133,3,5,5,5,3,5
134,3,4,5,4,4,5
135,3,4,4,4,2,5
136,3,5,4,4,2,5
137,3,5,4,4,2,3
138,3,4,3,4,2,5
139,3,4,3,4,3,3
140,3,3,5,4,3,4
141,3,4,3,2,3,3
142,3,3,3,3,4,3
143,3,4,4,5,4,4
144,3,5,4,4,2,3
145,3,4,4,4,4,3
146,3,5,4,4,2,3
147,3,4,4,4,2,4
148,3,6,6,5,3,6
149,3,5,3,3,2,6
150,3,6,6,4,2,3
151,3,5,6,4,1,5
152,3,4,2,3,1,3
153,3,6,6,3,2,6

154,3,6,5,4,1,3
155,3,4,5,5,1,6
156,3,6,5,6,3,3
157,3,6,6,6,2,6
158,3,4,5,4,3,3
159,3,4,4,4,3,6
160,3,4,3,3,1,3
161,3,4,2,2,2,3
162,3,4,3,4,2,4
163,3,5,5,5,2,3
001,4,6,6,6,6,6
002,4,4,3,2,4,2
003,4,6,6,6,6,6
004,4,5,5,4,5,5
005,4,5,5,5,5,4
006,4,5,5,4,5,5
007,4,6,5,6,6,5
008,4,4,4,3,6,2
009,4,4,5,5,6,4
010,4,6,6,6,6,5
011,4,1,2,2,3,3
012,4,5,4,4,5,5
013,4,5,5,4,6,5
014,4,1,2,2,4,3
015,4,6,6,6,6,5
016,4,2,2,1,3,3
017,4,5,5,6,5,5
018,4,6,6,4,5,5
019,4,5,5,4,6,5
020,4,6,5,5,6,5
021,4,6,6,6,6,6
022,4,6,6,6,6,3
023,4,6,5,5,5,4
024,4,6,6,5,4,5
025,4,4,4,1,3,5
026,4,5,4,4,4,4
027,4,3,4,3,4,5
028,4,4,4,4,5,5
029,4,3,3,3,2,2
030,4,5,5,5,5,5
031,4,6,6,5,5,5
032,4,6,6,5,5,5
033,4,6,6,6,6,5
034,4,6,6,6,6,6
035,4,6,6,6,6,5
036,4,6,6,6,6,5

037,4,6,6,6,6,6
038,4,2,2,2,4,5
039,4,5,4,2,3,5
040,4,5,5,4,5,5
041,4,6,6,4,5,6
042,4,6,6,5,5,5
043,4,6,6,6,6,5
044,4,6,6,6,6,6
045,4,6,6,5,6,5
046,4,5,4,4,3,4
047,4,6,6,6,6,6
048,4,5,5,2,3,5
049,4,6,6,6,6,6
050,4,5,5,3,3,5
051,4,6,6,6,6,6
052,4,6,6,6,6,6
053,4,4,3,3,4,5
054,4,5,5,5,5,5
055,4,6,6,6,6,5
056,4,5,5,5,6,4
057,4,2,2,2,2,2
058,4,4,4,4,4,4
059,4,6,5,5,6,4
060,4,6,6,6,6,6
061,4,6,6,6,6,5
062,4,3,3,3,4,4
063,4,6,5,5,5,3
064,4,5,5,6,4,2
065,4,4,3,2,6,3
066,4,4,3,3,5,3
067,4,4,4,3,5,3
068,4,6,5,5,6,4
069,4,6,6,6,6,6
070,4,6,5,6,6,5
071,4,6,5,6,5,4
072,4,5,4,6,5,4
073,4,5,4,6,4,4
074,4,5,3,2,3,3
075,4,2,2,2,3,1
076,4,3,2,3,3,2
077,4,5,2,5,3,3
078,4,3,3,3,4,3
079,4,6,5,5,5,5
080,4,5,4,5,5,4
081,4,5,3,5,4,3
082,4,5,3,4,3,2

083,4,6,5,5,5,3
084,4,6,6,6,6,5
085,4,4,2,2,2,2
086,4,4,3,3,4,5
087,4,2,2,2,2,3
088,4,6,6,6,6,4
089,4,5,5,4,5,6
090,4,5,3,3,5,2
091,4,6,5,6,5,4
092,4,5,4,5,4,3
093,4,5,3,5,5,3
094,4,6,6,6,6,5
095,4,5,3,3,5,3
096,4,6,6,6,6,2
097,4,5,3,4,5,3
098,4,5,3,5,5,3
099,4,5,5,5,5,3
100,4,6,5,5,5,3
101,4,6,5,6,6,4
102,4,4,4,3,4,4
103,4,6,5,5,5,4
104,4,3,2,2,4,1
105,4,6,6,6,6,5
106,4,4,3,3,5,3
107,4,5,4,4,4,4
108,4,2,3,1,1,4
109,4,4,2,2,5,2
110,4,6,6,6,6,6
111,4,6,6,6,6,5
112,4,6,5,5,5,3
113,4,6,6,5,6,5
114,4,3,3,3,5,3
115,4,5,3,5,4,2
116,4,5,5,4,4,4
117,4,6,5,6,6,2
118,4,6,6,6,6,5
119,4,6,4,5,6,2
120,4,3,3,3,6,3
121,4,6,6,6,6,1
122,4,6,5,5,4,4
123,4,4,4,4,4,3
124,4,6,4,5,4,3
125,4,4,3,3,3,5
126,4,4,4,4,4,3
127,4,6,4,5,3,4
128,4,5,3,3,4,2

129,4,4,3,3,3,2
130,4,6,5,5,6,5
131,4,2,2,1,2,5
132,4,3,2,2,2,1
133,4,4,3,4,5,3
134,4,6,4,5,6,3
135,4,6,3,5,5,3
136,4,5,5,4,6,2
137,4,5,4,5,6,2
138,4,4,4,4,4,4
139,4,6,4,5,6,3
140,4,6,6,6,6,6
141,4,6,5,6,6,2
142,4,6,4,5,6,3
143,4,6,5,6,6,3
144,4,6,6,6,6,3
145,4,6,5,6,6,3
146,4,6,5,4,6,3
147,4,6,5,6,6,3
148,4,6,6,5,6,6
149,4,6,6,6,6,4
150,4,4,6,5,4,3
151,4,5,5,5,5,3
152,4,3,3,3,3,3
153,4,6,6,6,6,6
154,4,5,5,5,5,3
155,4,5,5,5,5,4
156,4,5,4,5,5,2
157,4,6,6,5,5,4
158,4,6,6,4,6,2
159,4,6,6,5,5,4
160,4,4,5,4,4,4
161,4,4,3,3,4,3
162,4,5,5,4,4,5
163,4,6,6,4,5,5
001,5,3,3,3,1,2
002,5,2,2,3,1,3
003,5,3,2,3,1,3
004,5,3,3,3,2,3
005,5,4,3,4,2,4
006,5,2,2,2,1,2
007,5,4,3,3,2,3
008,5,2,2,3,1,1
009,5,2,2,2,1,1
010,5,2,2,2,1,2
011,5,1,1,1,1,1

012,5,4,4,4,2,3
013,5,3,3,4,2,3
014,5,1,1,1,1,1
015,5,5,4,4,2,3
016,5,1,1,1,1,1
017,5,3,2,3,2,2
018,5,4,3,3,2,3
019,5,3,3,3,2,3
020,5,3,3,4,2,3
021,5,5,5,3,2,4
022,5,4,3,3,2,2
023,5,4,4,4,2,3
024,5,4,2,2,2,3
025,5,3,2,2,1,3
026,5,2,2,2,1,3
027,5,4,2,2,2,3
028,5,4,2,2,2,3
029,5,2,2,2,1,3
030,5,4,3,3,2,4
031,5,4,4,4,2,4
032,5,4,2,2,2,3
033,5,4,2,3,2,3
034,5,4,2,2,1,2
035,5,4,2,3,2,4
036,5,4,3,3,2,4
037,5,4,4,3,3,4
038,5,1,1,1,1,1
039,5,4,3,2,1,3
040,5,4,2,3,2,3
041,5,4,2,3,2,3
042,5,4,2,2,2,3
043,5,4,3,3,2,3
044,5,4,3,3,2,2
045,5,4,3,3,2,3
046,5,4,2,2,2,3
047,5,4,3,3,2,3
048,5,4,2,2,2,3
049,5,4,3,3,2,3
050,5,2,2,2,1,3
051,5,4,3,3,3,3
052,5,4,2,3,2,4
053,5,2,2,2,1,2
054,5,4,3,3,4,3
055,5,4,3,3,3,3
056,5,4,3,3,2,3
057,5,4,2,2,1,2

058,5,4,3,3,1,3
059,5,4,2,3,2,3
060,5,4,3,3,1,3
061,5,4,3,3,2,2
062,5,3,2,2,1,2
063,5,5,4,5,2,5
064,5,4,3,2,2,3
065,5,5,2,2,1,3
066,5,3,2,2,1,2
067,5,3,3,2,1,3
068,5,5,2,3,1,3
069,5,4,3,3,3,3
070,5,4,2,2,1,3
071,5,5,3,2,1,3
072,5,5,3,3,2,2
073,5,4,3,2,1,2
074,5,3,1,1,1,1
075,5,3,2,2,1,1
076,5,3,2,3,2,3
077,5,5,2,2,1,3
078,5,3,2,2,1,3
079,5,4,3,3,2,3
080,5,4,2,3,2,3
081,5,4,2,2,1,3
082,5,3,2,2,1,3
083,5,5,3,4,2,4
084,5,4,2,2,1,3
085,5,3,2,1,1,1
086,5,4,2,2,2,2
087,5,4,2,1,1,2
088,5,5,3,3,2,3
089,5,4,2,2,1,2
090,5,3,1,1,1,2
091,5,3,2,2,1,2
092,5,3,1,3,1,3
093,5,5,3,2,1,3
094,5,5,2,2,2,3
095,5,4,2,3,3,2
096,5,5,3,3,2,4
097,5,4,2,2,1,2
098,5,3,2,2,1,3
099,5,5,2,2,3,3
100,5,4,2,2,2,2
101,5,5,2,2,3,3
102,5,4,2,2,2,2
103,5,5,3,2,2,3

104,5,4,2,2,1,2
105,5,5,2,2,1,3
106,5,5,2,2,1,2
107,5,5,2,2,1,3
108,5,3,2,1,1,2
109,5,3,2,2,1,2
110,5,4,2,3,2,3
111,5,5,2,3,2,3
112,5,5,2,2,1,3
113,5,5,3,2,1,3
114,5,4,3,2,1,3
115,5,5,2,4,1,2
116,5,5,2,3,1,3
117,5,4,2,2,1,2
118,5,5,2,2,2,3
119,5,5,2,2,1,3
120,5,3,1,2,1,2
121,5,5,3,2,2,3
122,5,4,2,2,1,3
123,5,4,2,2,1,3
124,5,3,2,2,1,1
125,5,3,2,2,1,2
126,5,4,2,2,1,1
127,5,4,2,2,1,2
128,5,5,2,3,1,2
129,5,4,2,3,1,1
130,5,3,2,1,1,1
131,5,3,1,1,1,1
132,5,3,2,2,1,1
133,5,4,3,2,1,2
134,5,4,3,4,1,3
135,5,4,3,2,1,3
136,5,5,3,3,1,3
137,5,4,3,2,1,2
138,5,5,2,3,1,2
139,5,5,2,2,1,3
140,5,5,3,3,3,3
141,5,3,2,2,1,2
142,5,3,2,2,3,1
143,5,4,2,3,1,2
144,5,5,4,4,1,3
145,5,3,2,2,1,1
146,5,5,3,3,1,3
147,5,5,3,3,1,2
148,5,5,4,4,2,4
149,5,4,2,2,1,3

150,5,5,3,3,1,4
151,5,5,4,4,1,4
152,5,4,2,2,1,3
153,5,5,4,5,4,5
154,5,5,4,3,2,4
155,5,5,3,3,2,5
156,5,5,4,4,2,5
157,5,5,5,5,2,5
158,5,5,3,3,3,4
159,5,5,4,4,2,5
160,5,4,2,2,1,3
161,5,2,2,2,1,2
162,5,4,3,2,1,3
163,5,5,3,3,1,5
001,6,5,6,5,5,6
002,6,5,6,6,6,6
003,6,5,6,6,6,6
004,6,4,5,5,5,4
005,6,5,6,6,6,5
006,6,6,6,5,6,5
007,6,6,6,6,6,6
008,6,4,4,4,5,5
009,6,5,4,5,4,5
010,6,6,6,6,6,6
011,6,2,2,1,1,3
012,6,6,6,6,6,6
013,6,6,6,6,4,6
014,6,3,2,3,2,3
015,6,6,6,6,6,6
016,6,6,3,1,2,6
017,6,6,5,4,4,5
018,6,5,5,4,3,5
019,6,5,3,4,4,5
020,6,6,5,6,6,6
021,6,6,6,5,6,6
022,6,4,4,2,5,5
023,6,6,5,6,6,5
024,6,6,5,6,4,5
025,6,6,4,2,2,5
026,6,4,3,3,2,4
027,6,5,3,4,3,4
028,6,5,4,3,2,5
029,6,4,2,3,3,5
030,6,5,5,6,5,6
031,6,6,5,6,5,6
032,6,5,4,3,3,5

033,6,5,5,5,4,5
034,6,5,4,4,4,5
035,6,6,5,4,5,6
036,6,4,5,4,4,6
037,6,5,5,4,5,5
038,6,2,2,1,1,5
039,6,5,5,2,4,5
040,6,4,4,5,4,5
041,6,4,5,4,4,5
042,6,5,5,4,4,5
043,6,5,5,4,4,4
044,6,6,6,6,5,6
045,6,5,4,4,4,6
046,6,4,5,5,4,5
047,6,6,5,5,5,6
048,6,6,5,2,4,5
049,6,6,6,6,6,6
050,6,3,3,2,2,3
051,6,6,5,5,6,6
052,6,5,4,5,5,6
053,6,4,3,2,2,5
054,6,6,5,5,5,5
055,6,6,5,4,5,5
056,6,6,5,4,5,5
057,6,5,4,4,4,5
058,6,6,5,5,5,6
059,6,5,5,5,4,5
060,6,6,5,6,5,6
061,6,6,5,6,4,6
062,6,4,3,3,3,5
063,6,6,5,4,3,5
064,6,6,6,6,5,5
065,6,4,4,4,3,4
066,6,4,3,4,3,4
067,6,5,4,4,4,5
068,6,4,5,4,5,5
069,6,6,6,4,5,5
070,6,5,5,4,4,5
071,6,6,5,4,4,5
072,6,6,6,5,5,6
073,6,5,5,4,5,5
074,6,4,4,3,4,5
075,6,4,4,3,3,5
076,6,4,4,3,3,4
077,6,5,4,4,4,4
078,6,5,4,3,4,4

079,6,6,5,5,4,5
080,6,5,4,4,4,5
081,6,5,4,3,4,5
082,6,5,3,3,3,5
083,6,6,6,5,5,5
084,6,5,5,6,4,5
085,6,4,3,2,2,5
086,6,5,4,3,3,4
087,6,4,4,4,3,4
088,6,6,6,4,5,5
089,6,4,4,3,4,5
090,6,5,4,3,3,2
091,6,5,3,3,3,5
092,6,4,2,2,2,4
093,6,5,4,5,4,6
094,6,6,5,1,3,5
095,6,5,5,5,6,5
096,6,6,5,4,5,5
097,6,5,5,4,4,5
098,6,4,4,4,3,5
099,6,6,6,5,5,5
100,6,6,6,5,4,5
101,6,6,5,5,6,5
102,6,5,3,4,3,5
103,6,6,6,5,5,6
104,6,5,3,2,2,4
105,6,6,5,4,4,5
106,6,4,4,4,2,4
107,6,5,4,4,3,4
108,6,4,3,1,5,5
109,6,5,4,3,3,4
110,6,6,5,6,5,5
111,6,6,5,4,5,5
112,6,5,5,5,4,5
113,6,5,5,4,4,5
114,6,5,4,4,5,5
115,6,6,5,4,4,5
116,6,5,5,4,4,5
117,6,6,6,5,4,5
118,6,5,4,3,4,4
119,6,5,4,4,4,3
120,6,5,5,3,3,5
121,6,6,6,5,5,6
122,6,5,4,4,4,5
123,6,5,5,5,4,5
124,6,5,4,5,4,5

125,6,5,4,4,3,5
126,6,4,4,4,3,5
127,6,4,4,4,3,4
128,6,5,4,3,3,4
129,6,4,5,4,3,4
130,6,4,3,2,3,4
131,6,3,2,2,2,4
132,6,5,3,3,3,4
133,6,5,5,5,4,5
134,6,5,5,4,4,5
135,6,6,5,5,4,5
136,6,6,4,3,4,5
137,6,5,5,4,4,5
138,6,5,4,4,3,5
139,6,5,4,4,3,5
140,6,6,6,5,4,6
141,6,5,5,4,4,5
142,6,5,5,4,4,5
143,6,6,6,5,5,5
144,6,6,6,5,4,5
145,6,6,6,5,4,5
146,6,6,4,4,3,4
147,6,6,6,4,4,5
148,6,6,6,4,3,6
149,6,5,5,4,3,4
150,6,5,6,5,4,5
151,6,6,5,5,4,5
152,6,5,4,5,3,5
153,6,6,6,6,6,6
154,6,6,6,6,4,5
155,6,6,6,6,4,5
156,6,6,6,6,4,6
157,6,6,6,6,4,6
158,6,4,5,5,3,5
159,6,5,5,5,3,5
160,6,4,5,5,3,5
161,6,4,4,4,3,5
162,6,5,5,4,3,5
163,6,6,6,5,4,6

APPENDIX L: EXAMINEE MEASUREMENT REPORT

Participant	Measure (logits)	Model S.E.	Infit MnSq	Outfit MnSq
001	1.56	0.20	1.15	1.04
002	0.05	0.20	1.95	1.88
003	1.08	0.20	0.80	0.84
004	0.61	0.20	0.92	0.94
005	0.61	0.20	1.14	1.12
006	0.37	0.20	0.91	0.87
007	1.20	0.20	1.14	0.93
008	0.01	0.20	1.28	1.15
009	0.61	0.20	1.21	1.30
010	1.16	0.20	1.57	1.38
011	-2.21	0.27	2.03	1.85
012	1.40	0.20	1.05	1.04
013	0.76	0.20	0.53	0.52
014	-1.87	0.26	1.55	1.47
015	1.65	0.21	0.91	0.74
016	-1.26	0.24	2.72	2.45
017	0.53	0.20	0.69	0.69
018	1.00	0.20	1.06	0.92
019	0.45	0.20	0.47	0.42
020	0.96	0.20	0.84	0.84
021	1.61	0.21	1.55	1.20
022	1.20	0.20	1.82	2.15
023	0.76	0.20	0.60	0.61
024	0.53	0.20	0.64	0.64
025	-0.46	0.21	1.48	1.45
026	-1.26	0.24	1.16	1.10
027	-0.42	0.21	0.71	0.75
028	0.17	0.20	0.92	0.86
029	-0.99	0.23	1.05	1.26
030	0.45	0.20	0.76	0.77
031	0.92	0.20	0.68	0.70
032	0.25	0.20	0.99	0.96
033	0.92	0.20	0.90	0.77
034	0.33	0.20	0.93	0.90
035	1.24	0.20	0.91	0.83
036	1.08	0.20	1.11	1.08
037	1.32	0.20	1.00	0.92
038	-1.67	0.25	2.63	2.37
039	0.09	0.20	0.94	0.93
040	0.25	0.20	0.56	0.60
041	0.29	0.20	0.85	0.87
042	1.20	0.20	0.97	0.92
043	1.04	0.20	0.80	0.78
044	1.24	0.20	1.00	0.91
045	0.57	0.20	0.71	0.69
046	0.57	0.20	0.75	0.78

Participant	Measure (logits)	Model S.E.	Infit MnSq	Outfit MnSq
047	1.16	0.20	0.49	0.52
048	-0.12	0.20	1.11	1.12
049	1.40	0.20	0.97	0.96
050	-0.60	0.22	1.90	2.03
051	1.44	0.20	0.48	0.53
052	1.52	0.20	0.75	0.86
053	-0.84	0.22	0.79	0.74
054	1.69	0.21	0.85	0.83
055	1.87	0.21	0.65	0.77
056	0.84	0.20	0.45	0.46
057	-0.60	0.22	1.67	1.65
058	1.00	0.20	1.25	1.32
059	1.08	0.20	0.77	0.77
060	1.28	0.20	0.90	0.82
061	1.20	0.20	1.30	1.10
062	-0.51	0.21	1.18	1.20
063	0.96	0.20	1.28	1.36
064	0.76	0.20	0.87	0.98
065	0.09	0.20	1.53	1.47
066	-0.46	0.21	0.96	0.96
067	0.37	0.20	0.78	0.81
068	1.00	0.20	0.66	0.77
069	1.04	0.20	0.64	0.69
070	0.57	0.20	0.67	0.66
071	0.29	0.20	0.76	0.77
072	1.00	0.20	0.78	0.81
073	0.53	0.20	0.58	0.61
074	-0.70	0.22	0.96	0.97
075	-0.89	0.22	1.68	1.83
076	-0.70	0.22	0.81	0.83
077	-0.12	0.20	0.66	0.62
078	-0.12	0.20	0.73	0.70
079	0.92	0.20	0.25	0.25
080	0.09	0.20	0.58	0.59
081	-0.20	0.21	0.77	0.76
082	-0.51	0.21	0.66	0.62
083	0.68	0.20	1.02	1.08
084	0.80	0.20	0.72	0.69
085	-1.74	0.25	1.48	1.45
086	-0.12	0.20	0.73	0.73
087	-0.74	0.22	1.22	1.15
088	1.16	0.20	0.64	0.62
089	0.33	0.20	0.91	0.96
090	-0.60	0.22	1.24	1.32
091	-0.08	0.20	0.76	0.71
092	-0.42	0.21	1.19	1.28
093	0.21	0.20	0.72	0.70
094	0.05	0.20	1.74	1.58
095	0.49	0.20	0.99	1.02

Participant	Measure (logits)	Model S.E.	Infit MnSq	Outfit MnSq
096	0.21	0.20	1.42	1.36
097	0.17	0.20	0.90	0.91
098	-0.56	0.21	0.84	0.72
099	1.12	0.20	0.93	0.98
100	0.05	0.20	0.94	0.94
101	1.08	0.20	1.17	1.07
102	-0.56	0.21	0.70	0.72
103	0.57	0.20	0.73	0.76
104	-1.21	0.23	1.62	1.51
105	0.41	0.20	1.27	1.23
106	-0.65	0.22	1.12	1.04
107	-0.12	0.20	0.30	0.32
108	-1.15	0.23	2.05	1.85
109	-0.42	0.21	1.31	1.28
110	1.00	0.20	0.84	0.72
111	0.84	0.20	0.59	0.58
112	0.33	0.20	0.73	0.74
113	0.57	0.20	0.55	0.52
114	-0.29	0.21	0.95	0.80
115	0.29	0.20	1.15	1.19
116	0.33	0.20	0.55	0.55
117	0.29	0.20	1.14	1.10
118	0.53	0.20	1.50	1.40
119	0.29	0.20	1.25	1.24
120	-0.29	0.21	1.37	1.26
121	0.49	0.20	1.34	1.37
122	-0.16	0.21	0.64	0.63
123	-0.08	0.20	0.72	0.75
124	-0.29	0.21	0.82	0.83
125	-0.51	0.21	0.34	0.34
126	-0.08	0.20	0.93	1.02
127	-0.29	0.21	0.52	0.56
128	-0.16	0.21	0.89	0.86
129	-0.38	0.21	0.75	0.76
130	-0.20	0.21	1.26	1.13
131	-1.61	0.25	1.73	1.59
132	-1.10	0.23	0.99	0.94
133	0.33	0.20	0.95	0.99
134	0.49	0.20	1.18	1.14
135	0.05	0.20	1.03	1.03
136	0.21	0.20	1.14	1.06
137	0.09	0.20	0.79	9.68
138	0.01	0.20	0.49	0.52
139	0.17	0.20	0.83	0.76
140	0.88	0.20	1.04	1.10
141	-0.08	0.20	1.10	0.93
142	0.01	0.20	1.16	1.23
143	0.61	0.20	1.03	1.05
144	0.76	0.20	0.99	1.02

Participant	Measure (logits)	Model S.E.	Infit MnSq	Outfit MnSq
145	0.41	0.20	1.07	1.11
146	0.25	0.20	0.72	0.65
147	0.49	0.20	0.75	0.76
148	1.24	0.20	1.31	1.33
149	0.35	0.22	1.16	1.09
150	1.01	0.22	1.25	1.40
151	0.86	0.22	1.01	0.99
152	-0.59	0.23	0.68	0.67
153	2.42	0.27	2.16	1.54
154	0.82	0.22	1.05	1.04
155	1.06	0.22	1.22	1.11
156	1.31	0.23	1.43	1.85
157	2.08	0.25	1.61	1.66
158	0.39	0.22	1.58	1.66
159	1.26	0.22	0.94	1.00
160	0.30	0.22	1.02	1.10
161	-0.48	0.23	0.65	0.70
162	0.06	0.22	0.33	0.35
163	1.11	0.22	0.88	0.88
M	0.32	0.21	1.01	1.00
SD	0.81	0.01	0.41	0.38

APPENDIX M: REVISED ANALYTIC RUBRIC FOR AN ACADEMIC BLOG POST

User instructions: This rubric is intended as an instructional tool as well as for self- and/or instructor-assessment of an academic blog—or discussion board—post written in response to a particular set of questions. For each category in the left-hand column, there are a series of data-derived descriptors that correspond to a numerical value (1, 2, 3, or 4). The higher the number, the greater the quality of written work within this genre. Total scores will range from 5 to 20 points.

	1	2	3	4
Task Fulfillment & Relevancy	Does not address the questions from the prompt Text has no relevance to the prompt	Addresses prompt questions in a cursory or glancing way, and fails to answer two of the questions Likely contains some irrelevant information	Addresses all but one of the questions from the prompt A few details are not relevant or the connection between those details and the prompt is unclear	Addresses all questions from the prompt All details are relevant and directly connected to the main points (i.e., no superfluous information)
Content	Ideas are not developed No supporting details or examples provided Potentially a very short response	Provides a few examples, though the ideas are not developed or explored fully Text lacks solid arguments and detailed reasoning	Contains good points supported with personal and/or concrete examples, though argument could be strengthened with additional detail. Ideas are logical and specific, though text may be repetitive	Has numerous, thoughtful ideas that are supported with a variety of detailed examples The depth and development of evidence show strong grasp of content area
Organization & Balance	Little discernible structure as writer toggles between points too many times. Ideas do not build upon each other well or progress clearly	Ideas separated into paragraphs, although within-paragraph organization may not always be clear Paragraphs are unbalanced, with points either over- or underdeveloped	Overall and within-paragraph organization is solid, and ideas progress logically Paragraphs are evenly weighted, though the conclusion may feel abrupt	Organization is clear & engaging; even explicitly stated in the opening paragraph Fantastic introduction and conclusion tie well-structured, balanced piece together
Genre Specific Features	Makes use of **none** of the following: acknowledgment of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience/opinion	Makes use of **one** of the following: acknowledgment of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience/opinion	Makes use of **two** of the following: acknowledgment of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience/opinion	Makes use of **three or more** of the following genre specific features: acknowledgment of a particular group of readers; emojis; hyperlinks; questions for readers; personal experience/opinion

	1	2	3	4
Language Use	Usage errors hinder comprehensibility of most sentences Numerous spelling and grammar mistakes make text difficult to read Barely controls simple syntactic structures; undetectable voice	Several word choice errors detract from clarity of expression Spelling and grammar mistakes give text a “rushed” quality Uses some simple structures correctly, but struggles with complex syntax Voice may be inconsistent	Utilizes some sophisticated vocabulary, but may include the occasional awkward phrase Contains a few grammar and spelling errors that generally do not impede comprehension Some mixture of syntactic structures, though not error-free	Appropriate, varied, and sophisticated word choice Nearly free of spelling and grammar errors A good mix of simple and complex syntactic structures Primarily employs active voice

APPENDIX N: RHETORICAL MOVES ANALYSIS

Moves	Sample 1: 19_Tech	Sample 2: 29_Tech	Sample 3: 17_Relig	Sample 4: 11_Relig	Sample 5: 30_HW	Sample 6: 9_HW
Addressing potential readership		Good afternoon friends. As you already know…	Dear readers			Hello everyone!
Identifying audience of post		…this blog targets those people who are interested on foreign language teaching and learning				
Stating importance of topic	Language learning has become a key concern nowadays. Within an international context in which globalization is always expanding, stepping outside of your mother tongue is crucial for many aspects of your life.		Religion has always been a hot topic in every country but especially in Spain where the Spanish Inquisition is known all over the world	Religion is a controversial topic, and it will always be.	Education is a controversial issue in every country	Today I am going to discuss with all of you…an interesting topic related to studying and learning.

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
clarifying what was just stated	This means that when you learn an additional language to a certain extent, you could see it reflected not only in your own communicating skills, but you could also earn a higher salary, reinforce cerebral connections, delay the appearance of mental illnesses…		and even some tourists when they come to Spain ask to visit places where happenings related to the Spanish Inquisition took place		Each person has an idea of what an acceptable education is, but also each human has an opinion about the finest way to educate.	In fact, this particular topic has been quite controversial for years for teachers and students as well as parents. If you have guessed "homework" then you are right.
Re-addressing readership			This post may come in handy for all of those in love with Spanish culture and tradition and who are aware of the current controversies related to the topic of religion in the country (and if you are not take a sit and enjoy the journey of discovering Spain's relation to religion (;).			…my readers…

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Reporting background information without citations	it is common to see second or foreign language classrooms in educational systems all over the world		Religion is part and parcel of the lives of several Spaniards for better or for worse.	The matter gets even more complicated when it comes to school implementation.	Homework is part of the education of children, teenagers, and adults;	Teachers from all levels, from primary to secondary education and beyond, send homework to students so as to help them study whatever it is that they are teaching.
Elaborating or justifying previous proposition	not to mention other types of programs such as immersion programs, Erasmus exchange programs, the emergence of quick methods to learn a foreign language by the hand of different enterprises, and also…		In the same way that language is a product of culture, culture is a product of tradition and Spain has a long tradition for religion, especially Catholicism			
Questioning of shared knowledge					however, is homework necessary or a waste of time?	However, is all homework helpful?

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Introducing main topic of post	the use of technology for language learning.	so this post will be aimed at the influence of technology on these fields	So much so that religion is still a subject that students can choose to learn as part of their curriculum, or rather their parents chose for them, I should add.	However, in this brief entry we will try to give a solid opinion regarding several aspects closely related to religion.	This post is going to revise the pros and cons of homework	
Reporting background information without citations, could be known personally	the emergence of technology has taken place at such a rapid pace that now a great part of the population has a connection to the internet, or some type of electronic device that could be used, at least, for communication	The way in which foreign languages are currently taught (and learnt!) has changed drastically over the last 20 years	When my parents were at high school they could not decide whether to study religion or not			
Exemplifying what has just been stated	this can also be seen in formal education: primary school, secondary school, high school, university, schools of languages, academies, etc.		(it was compulsory and what is more their schools were run by nuns).			
Justifying what has just been stated	it is even normal to see how young learners take their electronic devices to class	a fact that may have been influenced by globalisation and the fast development of technology				

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Connecting background information to present composition	such is the case that technology can't just be left aside		However, when I was a student at high school I remember being able to choose between religion or the history of religions. History fascinates me and I wanted to join in that class but my parents said no, it was catholic religion and that was it.			
Outlining structure of post					First, the way homework can be helpful with students will be revised, then, a contrastive analysis of the pros and cons will be done; additionally, the factor of age will be taken into account concerning homework and finally, a conclusion will be stated.	

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Qualifying of upcoming proposition					Depending on the mode of instruction, in-person, or online,	
Stating answer to first prompt	technology can be used in several ways for language learning	Nowadays, a foreign language classroom cannot be imagined without any sort of technological component, and that is thanks to globalization and the internet	Nowadays, the choice is laxer	The situation in Spain is quite diverse.	the quantity and variability of homework assignments are different	
Presenting first example in support of answer to prompt	it is common to see overhead projectors that enable teachers and students to have a different learning experience far from the classic teacher-centered and blackboard-centered second or foreign language class	because teachers and students have immediate access to millions of resources to teach and learn foreign languages respectively	still in public schools it is a subject that is part of the curriculum. In private school the situation could be even more dramatic in the sense that there may be no choice, it is mandatory	From my experience in a public school (and high school), there is a subject specifically devoted to Christian religion	In the era of COVID-19, online courses are a trend, and the idea of having a flipped classroom (a mode of teaching in which the student studies and does all the homework and comes to class ready to put it in practice) is each time more extended.	
Qualifying of proposition				although it is completely optional		

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Clarifying previous proposition	This adds a sense of interactivity and dynamism into the learning environment that boosts the students' attention and makes it more enjoyable					
Providing hyperlink to learn more about subject			If you are interested in expanding your knowledge on the matter, read this article (it is highly insightful) Church and State in Spain			
Presenting second example to support answer to first prompt	we can also see how learners can upload information or tasks for class into different collaborative or open spaces on the internet, such as wikis or forums				In both modes of education, teachers and professors use homework to revise the content from class, learn new content, and settle the knowledge students have been learning.	
Initiating presentation of second example		For example, I remember, when I was learning English at school, our teacher brought us a newspaper from her holiday in England				

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Justifying what has just been stated		and that kind of amazed me, as I saw that content as rather exotic.				
Utilizing question to arrive at crux of example		But, what about nowadays?				
Presenting crux of example		Newspapers in all languages are accessible for people from all over the world				
Clarifying previous proposition		and that is a great step forward for language learning, as this allows students to access to real information immediately				
Restating first example		Of course, I was giving the example of newspapers,				
Presenting third example to support answer to first prompt	students can use online dictionaries, corpora, or translation systems when they have doubts about language…	but this could be applied to any available information on the internet: music, videos, films, blogs or even books.				

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Clarifying previous proposition		In the past, you had to travel to another country to reach them, but now we can access to them thanks to Internet shopping, Spotify, Kindle Unlimited, Netflix or YouTube, just to provide some examples.				
Announcing transition to next prompt	but not only computers or overhead projectors are useful for the language classroom	Besides	You may get the impression that I am against the teaching of religion at high school or school but that is not the case.			
Stating answer to second prompt		students may benefit from these sources	On the contrary, I'd rather be learning about religion or values that may be useful in society (even if you are not a believer) than spend one hour in a subject called "help in education" which was the subject available as an alternative of religion when I was a student	Personally, I like this approach.	Homework assignments have many positive effects.	On one hand, we can find a wide variety of activities and exercises that are well design to help study

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Justifying what has just been stated			which basically meant spending two hours a week doing homework or taking photos for your Instagram account with your friends.	It gives students the opportunity to expand their knowledge on religion if, and only if, they want to.		
Clarifying previous proposition		especially for the improvement of their receptive skills	That is such a waste of time			
Presenting first example to support answer to second prompt	we can see language learning technology in smaller and more personal devices such as mobile phones or tablets”	as the content available on them may guide students on a better understanding of the language	From where I stand, children can't learn anything negative about the world from two hours of instruction about religion.	There are even Catholic schools,	Nowadays, I am teaching an online course and students learn the majority of the content of the class by doing assignments and homework. Then, they come to our online classes and we put in practice what they have been studying and learning outside of class.	such as those I like to call “practical activities”.

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Clarifying previous proposition						Those are exercises that require the true understanding of the content and through some thinking you can work out the answer to the activity. In other words, to apply the content you are supposed to learn.
Justifying what has just been stated	this generates a microcosmos that enables the learner to make the language learning experience something more immediate and unique		Thus, I do believe that the positives far out-weight the negatives.	and although I'm completely unfamiliar with them, I suppose that instruction will be much more focused on religion.	This mode of learning is flexible since they choose when they do their homework according to their time and pace.	This is a real challenge to students and as Franklin said and I quote “tell me and I’ll forget, show me and I may remember, INVOLVE me and I’ll understand”.

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Clarifying previous proposition						Understanding is the main aim of teaching so that students can develop their critical thinking and not just remember irrelevant data they don't even know how to apply in the real world.
Re-stating information presented earlier in the post	I have mentioned the internet, forums, wikis, language posts by other professionals, online translators, and the use of corpora, among others.					
Presenting second example to support answer to second prompt	But what about the so called 'apps'?	improve reading and listening comprehension			Moreover, it is helpful because in the actual classes they put in practice what they have learned, in real-life situations and actions.	

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Justifying what has just been stated	Apps are easy to download and install in your mobile phone or tablet so you can quickly start using them with a single tapping of your thumb.	and, therefore, improve their marks when practising these skills.			In this case, homework will be useful as long as the student does them.	
Exemplifying what has just been stated	Language learning apps such as Babbel or Duolingo adapt their level to the learners' necessities so that they self-scaffold their learning and make progress in an adequate pace while using 'conversational chunks,' bit of vocabulary and grammar that are learned implicitly, and at the same time, treating the four main skills of writing, listening, speaking, and reading					
Announcing transition to next prompt		Nevertheless	Nonetheless,	Nevertheless	Although I consider homework a beneficial way to learn,	On the other hand,

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Stating answer to third prompt		this could also be a double-edged sword, especially for students,	I can see why people may be against it and it depends on the teacher.	a deficient implementation can always happen, and I strongly believe that the Spanish way is far from perfect.	they can begin to being useless	we can find other activities that are useless
Presenting first example in support of answer to prompt		as they may not be able to differentiate between a correct use of language and an incorrect one.	I remember my religion teacher being all supportive about us teenagers and all the problems that we were facing during this difficult period	Those kids will only learn about "their" religion;	when the student does not understand the content and theme of the homework, because of their individual differences, they may feel anxious or even depressed as a consequence.	where all you have to do is to reply questions that you can actually find within the written text of the textbook.
Justifying what has just been stated		Hence, this could lead on unconsciously acquiring words and using them incorrectly	we even talked about sex in class and there were LGBT people in the class that felt welcomed and found a place to be themselves			Those are no help to students, in fact it is just annoying and frustrating to students to “copy and paste” the text’s words.
Clarifying previous proposition		something that could be extremely dangerous, as this errors might be difficult to eliminate in a future		that is, the main religion in Spain.		So obviously, those aforementioned exercises are the ones teachers should avoid at all costs.

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Suggesting remedy to counter-position		Despite this negative aspect, the teacher could provide students with a list of reliable sources that can be helpful to increase their knowledge.		What about Islam, Judaism, Hinduism, etc.?		
Justifying what has just been stated		Therefore, the teacher can feel "safe", as s/he has warned their students, and has even guided them to reliable sources of information				
Presenting second example to support answer to third prompt			However, I remember my best friend in a different high school in my town where the teacher of religion was very old-fashioned and tried to impose old values. (In this case, an example to support the counter possibility)	It is certainly a narrow perfective that, at worst, can lead to religious fanaticism.	Additionally, homework can be useless if the student does not do them and they are part of the course,	
Exemplifying what has just been stated			There was a gay boy in the class and he was even asked to leave the class.		which can lead to a loss of part of the content from class.	

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Reframing answer in light of example			In this situation I fiercely believe that religion instruction is not supportive of the student and even dangerous for the personal development of the student.			
Providing hyperlink to learn more about subject			By the way, for all those curious here is the opinion of the church on the matter: Spanish Church leaders criticise government plans on religion in schools.			
Summarizing experience and announcing authority on topic			This has been my experience with religion and my opinion based on my first-hand knowledge.			
Announcing transition to next prompt	In this sense		Nevertheless	Still		Besides all this

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Stating answer to fourth prompt	technology can serve the purpose of language learning from several areas.		If I was the minister of education and I could have a say in the matter I would change the current subject of religion for one called "religions" where students can learn about all the religions of the world (past and present) so that they can understand the differences, their own religion and understand some of the current social and political issues that are going on nowadays on the world and that are sprung from religious conflicts	the fact that religion is somehow present and not compulsory in syllabuses certainly helps students in such a complex stage of their lives.	Age is a key factor when using homework with students	we must take into consideration the different ages of the learners.
Presenting first example in support of answer to prompt	In class, it can make learning more interesting		This idea comes from my personal experience with religion.	This is a way of seeing things, of course. As valid as anyone's.	In my personal opinion, the younger they are, the less homework they should have	I stand to what I just say however, sometimes for young learners it is more important the fact that they do homework just to achieve a studying routine

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Exemplifying what has just been stated	with the use of, for example, PowerPoint presentations		I was raised a catholic and I was Christianized, did my first communion and even my confirmation. At that time I did it because I liked it and felt it. However, nowadays I'm seeing that the evangelic religion appeals to me more and makes much more sense than Catholicism even if I still like some things about Catholicism.		Children need to experience real-life and play,	which they will need for further studies than the activity itself.
clarifying what was just stated			Hence, I do not know where I stand and a subject telling me the differences among all religions may have helped today in my present crisis of faith,		although homework may help them to learn, it could not be as useful as it is with older people.	Nevertheless, that doesn't mean that the exercise should be detrimental to their learning.
Presenting second example to support answer to fourth prompt	It can open a whole world of possibilities as well				As I mentioned before, my students are in college, and doing homework helps them with their course and learning.	For older students sometimes teachers don't send homework so they can actually use that time on "active studying"

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Exemplifying what has just been stated	by exposing learners to situations in which real-life language is used and facilitated with subtitles and online videos					which can again be a double edge sword depending on the type of activities they are given.
Presenting third example to support answer to fourth prompt	and it can facilitate the retrieval of information in case of doubt.					
Presenting fourth example to support answer to fourth prompt	It can also make learning easier and foster self-learning if the students look for the target information of the target language for themselves to reinforce what they know, to check and reformulate their linguistic hypotheses, or to consolidate information.					
Presenting fifth example to support answer to fourth prompt	It can facilitate writing					
Exemplifying what has just been stated	by means of online posts in blogs and wikis, or					

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Presenting sixth example to support answer to fourth prompt	it can ease the provision of language assignments					
Exemplifying what has just been stated	by using the Cloud or other apps like Google Docs					
Clarifying previous proposition	which, in addition, allows the teacher and other peers to give feedback to other students, so they can learn from each other and make revisions in their second/foreign language use.					
Connecting examples to background information presented at beginning of post	All the aforementioned possibilities of use are backed up by the fact that, as it was mentioned at the beginning of this post, globalization is now the general rule in present-day societies, and this pushes (almost inevitably) people to interact with the internet in different languages than their mother tongue					

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Exemplifying what has just been stated	whether by online chats (e.g. WhatsApp, Facebook, etc.), PC apps that allow video-conferencing (e.g. Skype), or just commenting on social media (e.g. Facebook, Instagram, Twitter, etc.).					
Summarizing overarching stance on topic	Thanks to these opportunities for the use of technologies for language education, students can learn a language and develop the different macro- and micro-skills of language, but they can also help to develop intercultural awareness and other aspects of education that are often forgotten in pedagogy.		To cut a long story short, religion is good for students whether they are believers or not.	But if we can all agree on something, is that an improvable implementation is better than no implementation at all.	All in all, homework can have its advantages and disadvantages.	Moreover we must consider that the way of learning and studying is different for every student,

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
Restating position	In my opinion, the use of technology for language learning is a potential source of knowledge than can be appropriately used in the classroom or in language teaching in general.		As I have mentioned, it can help them understand why they are believers or why they are not and chose the religion that best suits their needs and ideas.		Depending on the student and its age, it can be more or less helpful.	some may benefit from doing exercises and activities while others simply don't.
Qualifying of position			However, I wouldn't make it compulsory but an option so that everyone can freely decide if it is good for them or not.		As far as I am concerned, homework is very useful when learning new content and practicing new material;	
Citing a famous quote			However, and to conclude this post, as Chesterton points out " let your religion be less of a theory and more of a love affair"			
clarifying what was just stated			In other words, we should be theorising whether to teach it or not, it cannot be taught, it has to be loved, felt and accepted.		thus, I believe homework assignments have more positive elements than negative ones. :)	Every student is different and everyone has diverse abilities and characteristics.

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
evaluating topic via metaphor		I would say that technology is like a car, if you use it to travel long distances, then it is good, if you use it just to brag out about how fast you can drive and you end up having an accident, then it is bad.				
clarifying what was just stated		That is, if used wisely, technology can be your best friend in language learning and teaching, but teachers have to guide students towards a wise use of them by practising what they preach.				
Calling for future research	More research into the language learning potential of technology should be carried out, and more research into how to properly include technology in the classroom must be done.					

Moves	Sample 1: 19 Tech	Sample 2: 29 Tech	Sample 3: 17 Relig	Sample 4: 11 Relig	Sample 5: 30 HW	Sample 6: 9 HW
clarifying what was just stated	Up to this point, what we know and what has been done is just the tip of the beautiful -language learning- iceberg.					
Providing reader questions		And you? What do you think about the influence of technology on language learning? Do you agree or disagree?				
Encouraging reader comments		Do not hesitate on posting your opinion on the comments!				
Addressing readership						Well, I hope you liked my insight on this topic and see you soon!